The Application of Explicit Semantic Analysis in Translation Memory Systems

Oumai Wang

A thesis submitted to Imperial College London for the degree of

Doctor of Philosophy

School of Professional Development

Imperial College London

London SW7 2AZ
DECLARATION OF ORIGINALITY

I confirm that the presented thesis is my work. All referenced works are acknowledged. Parts of the thesis have been presented for publication in advance of submission of the thesis.
COPYRIGHT DECLARATION

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.
Abstract

Although translation memory systems have become one of the most important computer-assisted translation tools, the development of systems able to retrieve Translation Memory (TM) files on the basis of semantic similarity has hitherto been limited. In this study, we investigate the use of Explicit Semantic Analysis (ESA), a semantic similarity measure that represents meanings in natural language texts by using knowledge bases such as Wikipedia, as a possible solution to this problem. While ESA may be used to improve TM systems, at present the evaluation of semantic processing techniques in the context of TM is not fully developed because the use of semantic similarity measures in TM systems has been limited. The study hence aims to evaluate ESA for its specific application in TM systems. The evaluation is performed within a knowledge management framework as this provides a suitable technical context. A software platform called the ESA Information Retrieval platform was designed to test the performance of ESA in TM system tasks using three different text genres: technical reports, popular scientific articles and financial texts. The aim of the evaluation was not only to improve our understanding of how ESA can be applied to TM systems, but also to examine certain textual factors that may have an impact on their performance. It was found that the use of ESA was able to create different ways of utilising translation suggestions. On the basis of the results obtained, both the existing problems of using ESA in TM systems and the future perspectives of TM systems are discussed. This study not only contributes to our understanding of employing semantic processing techniques in TM systems but also presents a new knowledge management perspective for the development of translation technology.
Acknowledgments

I would like to express my very great appreciation to my supervisors Dr Mark Shuttleworth and Dr Bettina Bajaj for their supervision, encouragement and support throughout the course of my Ph.D. I am most grateful that Mark allowed me to work with him for three years, and that he was willing to take me on as an MSc teaching assistant in my second year of study and was the co-author of my first paper. I very much appreciate all Bettina’s constructive and critical advice on my thesis. I would like to thank my examiners Dr Felicity Mellor, Professor Ruslan Mitkov and Professor Andrew Rothwell for their time and very important and judicious comments. Special thanks are also due to my proof-reader Susan Peneycad. I really appreciate your help and patience.

During my years as a Ph.D. student, I have had the pleasure of sharing in the knowledge of a number of people. Some of them have very patiently but enthusiastically explained various technical and mathematical issues to me. This work would not have been possible without their advice and support. The list includes Zhen Li, Xinyue Fu, Rui Zhang and Ji Qi.

I appreciate having had Tuan-Chi Hsieh, Dawning Hoi Ching Leung, Boshuo Guo, Emmanouela Patiniotaki and Albert Feng-shuo Pai as colleagues at Imperial College. Many of my friends who supported me throughout my Ph.D. study also deserve mention. They are Yufei Huang, Juan Ojeda Castillo, Xiang Weng, Yujie Sun, Jianfeng Yu, Ning Niu, Suwen Chen and Peng Zhang. I would like to thank Liam Watson and Naomi Anderson-Eyles for their excellent administrative support.

Most importantly, I would also like to thank my beloved family for the support they have provided me through my entire life. You have taught me to be a strong, independent person and have guided me to find the definitions of love, courage and perseverance in my life.
List of Abbreviations, Acronyms, and Notations

ACA: Aircraft Accident Report
ASL: Average Sentence Length
ATSL: Average Translation Suggestion Length
BLEU: Bilingual Evaluation Understudy
CAT: Computer-assisted Translation
CR: Conceptually Related Translation Suggestion
ESA: Explicit Semantic Analysis
FS: Formally Similar Translation Suggestion
GVSM: Generalised Vector Space Model
IR: Information Retrieval
KM/ KMS: Knowledge Management/ Knowledge Management System
LCS: Lowest Common Subsumer
LSA/LSI: Latent Semantic Analysis/ Latent Semantic Indexing
MT: Machine Translation
NLP: Natural Language Processing
OMCS: Open Mind Common Sense
POS: Part of Speech
QTR: Queries/Translation suggestions Ratio
RN: Reuters-21578 Collection
SaaS: Software as a Service
SciAm: Scientific American Articles
SD: Standard Deviation
TERL Translation Edit Rate
TF-IDF: Term Frequency-Inverse Document Frequency
TM: Translation Memory
TMS: Translation Memory System
TREC: the Text Retrieval Conference
TTR: Type/Token Ratio

XML: Extensible Markup Language

\( \vec{a} \): Vector

\( \mathbf{R} \): Matrix

\[ \sum \text{ (Summation)} \quad \sum_{m=1}^{n} a_m = a_1 + a_2 + a_3 \ldots a_n \]

For example, \( \sum_{m=1}^{3}(m+1) = (1+1) + (2+1) + (3+1) = 9 \)

\[ \prod \text{ (Product)} \quad \prod_{m=1}^{n} a_m = a_1 \cdot a_2 \cdot a_3 \cdot a_4 \cdot \ldots \cdot a_n \]

For example, \( \prod_{m=1}^{3}(m+1) = (1+1) \cdot (2+1) \cdot (3+1) = 24 \)

\( \in \) (is an element of) \( \quad \) If \( a \in S \), it means that \( a \) is an element of the set \( S \).

\( |a| \) \( \quad \) \( |a| \) is the absolute value of \( a \).
List of Figures and Charts

Figure 1.1: Example of a TMS layout .................................................................21
Figure 1.2: IR process .....................................................................................22
Figure 2.1: Wiig’s Knowledge Typology ........................................................41
Figure 2.2: Dalkir’s Integrated KM Cycle ......................................................45
Figure 2.3: Nonaka’s Knowledge Spiral Model .............................................48
Figure 2.4: The transfer and conversion of knowledge in the workflow of a TMS seen as a type of KMS .................................................................56
Figure 3.1: A fragment of WordNet hierarchy .............................................66
Figure 3.2: A fragment of WordNet hierarchy .............................................67
Figure 3.3: A fragment of WordNet hierarchy with trained probabilistic values ..............68
Figure 3.4: Semantic relations in ConceptNet .............................................74
Figure 3.5: A sample of semantic network in HowNet ................................75
Figure 3.6: Extract from the Wikipedia article on 'oxygen' .......................89
Figure 3.7: Category page for 'oxygen' .......................................................90
Figure 3.8: A basic proposition of 'ozone' ...................................................91
Figure 3.9: An example of the language bar on a Wikipedia page .............93
Figure 3.10: An example entry for 'oxygen' in BabelNet .............................97
Figure 3.11: Another example entry for 'oxygen' in BabelNet .....................97
Figure 4.1: Extract of an RN text .................................................................119
Figure 4.2: A screenshot of original Scientific American articles .................124
Figure 4.3: A screenshot of an original AAIB aircraft accident report ...........125
Figure 4.4: A screenshot of an original AAIB aircraft accident report ...........126
Figure 4.5: A screenshot of TextStats ..........................................................128
Figure 4.6: A screenshot of XML2sql ..........................................................130
Figure 4.7: A screenshot of the ESA IR platform ........................................131
Chart 5.1: Distribution of ESA similarity scores for the test collections .......156
Chart 5.2: The relation between QTR and ranges of query length ..............172
Figure 6.1: An example of a parse tree .......................................................188
Figure 6.2: An example of a Wikipedia page about ‘微信’ (WeChat) ...........197
Figure 6.3: The infrastructure of Translation Memory Systems as a type of KMS .................................................................203
Figure 6.4.1: A screenshot from the Chinese version of Xiaomi mobile ........207
Figure 6.4.2: A screenshot from the English version of Xiaomi mobile .........208
List of Tables

Table 1.1: Example of an inverted index .................................................................................. 24
Table 3.1: An illustration of the transformation of ‘bag’ to ‘bug’ ........................................... 60
Table 3.2: A HowNet entry of ‘手表/watch’ ............................................................................ 71
Table 3.3: Matrix presentation of ConceptNet concepts .......................................................... 76
Table 4.1: An overview of phase 1 .......................................................................................... 123
Table 4.2: An overview of phase 2 ........................................................................................ 132
Table 5.1: Statistics from each of the three test collections ...................................................... 142
Table 5.2: Topic models of test collections ............................................................................. 143
Table 5.3.1: The typical spread of ESA similarity scores of outcome one .............................. 147
Table 5.3.2: The typical spread of ESA similarity scores of outcome two .............................. 148
Table 5.3.3: The typical spread of ESA similarity scores of outcome three .......................... 151
Table 5.3.4: Statistics of Levenshtein distance scores of the valid translation suggestions ....... 152
Table 5.4: ESA similarity scores for valid translation suggestions from ACA, RN and SciAm  .......................................................................................................................... 154
Table 5.5.1: Details of formally similar translation suggestions ............................................... 164
Table 5.5.2: Details of conceptually related translation suggestions ....................................... 164
Table 5.6: Relationship between TTR, ASL, size of test collections and the ESA similarity  scores ............................................................................................................................ 168
Table 5.7.1: Relationship between QTR and the ESA similarity scores; query length from the ACA test collection ........................................................................................................ 170
Table 5.7.2: Relationship between QTR and the ESA similarity scores; query length from the RN test collection ........................................................................................................ 171
Table 5.7.3: Relationship between QTR and the ESA similarity scores; query length from the SciAm test collection ..................................................................................................... 171
Table 5.8: Relationship between QTR and the ESA similarity scores; query length from  representative ranges with the largest numbers of results ...................................................... 173
Table of Contents

Abstract................................................................................................................................. 4
List of Abbreviations, Acronyms, and Notations................................................................. 6
List of Figures and Charts ...................................................................................................... 8
List of Tables .......................................................................................................................... 9
Table of Contents................................................................................................................ 10
Chapter 1 Introduction ........................................................................................................ 15
  1.1 Thesis Overview ........................................................................................................... 15
  1.2 Background of the Study .............................................................................................. 17
    1.2.1 Knowledge Management ...................................................................................... 17
    1.2.2 Information Retrieval ........................................................................................... 19
    1.2.2.1 The Process of Information Retrieval ............................................................... 22
    1.2.2.2 Evaluation Measures ....................................................................................... 25
  1.3 Scope of the Study ........................................................................................................ 26
    1.3.1 Aims of the Thesis .............................................................................................. 28
    1.3.2 Main Research Questions .................................................................................... 29
  1.4 Outline of the Thesis .................................................................................................... 31
Chapter 2 Literature Review: Knowledge Management in Relation to Translation Memory Systems... 33
  2.1 Overview of Knowledge Management and Knowledge Management System .......... 33
  2.2 Understanding Knowledge in Knowledge Management ............................................. 34
    2.2.1 Definitions of Knowledge in the Context of Knowledge Management ............... 37
    2.2.1.1 Holistic Method ............................................................................................. 38
    2.2.1.2 Knowledge Hierarchy Method ....................................................................... 38
    2.2.1.3 Resource-based Method ............................................................................... 40
    2.2.2 Types of Knowledge ............................................................................................ 41
    2.2.2.1 Wiig’s Types of Knowledge ........................................................................... 41
    2.2.2.2 Nonaka and Takeuchi’s Explicit and Tacit Knowledge Dichotomy ............... 43
  2.3 Important Concepts of Knowledge Management ......................................................... 44
    2.3.1 Knowledge Management Process ......................................................................... 44
    2.3.1.1 Dalkir’s Integrated Knowledge Management Cycle ...................................... 45
    2.3.1.2 Nonaka and Takeuchi’s Knowledge Spiral Model .......................................... 47
    2.3.2 Knowledge Management Approach ...................................................................... 49
    2.3.3 Knowledge Management Systems ........................................................................ 50
    2.3.4 Knowledge Management Technology ................................................................... 51
  2.4 Translation Memory Systems as Knowledge Management Systems: A Conceptual Framework . 52
2.4.1 Knowledge and Knowledge Types Used in Translation Memory Systems as a Type of Knowledge Management System ............................................................... 52
2.4.2 The Workflow of Translation Memory Systems as a Type of Knowledge Management System ........................................................................................................... 53
2.5 Conclusion ........................................................................................................... 57

Chapter 3 Literature Review: Explicit Semantic Analysis and Other Existing Methods for Measuring Semantic Similarity ............................................................................ 58
3.1 Translation Memory Systems and Similarity Measures ........................................ 58
   3.1.1 Levenshtein Distance .................................................................................. 60
   3.1.2 The Advantages and Disadvantages of the Levenshtein Distance Method ...... 61
3.2 Measures of Semantic Similarity ........................................................................ 62
   3.2.1 Resources for Measuring Semantic Similarity ............................................. 63
      3.2.1.1 Using WordNet for Measuring Semantic Similarity ................................ 64
      3.2.1.2 Using HowNet for Measuring Semantic Similarity ................................. 70
      3.2.1.3 Using ConceptNet for Measuring Semantic Similarity ......................... 73
      3.2.1.4 Limitations of Thesaurus Methods ..................................................... 77
   3.2.2 Corpus-Based Methods ................................................................................ 78
      3.2.2.1 Point-wise Mutual Information ............................................................... 79
      3.2.2.2 Latent Semantic Analysis and its Variants .............................................. 79
      3.2.2.3 Vector Similarity Metrics ................................................................. 82
      3.2.2.4 Limitations of Corpus-Based Methods ............................................... 83
   3.2.3 Evaluation of Semantic Similarity Measures ............................................... 83
3.3 Explicit Semantic Analysis .................................................................................. 85
   3.3.1 Wikipedia .................................................................................................... 86
      3.3.1.1 Brief History of Wikipedia .................................................................. 86
      3.3.1.2 The Structure of Wikipedia ............................................................... 88
      3.3.1.3 Research on Wikipedia ..................................................................... 93
      3.3.1.4 Wikipedia Dumps ............................................................................ 97
   3.3.2 Using Wikipedia as a Knowledge Base to Measure Semantic Similarity ....... 98
   3.3.3 Formalisation of Explicit Semantic Analysis ............................................... 99
   3.3.4 ESA as a Generalised Vector Space Model ............................................... 100
   3.3.5 Implementations of Explicit Semantic Analysis ......................................... 102
3.4 Conclusion ........................................................................................................ 103

Chapter 4 Methodology ......................................................................................... 104
4.1 Overview of the Evaluation Methodology ......................................................... 104
   4.1.1 Description of the Evaluation Method ...................................................... 104
Chapter 1 Introduction

1.1 Thesis Overview

The Internet has made a vast amount of language resources available to the translation industry for translation and localisation. The Springer *Language Resources and Evaluation* journal defines a language resource as

'Language data and descriptions in machine readable form used to assist and augment language processing applications, such as written or spoken corpora and lexica, multimodal resources, grammars, terminology or domain specific databases and dictionaries, ontologies, multimedia databases, etc., as well as basic software tools for their acquisition, preparation, annotation, management, customization, and use.' (Springer, 2012)

Ferraresi points out that language resources can also be used for translation purposes (2009: p. 1). In contrast to traditional reference sources, such as dictionaries, galleries or encyclopaedias, online language resources are affordable, up-to-date and easily accessible to translators (Ferraresi, 2009). Many linguists have begun studying online language resources. For example, the 'Web as Corpus' approach was developed to explore ways to transform online language resources into corpora for linguistic purposes (Kilgarriff & Grefenstette, 2003).

In this study, a translation resource is considered a language resource used for translation purposes. Austermühl lists several types of online resource that serve as translation resources, such as general and specialised encyclopaedias, monolingual dictionaries, bilingual dictionaries, multilingual terminology databases and newspaper and magazine archives (2001: p. 69). At the beginning of the millennium, Austermühl acknowledges only one dimension of online language resource: the information directly read by translators; that is, those translation resources that are easily accessible for confirming terms and phrases or retrieving background information (Austermühl, 2001: p. 85–86). However, he does not consider the possibility that a translation resource could also consist of large-scale, machine-readable information used to augment particular applications, such as translation memory systems (TMS).

Recently, the semantic processing of natural language has become an active and dynamic field. Significant research is being done to improve the efficiency of building and using large-scale resources and converging techniques from knowledge engineering (Tokunaga et al., 2008; Gurevych & Zesch, 2013). It is reasonable to expect that many new techniques from natural
language processing (NLP) will continue to be used to improve translation technology. As Reinke suggests (2013) that translation memory systems can be enhanced by employing ‘linguistic knowledge’, as it may potentially improve the performance of retrieval algorithm of translation memory systems and may reduce post-editing efforts by incorporating statistical machine MT techniques. Explicit Semantic Analysis (ESA), originally used for the computation of the semantic relatedness of words in text categorization, provides a relatively simple and effective method for representing meaning in natural language texts by using semantics from knowledge bases such as Wikipedia. In the context of NLP, a knowledge base is defined as ‘a database containing knowledge representations that can support an information system or an NLP program or application’ (Mitkov, 2005: p. 743). Wikipedia, a large, collaboratively edited, multilingual and free Internet encyclopaedia, is used as the knowledge base for ESA. ESA has also generated attention for its potential in information retrieval (IR) because it does not require an ad hoc knowledge base to enable IR system semantic processing. Computer-assisted translation (CAT) tools, especially translation memory systems, may improve the efficiency of using translation resources if they are able to incorporate text semantics. Early research has enhanced the performance of translation memory with semantic processing techniques, such as finding semantically equivalent sentences through syntactic and semantic analysis with lexical database (Pekar & Mitkov, 2007), or the identification of rhetorical predicates (Mitkov & Corpas, 2008), or employing ontology-based resources (Yao, 2010).

It is feasible to use ESA in TMS tasks. As suggested by the Institute of Translation and Interpreting, commercial and professional translations are mainly translations of specialised texts (2010). However, translators from linguistics and humanities backgrounds may have great difficulty understanding and translating such specialised texts. ESA enables translation memory systems to use semantics from the domain knowledge of external knowledge bases to retrieve translation suggestions at the semantic level. In this study, Wikipedia is the knowledge base used to perform semantic measures by ESA because it covers a range of specialised subjects. This ability to identify texts from specialised domains means that an ESA-enhanced TMS has the potential to outperform traditional translation memory systems. However, the outcome of using ESA to retrieve translation suggestions has not been studied. This gap suggests that there is not a straightforward method for evaluating the performance of ESA or any type of semantic processing technique in the context of translation memory systems. This
is because employing semantic processing techniques is an underdeveloped method for CAT tools.

In this study, this gap is viewed as a knowledge management (KM) issue among translation resources because the implementation of ESA in TMS tasks is a dedicated process of employing a knowledge base for a particular application. In this thesis, I argue that translation memory systems should be seen as a type of knowledge management system (KMS). A set of experiments is conducted to evaluate the possibility of using ESA in TMS tasks. Here, 'possibility' refers to whether ESA is able to retrieve more translation suggestions from given document collections. This study contributes to the understanding of employing semantic processing techniques in TMS tasks and presents useful framework for the construction of a knowledge management framework for the language industry. The remainder of this chapter is structured as follows:

- Background of the study
- Scope of the study
- Outline of the thesis

1.2 Background of the Study

The background of the study consists of two main disciplines: IR and KM. This section briefly introduces these two disciplines for the convenience of the discussions in chapters 2 and 3.

1.2.1 Knowledge Management

The notion of 'management' has been presented in previous research to establish the relationship between translation, translation resources and computer tools. For example, 'terminology management' is used to describe a generic process for the 'documentation, storage, manipulation and presentation of specialized vocabulary', and is usually used for specific translation projects (Austermühl, 2001: p. 102-103). Similarly, Budin (2005) proposes a framework of translation management as follows:
'A comprehensive concept covering all procedures of the computational management of translation process by using a broad spectrum of computer tools, modelling these processes into operational workflow models, and including economic and human resource management aspects.' (p.103)

However, neither of the approaches attempts to explain the 'computational management of translation process.' Thus, they cannot explain how computer tools, such as TMS, fit in the operational workflow of a translation task. Therefore, KM was used to accomplish the objective of this study.

To provide a suitable technical context for this work, the elucidation of certain concepts from KM is necessary. In the second half of the 20th century, many companies (typically consulting and accounting firms) identified knowledge instead of land, capital, labour or materials as the key to determining business success (Mentzas, 2003: p. 1). Normally, KM is a loose term used to refer to a broad collection of practices and approaches related to generating, capturing and sharing knowledge that is relevant to an organisation's business (Mentzas, 2003: 2–3). A generic definition of knowledge management is given by Dalkir (2005: p. 3) as follows:

'The deliberate and systematic coordination of an organisation’s people, technology, processes, and organisational structure in order to add value through reuse and innovation. This value is achieved through the promotion of creating, sharing, and applying knowledge as well as through the feeding of valuable lessons learned and best practices into corporate memory in order to foster continued organisational learning.'

KM is used to form a working definition of knowledge that enables translation memory systems to measure translation units on the basis of semantic similarity.

A KMS is an information system that supports or enables activities of knowledge management (Hall, 2009; Alavi & Leidner, 2001). Baskerville and Dulipovici (2006: p. 90) note that the technical infrastructures of KMS are closely associated with artificial intelligence. The KMS should serve the general objectives of KM, namely 'knowledge reuse to promote efficiency and innovation to introduce more effective ways of doing things' (Dalkir 2006: p. 166). Different technologies, such as data mining and content management systems, are also employed for KM purposes (Dalkir, 2005: p. 217). As a type of CAT tool, translation memory systems functions primarily as an information-retrieval platform that searches a database of previously translated text fragments (i.e., translation suggestions) to retrieve translation units similar to the one
currently being translated (Trujillo, 1999: p. 60–61). The features of ESA represent an external category of knowledge in a machine-readable format, which allows the KMS to measure semantic similarities between a query and a translation suggestion. The ESA technique enables this function and is expected to improve the efficiency of using TM files. The aim is to involve the KMS in building an operational framework for the discussion of semantic processing techniques in translation memory systems. Chapter 2 will demonstrate that the knowledge should be stored in a knowledge base; thus, translation memory systems may be viewed as a type of KMS. The implementation of ESA reflects an approach where the KMS uses TM files effectively.

1.2.2 Information Retrieval

In practice, translation memory systems have been called a 'translators' workbench' comprising a set of features integrated into a single working environment; however, the core function of translation memory systems is the process of retrieving previous translated texts (Trujillo, 1999: p. 59). This suggests that translation memory systems can be seen as a special type of IR. Because of the similarities in the technical architectures of IR systems and translation memory systems, any technique that enhances the performance of IR should also be applicable to TM tasks. Therefore, the improvement of the information-retrieval task will improve the actual performance of the translation memory systems. This section introduces IR and its relevance to translation memory systems.

IR is a branch of computer science that has been applied in many high-demand areas, such as online search engines, and can consist of many components, such as text classification and similarity measures (Tzoukermann et al., 2005). Baeza-Yates (2011) defines information retrieval as follows:

'Information retrieval deals with the representation, storage, organisation of, and access to information items such as documents, Web pages, online catalogues, structured and semi-structured records, multimedia objects. The representation and organisation of the information items should be such as to provide the users with easy access to information of their interests' (p.1).

Translation memory systems primarily provide text segments, particularly sentences, as translation suggestions rather than other kinds of information, such as images. Therefore, I will
focus on the text-retrieval task. In this study, different experiments evaluated the use of ESA in translation memory systems to retrieve texts.

TMS text-retrieval tasks are implemented in a way similar to IR system implementation. Baeza-Yates (2011: p. 58) represents an IR system formally as a quadruple:

\[ [D, Q, F, R(d_i, q_j)] \]

**Document collections (D):** In the context of IR, a document \((d_i)\) is generally the unit of information indexed in the system and available for retrieval, while a document collection \((D)\) is a set of documents that is employed to satisfy user requests (Jurafsky & Martin, 2009: p. 801). An IR system needs to represent documents in machine-readable formats, sometimes also called a logical view of the documents.

**Queries (Q):** In the context of IR, a query denoted as \(Q\), is the information requested by users. Similar to representations of documents, an information retrieval system also needs to provide the representations of queries.

**Framework (F):** An IR system requires a model to establish the relationship between queries and documents.

**Ranking (R\((d_i, q_j)\)):** An IR system typically returns more than one answer from document collections. Ranking is the algorithm that establishes an order of the retrieved \(d_i \in D\) based on \(q_j \in Q\). Documents appearing at the top of this order are considered more relevant to the user's query (Baeza-Yates, 2011: 58).

Accordingly, these elements have counterparts in translation memory systems. In this study, the document \((d_i)\) referred to a translation unit comprising marked-up text associated with its corresponding translation. Thus, document collections \((D)\) are TM files. A translation suggestion is the sentence (i.e., \(d_i\)) retrieved by translation memory systems from TM files (i.e., \(D\)). A query \((q_j)\) is the newly segmented source text (i.e., sentences to be translated). \(F\) represents the similarity measures to instruct translation memory systems to select relevant texts from document collections (i.e., TM files).
The layout of translation memory is a type of interface to access the IR system.

As the layout shows, the text highlighted in green indicates that it is to be translated. The sentence is seen as a query by the TMS. The translation suggestion is shown in the 'Fuzzy Matches' window, as the TMS matches a sentence from the TM files. In this case, the translation suggestion is a document retrieved by the TMS. As shown, the fuzzy matching score is 37%. Currently, most similarity measures used by translation memory systems are edit-distance methods to measure the string similarities between queries and translation suggestions from TM files. As will be demonstrated in Section 3.1, translation suggestions are assigned similarity scores by similarity measures and are presented in the order of these scores. However, in practice, ranking is less important in translation memory systems than in IR systems. Usually, not many translation suggestions can be found with TM files because of the size of the TM files or the ability of the matching methods.
1.2.2.1 The Process of Information Retrieval

The process of IR normally consists of five steps: 1) document pre-processing, 2) building the inverted index, 3) query processing, 4) matching queries to documents and 5) ranking; all of these are explained below. The following software architecture is used to illustrate how an IR system works:

![IR process diagram](Baeza-Yates, 2011:p.8)

**Document Pre-processing:**

In an IR system, each document and query must be represented. The first step of an information retrieval task is to pre-process the documents and queries. To accomplish this, various natural language processing (NLP) techniques are used. The process of stemming changes content words (such as verbs, nouns and adjectives) into their stems by removing
morphological variations. In contrast, function words, such as a/an/the, are normally removed, as they have less semantic meaning for retrieval (Tzoukermann et al., 2005: p. 530–531). For East Asian languages, word segmentation is also required, as words are not delimited in these languages (Ke, 2008: p. 166). For example, '斯柯达是一家位于捷克的汽车制造公司 (BT: Škoda Auto is an automobile manufacturer based in the Czech Republic)'. Unlike English, there are no white spaces between words in this Chinese sentence. Rather, readers rely on background knowledge to recognise individual words: '斯柯达 (Škoda) /是 (is) /一家 (a) /位于 (located in) /捷克的 (Czech Republic) /汽车制造公司 (automobile manufacturer)'. Usually, function words, such as '是 (is)', and '的 (of or 's'), are removed. In the context of IR, a term is a lexical item (or a phrase) that appears in a document collection. The pre-processing gives the representation of documents as a set of index terms (Jurafsky & Martin, 2009: p. 802).

Building the inverted index:

Index terms are used to give an indication of the document content, and IR tasks normally take place over the index. After index terms are produced in document pre-processing, IR systems build an inverted file called an inverted index. We can imagine a two-column table (cf. Table 1.1 below) in which selected index terms appear next to the location of occurrence of these words in the documents (Tzoukerman et al., 2005: p. 531–533). For example, we have the following six index terms and their numbers of occurrences in a document collection:

Knowledge Management, 89; Translation Memory file, 150; CAT, 15;
Knowledge Management System, 67; XML 4; CAT, 8; Information Retrieval, 29.

The numbers of occurrences are given next to the indexing terms. Because 'CAT' appears twice in the inverted index, it has been identified from two documents. Hence, the inverted index is produced as follows:
Table 1.1: Example of an inverted index

<table>
<thead>
<tr>
<th>Term</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translation Memory file</td>
<td>150</td>
</tr>
<tr>
<td>Knowledge Management</td>
<td>89</td>
</tr>
<tr>
<td>Knowledge Management System</td>
<td>67</td>
</tr>
<tr>
<td>Information Retrieval</td>
<td>29</td>
</tr>
<tr>
<td>CAT</td>
<td>15, 8</td>
</tr>
<tr>
<td>XML</td>
<td>4</td>
</tr>
</tbody>
</table>

Term weighting is often conducted because not all terms are equally important in describing the document contents (Manning, 2009: p. 117). Term frequency (TF) and inverse document frequency (IDF) are the two most important factors; therefore, a TF-IDF weighting scheme is employed in most IR tasks (Baeza-Yates, 2011: p. 67–75). The TF-IDF weighting is computed as follows:

First, term frequency is the number of occurrences of each term in all of the documents. TF weight $tf$ is normally computed as a logarithmic scale in base 2, where $j$ stands for the number of different documents and $i$ for the number of different terms. TF weight is defined as follows (Baeza-Yates, 2011: p. 73):

$$tf_{i,j} = \begin{cases} 
1 + \log tf_{i,j} & \text{if } tf_{i,j} > 0 \\
0 & \text{otherwise}
\end{cases}$$

where $tf_{i,j} > 0$ means that term $i$ appears at least once. Otherwise, its TF weight is 0.

Second, inverse document frequency is another approach to assigning weight to discriminative terms. The intuition is that a more important term would appear in many documents from a document collection. We denote $N$ as the number of documents in a document collection and $n_i$ as the number of documents that contain term $i$. IDF weight is defined as follows (Baeza-Yates, 2011: p. 74):
\[ \text{idf}_i = \log \left( \frac{N}{n_i} \right) \]

The TF-IDF weighting is the combination of idf weight and tf weight. The TF-IDF weight \( w_{i,j} \) is defined as follows (ibid.):

\[ w_{i,j} = \text{idf}_i \times t f_{i,j} \]

**Query processing and matching queries to documents:**

After the inverted index is created, the IR system is able to process a query sent from the user interface. Similarly to how documents are represented, queries must be pre-processed. Queries are represented to match the index terms so that the IR system can match the relevant documents (Tzoukerman et al., 2005: p. 530). Baeza-Yates (2011: p. 61) regards the Boolean, vector space and probabilistic paradigms as three classic IR models. In this study, I focused only on the vector space model due to its relevance to ESA. This will be demonstrated in Section 3.3.

**Ranking:**

An IR task may yield a set of documents that are ranked in a particular order in response to a user query. An effective ranking algorithm lists the more relevant documents first.

**1.2.2.2 Evaluation Measures**

IR is essentially about finding the relevant information about a topic. IR systems deal with documents in natural language, which means they are not structured like other forms of data. In this context, it is vital to point out the differences between IR and data retrieval, which only processes well-defined data. Data retrieval systems only provide absolute answers about factual data (Baeza-Yates, 2011:p.5). For instance, there is only one definite answer to the question of the number of employees of the University of Oxford. On the other hand, the retrieved IR results to a non-factual query might be inaccurate and contain errors.
Normally, two retrieval metrics evaluate the performance of IR systems; these are **precision** and **recall**. In this study, evaluation measures consider not only the IR system, but also the performance of the TMS and semantic similarity. This issue will be examined in Chapter 4.

### 1.3 Scope of the Study

The possibility of using ESA in TMS tasks can be multi-dimensional. For example, a matching method may improve the quality of translation suggestions. This study will focus on the extent to which ESA can improve the efficiency of TM files in the retrieval of more translation suggestions. ESA will be compared to the Levenshtein distance method, which is the current similarity measure method used in translation memory systems. Their performance will be evaluated using large quantities of data. This research will establish a comparable quantified evaluation measure to demonstrate the performance of ESA and the Levenshtein distance method.

The scope of the study comprises the performance of semantic similarity measures that are supported by certain types of machine-readable resource, rather than the human perception contained in reference works. The ability of translation memory systems to retrieve translation suggestions and the impact of translation suggestions on translations are two completely different enquiries. The impact of translation suggestions is determined largely by the translator's proficiency, which is not included in the scope of the study. For such enquiries, translators' performance would be examined under conditions that provide selected, relevant references.

It should be noted that the main reason why English to Chinese is the primary language-combination used in this study is because of my own language proficiency: I am a native speaker of Mandarin Chinese and English is for me an acquired language. This language pair consists of two languages that are not at all closely related to each other, which means that translation between them will be likely to entail significant amounts of syntactical restructuring and a lower likelihood of one-one equivalences being available on the word level. While this study does not aim to be language specific in nature, it is possible that the conclusions may have a greater validity for language pairs such as English to Chinese that consist of languages that are not closely related.
The present research will focus on texts with the following features:

1) The texts will contain densely packed information including highly diverse terms, as well as standardised syntax or normative forms to express complex meanings produced by and intended for specialists. Examples of these kinds of text will be taken from scientific, technical and legal domains. Take the following sentence for example: 'One way to toughen polymers is to incorporate a layer of rubber particles and there has been extensive research regarding the rubber paraphrasing method of polylactide (PLA).’ This sentence contains technical terms, such as ‘polymers’, ‘polylactide’ and ‘rubber paraphrasing method’. Translators must be aware that some words are translated differently according to topic. For example, 'toughen' is translated as ‘锻炼 (BT: exercise)' in common Chinese texts but is more likely to be translated as '增韧 (BT: strengthen)' in texts relating to chemistry.

2) The production of the translation will fall within industrial procedures, namely, translators receiving payment and working against deadlines. This will ensure that the translations are produced for those who require good quality work completed within a limited time.

3) The texts will be of sufficient length; that is, they will include numerous textual and stylistic phenomena, such as the ambiguity of natural language, anaphoric reference and ellipsis, which can cause problems for natural language processing applications, such as machine translation systems. Take the following sentence: 'He gave a lecture about liver cancer at the hospital last January'. This phrasing is ambiguous because it may refer to January's lecture or the cancer cases that same month (Glasman-Deal, 2010:p56).

4) The authors of the source texts are human beings using natural language, and the recipients are also human, generally from a community of specialists. This ensures that the texts are not products of text generation applications and that the translations do not exceed the processing ability of a human being. This means that certain practices involved in transforming information from linguistic databases or sources are not included in this research. For example, items from a localised corpus or database (e.g., WordNet) translated from one language to another or published by dictionaries (e.g., the Oxford English Dictionary) will not be discussed in the context of this notion of translation.
The above features will regulate the selection of texts that are suitable for professional translators and form reliable and realistic criteria for evaluating ESA in translation memory systems.

This research will not develop ESA algorithms from a developer's perspective, although necessary paraphrasing methods will be used to ensure a successful implementation of ESA in the context of the study. Different implementations of ESA may perform differently. It should be noted that the results produced in the experiments do not aim at representing the best possible performance of the ESA technique as it may be limited by the test collections available to the researchers and the technical capacity of the ESA implementation created by its original developers. The performance of ESA in this work will be based on one IR platform selected from different research teams. This is discussed in more detail in Section 3.3.3.

1.3.1 Aims of the Thesis

The purpose of this thesis is to study the application of ESA in TMS tasks. As explained in Section 1.2, the use of ESA in TMS tasks will be conceptualised in a knowledge management framework, and the use of semantic processing techniques will be discussed. According to the scope of the study specified in Section 1.3, the thesis can be summarised in the following aims:

1. **Construction of an operational framework for discussion of the use of semantic processing techniques in translation memory systems:** The argument will be made that translation memory systems are a type of KMS, and ESA is used to improve the workflow of translation memory systems as a type of KMS. This aim is addressed in Section 2.5 after key notions of KM studies are revisited.

2. **Investigation of semantic similarity measures:** The thesis will assess semantic similarity measures and demonstrate the feasibility of using ESA in TMS tasks. Chapter 3 reviews some representative methods for measuring semantic similarity.

3. **Design of a software platform for evaluating the possibility of using explicit semantic analysis:** Because no prototype is available, a software platform will be created to test the performance of ESA in TMS tasks. It will be an original work, as considerable effort will be made to design a test suitable to reflect the performance of
ESA in TMS tasks with low costs. The core function of this software will be to conduct pseudo-TMS tasks by measuring semantic similarities between queries and texts from document collections according to the ESA technique. In this study, the pseudo-TMS task will instruct the software platform to match a query and sentences from a document collection (i.e., source text segments of translation suggestions), while ignoring whether they are actually translated in the targeted language. The procedure for preparing tools and materials is described in Chapter 4.

4. **Defining of measures for evaluating Explicit Semantic Analysis**: An argument is made that the possibility of utilising ESA for TMS tasks is based on whether ESA methods enables TMSs to retrieve a wider range of potentially useful translation suggestions. Two research questions are proposed in Section 1.3.2. The answers to these research questions provide a comprehensive understanding of the use of ESA in translation memory systems.

5. **Examination of relevant issues regarding involvement of Explicit Semantic Analysis in translation memory systems**: Based on an analysis of significant findings of evaluation experiments (Section 6.1), two issues relevant to translation memory systems are discussed:

- Bottlenecks in using ESA are shown (Section 6.2);
- The future perspective of translation memory systems is also outlined (Section 6.3).

### 1.3.2 Main Research Questions

ESA is the subject of ongoing research, as will be discussed in Section 3.3.2. The use of ESA in TMS tasks is still far from the product stage, so this study will not create an ESA-enhanced TMS as a prototype for real end-user field-testing. Regarding the current development of ESA, a more realistic approach will take ESA as a component of semantic similarity measures in the TMS workflow (Hirschman & Mani, 2005: p. 417). Therefore, the experiment will focus on the most important function of ESA in TMS: that is, improving the efficiency of using TM files by
retrieving translation suggestions that are potentially useful for translators. This evaluation will seek to answer the following research questions:

**Question A:** To what extent is the ESA similarity score useful for translation memory tasks?

**Question B:** How do textual factors affect the performance of ESA?

Question A will be evaluated through an automated process based on a large number of texts, which will mainly reflect the performance of ESA in TMS tasks. The performance of ESA will also be compared with the Levenshtein distance method. Question B will be based on statistical analyses of evaluation measures and human observations from ESA performed on the test collections, which contain rich examples of a number of specific features. Textual factors are the surface features that are 'readily identifiable linguistic features without syntactic or semantic parsing' (Xiong, 2009). More detail on textual factors will be provided in Section 4.1.4.3. These two research questions will be used to develop a comprehensive understanding of the possibility of employing ESA in TMS tasks.

The evaluations of NLP applications vary across different levels of maturity: user-centred evaluations are used for mature technology, while technology-based evaluations are suitable for technology that is still being developed (Hirschman & Mani, 2005: p. 415). ESA and other semantic similarity measures have not yet been embedded in a known commercial TMS. The evaluation of ESA will not take place in a user-centred style (Hirschman & Mani, 2005: p. 416). The user-centred style evaluation involves a judgement of the quality of ESA translation suggestions as machine translation systems, an estimation of the post-editing costs of translation suggestions retrieved by ESA and a testing of end users' acceptance of ESA-enhanced translation memory systems based on certain forms of software quality standard (Hirschman & Mani, 2005: p. 421–424).

Although no prototype software will be used, the experiment will conduct tests in a simulation in which translation memory systems perform the IR tasks (i.e., pseudo-TMS tasks). The following elements will be required for the simulation:

1. Document collections
2. Potential queries
3. ESA implementation
4. Wikipedia dump
The document collections and translations should satisfy the features mentioned in Section 1.3 and should relate to different genres. An ESA implementation is a software platform that can measure the semantic similarities of text fragments at the sentence level. A Wikipedia dump will be used to provide the required training data for ESA. The details of ESA processing are provided in Section 3.3, while those of the experiment design are presented in Chapter 4.

### 1.4 Outline of the Thesis

Having specified the scope of the study, the remainder of this thesis is structured as follows:

**Chapter 2:**

In this chapter, the conceptual framework of KM is reviewed. After reviewing a number of important concepts from the KM literature, the relation between knowledge management systems and translation memory systems is discussed. In the second half of Chapter 2, I provide a new perspective for conceptualising TMS workflow within the framework of KM: namely, translation memory systems are a type of KMS that assists in translation by offering required information. TMS as a type of KMS is an operational framework for the technical discussion of semantic processing techniques such as ESA.

**Chapter 3:**

Chapter 3 presents the technical background of ESA. Similarity, semantic similarity and similarity measures are all defined in the first section of this chapter. In the rest of the chapter, I systematically examine two types of method for measuring semantic similarity: thesaurus-based methods and corpus-based methods. As a semantic similarity measure, ESA is a kind of corpus-based method.
Chapter 4:

The first part of the chapter explains the research questions for the evaluation and the reasons why they are proposed. In Section 4.1, special features and difficulties of evaluating ESA in translation memory systems are shown. Some similar evaluation methods are also reviewed. The second part of the chapter describes the experiments, including tools and materials used and procedures.

Chapter 5:

Chapter 5 has three sections. The first section shows the details of the test collections used in the evaluation. The following sections provide results corresponding to each sub-question of the experiments.

Chapter 6:

This chapter comprises the analysis of the results presented in Chapter 5. There are four significant findings from the results. In Section 6.2, two problems that may affect the performance of using ESA are discussed. This mainly concerns Chinese translation. Section 6.3 outlines the perspectives for translation memory systems in the future. Some possible consequences of translation memory systems employing the notion of KMS are presented, and the possibility and feasibility of translation memory systems becoming a type of media for both the translation and non-translation industry is also discussed.

Chapter 7:

As conclusions, chapter 7 assesses the outcome of each aim and recommends four possible directions for future work.
Chapter 2 Literature Review: Knowledge Management in Relation to Translation Memory Systems

In this chapter, I review the framework of KM and argue that translation memory systems are a type of KMS that assists in translation by offering required information. The TMS workflow is presented to demonstrate how ESA could enhance the performance of translation memory systems. This chapter explains how we understand knowledge; how we manage it; what is the relation between TMS, KM and semantic processing techniques.

2.1 Overview of Knowledge Management and Knowledge Management System

KM is used as an operational framework to provide necessary definitions of concepts and to evaluate ESA for its use in TMS tasks. The KM framework is used to avoid defining all necessary concepts from scratch. These concepts, such as knowledge, management and knowledge base, are also used in different disciplines. It would cause confusion if these terms were not used in the proper context. Concepts such as knowledge, types of knowledge, KM process and KM approaches are defined within the conceptual framework of KM. They are the essential concepts to understand KMS. To evaluate the performance of ESA in TMS tasks, the framework should consider a blended understanding of computational processing of knowledge and the use of TMS by human translators. To understand translation memory systems as a type of KMS will ensure that the two perspectives are covered by this study.

KMS and KM are interrelated. KMS can only be understood properly within the context of KM because KMS is based on several concepts from KM, such as KM approach and the KM process. These KM concepts are used to outline the technical architecture of KMS and to describe the workflow of translation memory systems as a type of KMS. Thus, it is necessary to review the KM literature prior to the technical implementation of ESA. Key concepts of KM are reviewed and their relevance to the study is discussed. This study does not intend to address the research issues in KM or KMS; thus, the KM literature review is not intended to reflect the latest trends in KM.
KM is a generic concept that refers to the process of creation, sharing and application of knowledge (Stevens et al., 2010:p.131–132). Due to the multidisciplinary nature of KM, most KM theorists agree that it is impossible to have a uniform definition of KM (Anand & Singh, 2011; Baskerville & Dulipovici, 2006). Hlupic et al. (2002) identify 18 definitions, and Dalkir (2005) reports 72 working definitions. A comprehensive survey by Baskerville and Dulipovici (2006:p.85–86) found that the theoretical foundation of KM comes from 18 theories of different disciplines including organisational behaviour and artificial intelligence.

The perspective of KM for this study is based on the study of organisational behaviour, which explores the empirical process of knowledge used by people or groups (Dalkir 2005:p.5). This empirical process can be described as follows:

‘focusing on determining, organizing, directing, facilitating, and monitoring knowledge-related practices and activities required to achieve the desired business strategies and objectives’ (Wiig, 1993:p.38).

Numerous conceptual models have been developed to describe, analyse and support the implementation of KM. The conceptual framework of KM specifies how KMS is constructed.

### 2.2 Understanding Knowledge in Knowledge Management

Knowledge referred to in this study must be understood before the practices of KM are discussed. A working definition of knowledge helps identify manageable objects or processes for a KMS. In this study, the knowledge involved in the TMS workflow should be understandable to both human beings and computers. ‘Understandable’ knowledge means machine-readable information for computers; it is also the information used to assist the translation process. For convenience, three categories of knowledge need to be distinguished:

1) The knowledge that is manipulated directly by the TMS;

2) The knowledge that is used within the TMS to enhance its performance;

3) The knowledge that is used by translators to employ translation suggestions.
This chapter centres on the study of the first category of knowledge. But it is also useful to examine the other two, because they are also related to each other in the framework of translation memory systems as a type of KMS.

Apart from TM files, the ESA-enhanced TMS requires a **knowledge base**, which is defined as a database that contains concisely and formally structured information to enable particular applications (Jones, 2009:p.143; Mitkov, 2005:p.745). In this case, the **second category of knowledge** is the information stored by the knowledge base. From the perspective of computing, Raynor (2009:p.154) defines knowledge as follows:

> ‘the values of the parameters and rules that have been learned (estimated) from data in a model represent the “knowledge” of a model’

According to this definition, knowledge can be manipulated by computers in different ways, for instance ontology (Sowa, 2000:p.132; Jones, 2009:p.144). Ontologies store information to enable inferences by specifying ‘a collection of names for concept and relation types organised in a partial ordering by the type-subtype relation’ (Sowa, 2000:p.493). It is the knowledge that enhances the performance of translation memory systems. However, the second category of knowledge is knowledge that is not often involved in the TMS workflow, as I will show in Section 2.4.2.

The knowledge involved in the TMS workflow is also different from the **third category of knowledge** that is used by translators. The understanding of knowledge in translators’ minds is based on the context of translation training or descriptive translation studies, which goes beyond the scope of this study. Several researchers from descriptive translation studies and translator training have presented different ideas about the importance of knowledge in technical translations, especially knowledge other than linguistic knowledge (Kim, 2006; Collombat, 2006; Wilss, 1994&1996; Delisle, 2003; Kastberg, 2003; Robinson, 2003).

Wilss’ study on translation practices (1994) was deeply influenced by Levý’s (1967) theory of treating translation as decision making. Levý suggests that there are alternative options for solving every translation problem and that the chosen options affect the choices between the options that follow. Wilss (1996) argues that translation is a knowledge-based activity to solve translation problems, and that translators need two types of knowledge: declarative knowledge (knowing what) and processual knowledge (knowing how). Translation practices are formed by using processual knowledge as a set of skills to process the semantic information of translation.
Wilss’ study on translation practices (1994) was deeply influenced by Levý’s (1967) theory of treating translation as decision making. Levý suggests that there are alternative options for solving every translation problem and that the chosen options affect the choices between the options that follow. Wilss (1996) argues that translation is a knowledge-based activity to solve translation problems, and that translators need two types of knowledge: declarative knowledge (knowing what) and processual knowledge (knowing how). Translation practices are formed by using processual knowledge as a set of skills to process the semantic information of translation.

To some extent, Wilss’ (1996) explanation of the relationship between knowledge and translation can be backed up by Kim’s (2006) research on training translators. Kim (2006:p.287) has conducted thinking-aloud protocol research on three groups of Korean speakers: translation students, professional translators and English-language learners. Kim has found that translation students who have better awareness of the subject matter outperformed professional translators who mainly rely on dictionaries in terms of presenting meanings of source texts (2006: p.291–293). Translation students preserved rhetorical styles at a nearly professional level.

Delisle (2003) states that extra-linguistic knowledge is cognitively inseparable from the translation process. The extra-linguistic knowledge suggested by Delisle is a type of ‘pragmatic’ understanding, stated as ‘the interests if the participants in communication, their lines of reasoning, position, wishes, weakness, interaction, etc. the better the chances of understanding the sender’s discourse more accurately.’(2003:p.88). On that basis, Collombat also argues that general knowledge is as important as knowledge of the subject in technical translation and should be used as a ‘basic translation problem-solving tool’ (2006:p.61). In the field of translation training, Kastberg (2003) contrasts different approaches to integrating subject knowledge in training curricula and proposed a ‘personal knowledge management’ framework to help translation students gain sufficient knowledge for the real-world environment.

Robinson proposes that three types of intellectual activity are required in the translation process, namely abduction, induction and deduction, (2003:p.86–89). Abduction is a term coined by Charles Peirce (Robinson, 2003:p.87) to refer to the act of coming to a rough conclusion based on intuitive understanding. Induction and deduction are two types of logical reasoning: ‘induction beginning with specifics and moving toward generalities, deduction beginning with general principles and deducing individual details from them’ (Robinson, 2003:p.87). Robinson (2003:p.86) argues that translators may need different sets of skills and sources of information to conduct these activities.
Concepts of knowledge were not clearly defined in the works discussed above. Knowledge was used interchangeably with some concepts such as, 'intelligence', 'valuable information', and 'problem-solving skills'. The lack of working definitions and typologies of knowledge results in many conceptual obstacles. Many translation scholars often only address a certain perspective of knowledge such as the usefulness of knowledge of a subject matter. Without a clear definition of knowledge, it is difficult for translation scholars to have a specific understanding of knowledge in the use of TMS.

These drawbacks suggest that translation studies do not have a working definition of knowledge in the use of TMS. We must determine how to present the knowledge that functions in the TMS workflow. More importantly, we must focus on a perspective of knowledge that clarifies how knowledge manipulated by TMS is related to the knowledge manipulated by ESA. Such a perspective of knowledge is found in the context of KM. The next section approaches the first category of knowledge from the perspective of KM.

2.2.1 Definitions of Knowledge in the Context of Knowledge Management

Knowledge is a subject with many dimensions, the definitions of which are discussed here in the context of KM. Knowledge is also studied in other disciplines such as terminology, which focuses on the relations between language for special purposes and extra-linguistic knowledge (Bajaj, 2003:p.81-85), and philosophy, which addresses the moral and ethical properties of knowledge, and the ultimate source of knowledge or the mind and consciousness perceiving knowledge (Alavi and Leidner 2001:p.107–108). However, their perspectives are not used here, as Alavi and Leidner point out that such perspectives cannot be factors in building KMS for any kind of purpose (2001:p.108).

In the context of KM and KMS, knowledge is understood in a pragmatic way, rather than being theoretical or epistemological. Most KM researchers have reached a consensus that knowledge is a valuable, intangible object and a manageable factor that brings benefits such as the improved process of decision making (Dalkir, 2005), improved skills for work (Singh et al., 2006) and innovation (Davenport & Prusak, 2005). But concrete definitions of knowledge given by KM theorists differ. It is unnecessary to examine them individually. Instead, I categorise them according to the different methods that KM theorists have used to define
knowledge, because the same methods provide fairly similar definitions. The three methods for defining knowledge are the holistic method, the knowledge hierarchy method and the resource-based method.

2.2.1.1 Holistic Method

The first method used to define knowledge is the holistic method, which is an intuitive attempt to see knowledge as a container of many useful components. The components mentioned by KM researchers include truth, belief, perspectives, concepts, judgements, expectations, methodologies, context, reflections, insights, skills, information, memories and values (Wiig, 1993; Spek & Spijkerver 1997; Bender & Fish, 2000). From an empirical perspective, knowledge includes all these components. However, it is hard to evaluate the usefulness of the definitions of the holistic method because they are essentially a sum of important intellectual elements in the human mind. The components used by researchers are based on their own needs, but are also identified in a fairly arbitrary manner. Most KM theorists do not have a well-formed methodology to explain why they include certain aspects of the intellect and exclude others. In addition, these definitions usually do not analyse the relationship between different elements clearly, if at all. We are not expected to develop a solid framework for managing knowledge using this method.

Due to the drawbacks of the holistic method, we need a more logical and rational method to define knowledge. Some theorists have attempted to define knowledge by distinguishing it from its lower forms, namely information and data. Thus, the knowledge hierarchy method was developed.

2.2.1.2 Knowledge Hierarchy Method

A group of scholars define knowledge with the knowledge hierarchy diagram, which ranks data, information and knowledge (Beccerra-Fernandez, 2008; Davenport & Prusak 1998; Hicks et al.,
The basic rationale of the method is to contrast knowledge with its lower forms, i.e. data and information, and sometimes with its higher form, i.e. wisdom (Hicks et al., 2006). A well-known example is Davenport and Prusak’s (1998: p.2-4) definition of knowledge:

**Data** is a set of discrete, objective facts about events. In an organizational context, data is most usefully described as structured records of transactions.

**Information** is meant to change the way the receiver perceives something, to have an impact on his judgment and behaviour.

**Knowledge** is a fluid mix of framed experience, values, context information, and expert insight that provides a framework for evaluating and incorporating new experiences and information.

There are many similar definitions. For example, Long et al. (1998) defines knowledge as ‘information combined with experience context, interpretation and reflection’. Information is often related knowledge that is seen as its higher form. For example, Nonaka and Takeuchi (1995:p.58) try to clarify the relationship between knowledge and information:

1) Knowledge, unlike information, is about beliefs and comments. Knowledge is a function of a particular stance, perspective, or intention.

2) Knowledge, unlike information, is about action. It is always knowledge “to some end”.

3) Knowledge, like information, is about meaning. It is context-specific and relational.

The knowledge hierarchy method provides clearer definitions than the holistic method and can be used on many occasions. However, these definitions are vague in identifying an absolute boundary between information and knowledge, and in explaining what turns information into knowledge. The answers depend on how we understand information. The general idea of the knowledge hierarchy method is flawed, as it clarifies the relationship between different forms of information or knowledge rather than directly stating the unique property of knowledge. Another alternative to the knowledge hierarchy method is the resource-based method.
The third way of defining knowledge is to focus on the functions of knowledge as a type of ‘resource’ (Mentzas, 2003:p.19–20). Nonaka and Takeuchi (1995) argue that knowledge is a factor of production. Dalkir (2005) define knowledge by identifying the functions of knowledge as follows:

'Subjective and valuable information that has been validated and that has been organised into a model (mental model); used to make sense of our world; typically originates from accumulated experience; incorporates perception, beliefs, and values' (p.336)

However, Wiig’s definition of knowledge (1993:p.75) as ‘the principle force that determines and drives the ability to act intelligently’ is more applicable in the context of this study for two reasons.

First, Wiig’s understanding of knowledge is more compatible with the context of translation and potentially with translation memory systems. Both Robinson and Wilss consider knowledge an indispensable factor for rational translation behaviours such as logical reasoning or other uses of language. Another general purpose of TMS is to help translators translate more intelligently with translation suggestions.

Secondly, Wiig (1993) provides a sensible conceptual recognition that knowledge is primarily something that guides people’s behaviours. If knowledge is defined as a certain type of information, then the link between knowledge and human behaviours through skills is ignored. People’s skills or competences are not always easily expressed in language, images or any other media, and are not seen as a form of information.

In addition, Wiig’s (1993) understanding of knowledge also suggests that knowledge can be an asset, considering knowledge the same as physical resources in creating value (Mentzas, 2003:p.19; Nonaka, 1991). The value of knowledge in our context is arguably that of machine-readable resources to improve the performance of translation memory systems.

A proper definition of knowledge enables the review of different typologies of knowledge, KM process and KM approach. The next section will discuss the types of knowledge.
2.2.2 Types of Knowledge

Different typologies categorise knowledge differently, according to the researchers’ understanding of knowledge and settings of tasks. A working definition of knowledge is not equal to the typology of knowledge, even though it is necessary to have a working definition at first. The previous section reviewed numerous definitions of knowledge and adopted Wiig’s definition as a resource that generates the ability to act intelligently. On this basis, I review only the typologies that categorise knowledge from Wiig’s perspective. Thus, only two typologies are discussed below.

2.2.2.1 Wiig’s Types of Knowledge

Wiig (1993) has done comprehensive research into many perspectives of knowledge and offers a typology of knowledge. Wiig categorised knowledge into epistemological and ontological dimensions. Under the epistemological dimension, Wiig identifies three forms of knowledge: public knowledge, shared expertise and personal knowledge (1993:p.144–149). Under the ontological dimension, Wiig defines four types of knowledge: factual, conceptual, expectational and methodological (1993:p.150–152). Dalkir uses a table to illustrate the different kinds of knowledge defined by Wiig:

<table>
<thead>
<tr>
<th>Form of Knowledge</th>
<th>Type of Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>Factual</td>
<td>Conceptual</td>
</tr>
<tr>
<td>Measurement, reading</td>
<td>Stability, balance</td>
</tr>
<tr>
<td>Shared</td>
<td>Expectational</td>
</tr>
<tr>
<td>Forecast analysis</td>
<td>When supply exceeds demand, price drops</td>
</tr>
<tr>
<td>Personal</td>
<td>Methodological</td>
</tr>
<tr>
<td>The “right” color, texture</td>
<td>“Market is hot”</td>
</tr>
<tr>
<td>Company has a good track record</td>
<td>A little water in the mix is okay</td>
</tr>
<tr>
<td>Hunch that the analyst has it wrong</td>
<td>Check for past failures</td>
</tr>
<tr>
<td>What is the recent trend?</td>
<td>Look for temperatures outside the norm</td>
</tr>
</tbody>
</table>

Figure 2.1: Wiig’s Knowledge Typology (Dalkir, 2005:p.65)
The epistemological dimension of knowledge is loosely based on three different contexts of accessing knowledge. Public knowledge is knowledge that comes from sources in the public domain, such as newspapers. Shared expertise is knowledge possessed by specialists and shared within professional communities. Personal knowledge is an individual’s use of common sense.

The ontological dimension of knowledge is based on the level of abstractness. Factual knowledge is data in the traditional sense; for instance, ‘the temperature in London on 18 January 2013 was 2 °C’. Conceptual knowledge is people’s knowledge about concepts or systems, such as their knowledge of the differences between weather and temperature. Expectational knowledge is more abstract and tacit and is about having rational hypotheses or predictions based on one’s expertise. For instance, an experienced adventurer is expected to have clearer expectations before going to the jungle because he or she has better survival skills. The fourth type of knowledge, methodological knowledge, is about decision-making methods and systematic approaches or strategies, which equate more with wisdom in the traditional sense.

As a detailed and comprehensive framework of categorising knowledge, Wiig’s typology is seen by Dalkir (2005:p.65) as pragmatic. Dalkir believes it could be a component of many other future KM studies even though it was proposed in 1993. However, Dalkir also acknowledges that Wiig’s typology might be impractical when it is used in actual KM implementation (2005:p.66), because the typology is too complex.

Wiig’s typology of knowledge has two other serious problems when implemented. First, the knowledge forms and types are sometimes not precisely distinguished in Wiig’s context. Wiig does not consider the factor of people’s capability for processing knowledge. The boundaries between factual, conceptual, expectational and methodological knowledge are not always absolute. This becomes noticeable in the context of using computer-aided translation tools. For example, entries from Wikipedia or Britannica Online Encyclopaedia are available to the public, but their usefulness to translators depends on the translators’ needs, their ability to use information or knowledge, their knowledge background and particular translation problems. One entry may provide different knowledge types simultaneously, such as factual knowledge about terminology and conceptual knowledge clarifying some concepts in specialised texts. Wiig’s typology does not fully explain the complex situations one may face in reality.
The second problem of Wiig’s typology is that it is too complex to use in analysing some real problems. Knowledge is always observed in both form and type. As Figure 2.1 shows, Wiig’s typology proposes 12 kinds of knowledge. But this creates another problem: does a KMS really need to process that many kinds of knowledge, especially when it comes to translation memory systems? Some kinds of knowledge listed in this table also conflict with Wiig’s understanding of knowledge. For example, when factual knowledge is in public form, such as ‘PRC stands for People's Republic of China’, it is not knowledge but data. Although it might be useful, it does not correspond to the same notion of knowledge proposed by Wiig.

These two problems indicate that the typology of knowledge needed in this study should effectively demonstrate the different knowledge types involved in KMS. It should be practical and logical. The explicit knowledge and tacit knowledge highlighted by Nonaka and Takeuchi (1995) will therefore be discussed next.

2.2.2.2 Nonaka and Takeuchi’s Explicit and Tacit Knowledge Dichotomy

Nonaka (1991) presents his earlier vision of the knowledge conversion model based on various cases of innovative Japanese manufacturing companies in the 1980s. Nonaka points out that Western companies were only aware of the importance of explicit knowledge, which is composed of ‘formal and systematic’, ‘quantifiable data, codified procedures, [and] universal principles’ (1991:p.91-93). Explicit knowledge is defined by Dalkir (2005: p.334) as ‘rendered visible (usually through transcription into a document); typically, captured, and codified knowledge’.

Nonaka (1991) regards tacit knowledge as the key to creating knowledge. Tacit knowledge is fundamentally different from explicit knowledge, which corresponds to our common understanding of knowledge. Tacit knowledge is embedded in individual experiences in forms such as insights, intuitions and hunches (1991:p.95). However, Nonaka recognises that ‘tacit knowledge is highly personal. It is hard to formalize and, therefore, difficult to communicate to others’ (1991:p.96).

This study uses the explicit and tacit knowledge dichotomy as its knowledge typology. Hence, the knowledge involved in the TMS workflow can fall into these two categories. Pym (2003) points out that the tacit knowledge used by translators includes the translators’ competences,
such as a set of skills for solving linguistic, cultural, terminological and text-related problems, the ability to locate and evaluate translation-relevant resources, and certain translation strategies and preferences of style for using knowledge. This type of tacit knowledge is embedded in every translator’s mind and is normally acquired by their past experience such as language or translation training sessions or other professional experience. Meanwhile, explicit knowledge is also required on factual issues, translation strategies and techniques pertaining to specific topics to improve their translations. For example, translators should have knowledge about the grammar of both source and target languages. The bilingual aligned translation suggestions is also explicit knowledge, as it is knowledge presented in an accessible form and can be used directly.

2.3 Important Concepts of Knowledge Management

Once we have the knowledge that TMS aims to process, we can develop concrete methods of managing knowledge. To do so, two concepts need to be reviewed: the KM process and the KM approach. The KM process, also called the KM cycle, is a practical model that specifies activities implemented in the practices of KM (Anand & Singh, 2011:p.934–935; Dalkir, 2005:p.25–26). In this study, a theoretical KM model is also included in the KM process, as it may also guide the implementation of KM. The KM process generates the workflow of KMS, while the KM approach is a strategy for managing knowledge (Dalkir 2005:251; Mentzas, 2003:p.4–5, 20).

2.3.1 Knowledge Management Process

The KM process has been well documented in the literature. Eleven influential kinds of KM processes have been identified (Wiig, 1993; Nonaka & Takeuchi, 1995; Meyer & Zack 1996; McElroy, 1999; Bouthillier & Shearer 2002; Bukowitz & Williams, 2000; Lee et al., 2005; Dalkir 2005; Dagnfous & Kah 2006). These processes are different from each other in terms of the different activities involved. Some activities, such as knowledge creation (or acquisition or capture), exist in most KM processes, while activities such as knowledge validation are recognised by only a few KM scholars (McElory, 1999). There is no such thing as the ‘best knowledge management process’. As Probst (1998) suggests, the choice of the KM process
should depend on the degree of relatedness to the context of the study (Probst, 1998:p18). The KM process selected in this study should effectively demonstrate the workflow of KMS for translation Therefore, only two KM processes are potential candidates: Dalkir’s Integrated KM Cycle and Nonaka and Takeuchi’s Knowledge Spiral Model.

2.3.1.1 Dalkir’s Integrated Knowledge Management Cycle

Dalkir synthesises four different practical KM processes into an integrated process he calls the Integrated KM Cycle (2005:p.43). The Integrated KM Cycle contains three stages, as shown in Figure 2.2 below:

![Dalkir’s Integrated KM Cycle](image)

Figure 2.2: Dalkir’s Integrated KM Cycle (Dalkir, 2005:p.78)

1. Knowledge capture and/or creation. The KM process begins with the identification of knowledge and the codification of knowledge, i.e. knowledge is stored in forms that can be easily shared and communicated. Once knowledge has been codified, the next activity is to evaluate the value of the newly collected
knowledge and its relevance to the work or the goals of the organisation (Dalkir, 2005:p.43).

2. Knowledge sharing and dissemination. If newly codified knowledge is valid and sufficiently valuable, the next activity is contextualisation. The knowledge is delivered to the relevant personnel according to the attributes and the content of knowledge and the specific requirements. Another way of contextualising knowledge is to distribute it and to maintain the knowledge of authors or creators, e.g., subject experts, once again. (Dalkir, 2005:p.44).

3. Knowledge acquisition and application. In this step, the knowledge is arguably more accessible. It is hard to generalise the ways of using knowledge due to the heterogeneous situations. Dalkir does not mention discrete activities for this step. She simply reminds us that the knowledge we manage is eventually for the end users, not the organisers of the KM process, and that the KM process does not end at this step. A successful KM process may contribute new knowledge, which is also used for another KM process (Dalkir, 2005:p.44).

Dalkir’s well-balanced model systematically includes the major activities in general KM processes. It also recognises that the knowledge process is a cyclic rather than a sequential workflow. Many earlier KM processes, such as Wiig’s process, consider knowledge application as the last step of the KM process and ignore managed knowledge (particularly in the form of a database) as a possible source of new knowledge.

However, the Integrated KM Cycle is not a silver bullet for all situations. Dalkir’s model of the KM process is rooted in prior studies. The notion of the KM process as a cycle is inherited from Nonaka and Takeuchi’s Knowledge Spiral Model, although Dalkir does not mention it in her textbook. Dalkir is not the first to synthesise different KM processes. Probst coins the term ‘knowledge management building blocks’ to refer to the major phases in the KM process (1998:p.19). Probst’s (1988:p.20–28) KM process has eight KM building blocks: knowledge goals, knowledge measurement, knowledge identification, knowledge use, knowledge acquisition, knowledge development, knowledge distribution and knowledge preservation. Probst explains the activities of each KM building block comprehensively, but does not specify
the internal relations between these KM building blocks. One of the most complicated KM processes is proposed by Bouthullier and Shearer (2002), containing six blocks of activities with 13 interactions. While it might be comprehensive, it also raises the issue of practicability.

The Integrated KM Cycle has been created to demonstrate the concept of the KM process. Dalkir has made a great effort to involve most of the common activities identified in other KM processes, but this makes the Integrated KM Cycle more theoretical than practical. The limitations of the Integrated KM Cycle include the lack of flexibility and the lack of consideration for the use of tacit knowledge. All three steps are equally important in Dalkir’s model. This makes the Integrated KM Cycle not particularly flexible to the situation in TMS workflow. The phase of knowledge sharing and dissemination might be unnecessary, as will be explained in Section 2.3.1.2.

The unsuitability of the Integrated KM Cycle, and its many other similar versions, suggests that the breadth of the KM process is not as important as its practicability and flexibility. The KM process should also consider capturing tacit knowledge, which is crucial for describing the use of knowledge in the TMS workflow. The next Section introduces another KM process that can fulfil the disadvantage of the Integrated KM Cycle used in TMS tasks.

2.3.1.2 Nonaka and Takeuchi’s Knowledge Spiral Model

Nonaka and Takeuchi’s Knowledge Spiral Model (1995) is simple and robust. It is a general model that describes the conversion of knowledge. Nonaka and Takeuchi uses this model to reveal the key to creating new knowledge, by converting tacit knowledge embedded in people’s minds to explicit knowledge owned by others. This model is a four-step knowledge management process (Nonaka & Takeuchi, 1995: p.57):

1) Socialisation: one shares the tacit knowledge with others;
2) Externalisation: tacit knowledge is articulated as explicit knowledge by the individual;
3) Combination: the discrete pieces of explicit knowledge are organised into new systematic and codified knowledge;
4) Internalisation: the formalised explicit knowledge becomes new individuals’ own knowledge, and can also be used as a source for creating new knowledge.
Some important features of the Knowledge Spiral Model make it a framework that can easily describe all activities in the TMS workflow. Its simplicity makes it more flexible to use with other theoretical frameworks and allows it to have technical extensions. Its robustness means that one does not need to follow all the steps presented in this model and that it can easily be modified to new situations. These two features overcome the criticism of its limitations because of its cultural origin in the Japanese manufacturing industry (Glisby & Holden, 2003). Another more important feature of the Knowledge Spiral Model is that it is the first model that describes the knowledge of workers’ mentality and emphasises the relationship between tacit and explicit knowledge. Most KM processes only manage the knowledge that can be systematically recorded and ignore the knowledge that is also closely related to human skills. Although the Knowledge Spiral Model was not originally used as a model of the KM process, it offers a very distinct perspective from other KM processes. The four steps of the Knowledge Spiral Model are also broad enough to cover most activities in the KM process and is therefore, a suitable model for managing the knowledge involved in the TMS workflow.

The advantages of the Knowledge Spiral Model make it particularly useful for describing TMS, which is used for translators working at the individual user level. Therefore, I adopt the Knowledge Spiral Model as a KM process to describe the TMS workflow.
2.3.2 Knowledge Management Approach

The KM approach is the strategy used to complete the KM process. Different KM approaches have different preferences to the types of technology used in the KM process. Information technology has different roles in different KM approaches because KM approaches have different goals and understandings of knowledge related to individuals and to the organisation (Srikantajah & Koenig, 2000). Two generic KM approaches are reviewed in this section: the process approach and the product approach.

The process approach, also called the personalisation or collaboration strategy, emphasises the process of knowing and the social nature of knowledge (Mentzas, 2003:p.24). Mentzas (2003:p.4) describes the central notion of the approach as follows: it ‘mainly understands KM as a social communication process, which can be improved by collaboration and cooperation support tools’. The process approach regards knowledge as deeply embedded in the minds of the people who develop it. The basic principle of this approach is to create an environment or network to share creative ideas. To achieve this, the approach requires the organisation to use E-learning and communication tools such as video-conferencing, e-mail, wikis and workflow management tools (Mentzas, 2003:p. 8). The process approach is successful in consulting companies. For instance, McKinsey relies on a relatively small elite team to provide creative solutions (Mentzas, 2003:p.10).

Despite the success of the process approach in many contexts, it may not be applied in the process of TMS. The translation tasks specified in this study are not creative in terms of giving unique translations or providing innovative translation techniques or methods. The aim of non-literary translation is primarily to render source texts into semantically equivalent target texts (Byrne, 2006:p10). The knowledge in the TMS workflow does not primarily address the social communication process, even though some commercial products offer certain features, such as project management or extracting online resources provided by memoQ. Many non-functional features can be embedded into a translator’s workbench, but they cannot change the nature of TMS as a special form of database system. The primary function of translation memory systems is its retrievable translation units. Other non-functional features, such as the interface of human translators, are important but are not the focus of this study. We need an approach
that aims to improve the efficiency of using knowledge as a resource. Therefore, the more relevant KM approach is the product approach.

The product approach is also called the codification strategy (Hasen et al., 1999:p2). It regards knowledge as something that can be separated from people and can be processed as independent objects, for instance documents (Hasen et al., 1999:p2). The main notion of management in the product approach refers to the activities to capture, store and retrieve knowledge by using KMS or knowledge base, i.e. improving the locating of knowledge (Hasen et al., 1999:p3). The product approach is a computational process, relying heavily on support from computer science, especially knowledge representation and information retrieval (Hasen et al., 1999:p3).

Knowledge in the product approach does not conflict with the definition of knowledge provided by Wiig. Even though knowledge is abstract and indivisible, it always needs to be in a discrete form in order for people to use it, such as when they read a book or use an encyclopaedia. The form or medium can be managed and evaluated in a quantitative or qualitative manner as long as the context is specified. For example, one can say that Mona Baker’s textbook *In Other Words: A Course Book on Translation* (2011) offers a better quality of knowledge than most online ‘translator’ tips if one wants to seriously acquire ability in translation. But knowledge can be managed by managing the medium that contains the knowledge.

### 2.3.3 Knowledge Management Systems

KMS must be clearly defined before describing translation memory systems as a type of KMS. This section examines KMS and the roles of KM technology in KMS. The nature of KMS is dynamic, as the function of KMS depends on the technologies used (Dalkir, 2005:p217-219). Given that the product KM approach centralises database use, Dalkir (2005) defined KMS as:

> 'Centralized databases in which employees enter information about their jobs and from which other employees can seek answers. This system often relies on groupware technologies, which facilitate the exchange of organizational information, but the emphasis is on identifying knowledge sources, knowledge
analysis, and managing the flow of knowledge within an organization—all the while providing access to knowledge stores’ (p.352)

However, this definition cannot be used as the definition of KMS in this study because the ‘knowledge’ managed by KMS for translation purposes is not directly associated with the knowledge we defined in KM. If TMS is viewed as a type of KMS, then the KMS should centralise a knowledge base that contains information from some domains in machine-readable format to assist the translation process. The next section provides an overview of KM technology and its roles in KMS.

2.3.4 Knowledge Management Technology

KM technology enhances the performance of KMS by enabling particular functions of KMS. Dalkir’s categories of KM technologies, based on their roles in each particular phase of the KM process, are as follows: knowledge creation and capture phase, knowledge sharing and dissemination phase and knowledge acquisition and application phase (2005:p.220). The KM process proposed by Dalkir is the Integrated KM Cycle, which has been criticised for its oversimplification in Section 2.2.3.1. Although Dalkir is aware of content management tools, she only mentions Extensible Markup Language (XML) tags and taxonomies as ways of storing structural knowledge or content. Dalkir also makes a more serious omission by failing to suggest retrieval approaches. Koenig and Neveroski point out that the emphasis on effective IR function is crucial for the implementation of KM (2008:p.248). Once the database becomes large, it is impossible to make use of it without proper retrieval tools. LaBrie and St Louis also remind us that even if KMS database has high quality content, the output might be low quality if a KMS does not have a proper mechanism to retrieve that knowledge (2003: p.2552).

An alternative is proposed by Alavi and Leidner (2001), who recognise that a KMS should have four components of technology: knowledge creation, storage/retrieval, transfer and application (p.115). Alavi and Leidner’s idea is more relevant to TMS, because they recognise the storage and retrieval of knowledge as one phase of the process. They also clarify the fact that the system normally processes three types of knowledge: structured information (such as written documents), codified human knowledge, and documented organisational procedures
and processes, as well as tacit knowledge acquired by individuals and networks of individuals (p.118). Unfortunately, Alavi and Leidner do not explain what the technology involves.

As the use of KM technology depends on the structure of KMS, the next section will modify the KM and KMS paradigm in the TMS context.

2.4 Translation Memory Systems as Knowledge Management Systems: A Conceptual Framework

In this section, I illustrate the TMS workflow as a type of KMS. I focus on the most essential function of TMS as indicated by Kay (1980:p.11–16): retrieving the previously stored translation suggestions to avoid repetitive and mechanical work. Therefore, other common features embedded in commercialised translator's workbench, such as translation quality assurance, project management, terminology management tools and machine translation components, are omitted from this study.

2.4.1 Knowledge and Knowledge Types Used in Translation Memory Systems as a Type of Knowledge Management System

Wiig’s understanding of knowledge is most suited to the purpose of this study, because Wiig defines knowledge as the principle force that drives people to act intelligently (1993:p.75). Based on Wiig’s perspective, knowledge is the translator’s competence, capability, proficiency or ability to accomplish translation tasks. However, this is a generic and abstract notion that cannot be used directly to reveal the TMS workflow. Knowledge should be placed in an appropriate typology, namely the dichotomy of tacit and explicit knowledge.

The tacit knowledge involved in the TMS workflow is different from the knowledge required for the translation process. Tacit knowledge, as defined in Section 2.2.2, is knowledge that is hard to express and is internalised by people, and is usually about the process of performing particular skills or demonstrating expertise. Thus, the tacit knowledge in the translator’s mind

---

1 Section 2.4 is based on author’s conference paper Knowledge management issues in the workflow of translation memory systems (Wang, 2013).
is essentially a series of mental activities or mental states. Translation scholars are aware of the tacit knowledge embedded in translators in different forms, such as translation competences and translators’ cognitive systems, to explain the process of translation (Pym, 2003:p482.). The tacit knowledge in the TMS workflow is how translators use translation suggestions. Tacit knowledge is always the knowledge that cannot be shared or used directly by other translators. For this study, we do not need to include every type of tacit knowledge that is manipulated by TMS. The use of tacit knowledge should depend on the explicit knowledge and the knowledge capture process, which converts tacit knowledge into explicit knowledge.

Explicit knowledge in translation (i.e. the third category of knowledge mentioned in Section 2.2) is also complex. It can involve many forms of systematic knowledge from several disciplines. For example, Shei (2005) reports that translators need many kinds of knowledge, such as text analysis skills, translation methods and strategies, understanding readership and translation norms. However, the explicit knowledge manipulated by TMS is fairly simple: translation suggestions in the form of target texts aligned with source texts. Technically, the explicit knowledge that TMS has, is stored mainly in various machine readable forms such as Translation Memory Exchange (TMX), which is an XML-based format (GALA, 2011).

### 2.4.2 The Workflow of Translation Memory Systems as a Type of Knowledge Management System

Translation memory systems as a type of KMS should still fall under into the category of CAT tools, because the translation process is still partly conducted by humans. As mentioned in Section 2.3.1.2, the workflow of translation memory systems as a type of KMS can be analysed using the Knowledge Spiral Model, which focuses on the conversion between tacit knowledge and explicit knowledge (Nonaka & Takeuchi, 1995).

An explanation of the tacit knowledge and explicit knowledge involved in the TMS workflow indicates that users benefit from explicit knowledge, i.e. bilingual aligned translation suggestions retrieved by TMS. Tacit knowledge refers to how translators use translation suggestions. It also suggests that the product KM approach is more applicable than the process KM approach, because translation memory systems as a type of KMS primarily functions as a database rather than a tool to support innovative thinking. Translation memory systems as a
type of KMS is different from the content management system, which primarily aims to store, edit and review documents (Dalkir, 2005:p214). KMS is a holistic system used to complete all the activities required for the KM process.

The Knowledge Spiral Model should be modified when it is used to analyse the use of knowledge in the TMS workflow. As mentioned in Section 2.2.3.1, the ‘knowledge transfer’ step may not exist in the workflow of translation memory systems as a type of KMS. Knowledge transfer refers to the sharing and dissemination of explicit knowledge after it is codified (Anand and Singh 2011:p.935). The scope of this study does not involve all the details of the actual uses of TMS. My focus on the TMS workflow is at the individual level as an IR task. For example, although the exchanges of TM files are an important part of translation project management, this feature is not considered in the workflow in this study. Therefore, the workflow of translation memory systems as a type of KMS can be explained as follows:

**Knowledge Capture**

When translators use TMS, human-produced translations should be seen as tacit knowledge of translation captured by TMS. TMS does process tacit knowledge indirectly, and manipulates translation suggestions that contain the tacit knowledge about translation. Each translation suggestion is formed as a bilingual aligned text fragment that contains tacit knowledge about translation. The tacit knowledge embedded in the newly generated translation suggestion is captured by TMS as it updates TM files.

**Knowledge Codification**

The codification step involves converting the tacit knowledge into explicit knowledge. Once the tacit knowledge is captured, it is codified, which means the newly captured translation suggestion is stored in certain formats such as TMX. By doing so, the tacit knowledge is saved and the structured explicit knowledge can be used.
Knowledge Application

Knowledge application in TMS means the codified explicit knowledge is reused to improve the productivity of translation. The translation suggestions are retrieved as translation suggestions by the TMS according to various similarity measure methods.

Knowledge Creation

Ideally, the KM process can be continued as a mutually beneficial relationship. Translators using the translation suggestions to produce new translations, and the TMS assists translators more effectively as the scale of translation memory knowledge grows.

The TMS workflow is a KM process during which explicit and tacit knowledge is reciprocally converted at every stage, and different categories of knowledge can also be involved. The conversion of different types and categories of knowledge in the TMS workflow is displayed in Figure 2.4 below.

The KM process as set out in the figure below, suggests that as a type of KMS, translation memory systems primarily assists the translation process by retrieving codified explicit knowledge. Translation memory systems employ a successful technical approach to capturing translators’ tacit knowledge, and it presents the explicit knowledge properly in a bilingual parallel form for translators. The technical implementation is relatively easy, as the use of translation suggestions also depends on the translator’s professional proficiency, which makes good use of reference information retrieved by TMS.
The above KM process suggests that as KMS, TMS primarily assists translation by retrieving the codified explicit knowledge. However, TMS does not perform equally effectively in every step of the KM process. Translation memory systems as a type of KMS finds a successful technical approach to capture translators’ tacit knowledge, and it presents the explicit knowledge properly in a bilingual parallel form for translators. The technical implementation is relatively easy, as the use of translation suggestions also depends on the translator’s professional proficiency, which makes good use of referential information.

Translation memory systems as a type of KMS can be severely affected by two technical problems: the small size of TM files and the low efficiency of retrieving translation suggestions (Macken, 2010: p.197). The size of TM files can be increased fairly easily by using bilingual alignment tools (Macken, 2010). But the second problem, which is associated with the KM process of TMS, may require greater effort to ensure improvement.

The low efficiency of using translation suggestions, the explicit knowledge captured by TMS, is mainly caused by the over simplistic technical approach used in knowledge codification and application. In the knowledge codification and application steps, XML-based formats, such as TMX do not store substantive content or other more descriptive features of translation suggestion in the repository of TMS. This is because TMS can only retrieve translation suggestions by the relatively simple similarity measure methods, such as editing distance,
which is essentially based on calculating the number of words in common (Trujillo, 1999:p.64). These methods are employed because they do not require additional resources, such as linguistic knowledge or world knowledge. However, as a result, only key words found in sentences or phrases are used as the basis of retrieval; this is only effective for retrieving highly repetitive texts contained in long-lasting projects (Trujillo 1999:p.64–69). TMS cannot match translation suggestions, which are the same at semantic level but are substantially different in the use of words. TMS encounters the classic ‘vocabulary mismatch problem’ proposed by Furnas in 1987.

Translation memory systems as a type of KMS lacks the capability to process the semantics contained in translation suggestions, particularly the capability to retrieve the required text fragments based on semantic similarity between sentences. TMS as a type of KMS requires a set of technologies to enable this capability. We need to find a method that can store and represent knowledge from some domains in a computable format and can support KMS to compute semantic similarity for the effective retrieval of translation suggestions. Thus, the ESA algorithm, which presents large-scale domain knowledge for measuring semantic similarity, is highlighted as the most crucial factor for TMS to function as KMS.

2.5 Conclusion

In this chapter, the framework of KM was reviewed and used to illustrate the TMS workflow from a conceptual perspective. TMS is a type of KMS that centralises databases and involves a complete KM process. However, TMS currently does not perform the knowledge codification step that well, thereby resulting in low efficiency when it comes to using TM files. We can consider that the improvement of translation memory systems as a type of KMS will lie in improving the capability to retrieve translation suggestions on the basis of semantic similarity. Enabling semantic similarity is an attempt to manage knowledge in order to enhance the KM process of TMS, as such an implementation can only be supported by certain knowledge bases. The next chapter will review relevant semantic similarity measures from a technical perspective.
Chapter 3 Literature Review: Explicit Semantic Analysis and Other Existing Methods for Measuring Semantic Similarity

This chapter first examines the notion of similarity measure and its relation to TMS. It then defines semantic similarity. Based on this definition, different methods of measuring semantic similarity, especially those suitable to sentence-length text fragments, are reviewed. ESA, the proposed technique of measuring semantic similarity, is formally introduced.

3.1 Translation Memory Systems and Similarity Measures

Before discussing semantic similarity and similarity measures, it is necessary to understand similarity and its relation to TMS. Similarity is a complex issue that is discussed in different disciplines such as linguistics, philosophy, information retrieval and artificial intelligence groups (Islam & Inkpen, 2008; Mihalcea & Hassan, 2015). An intuitive approach to similarity might focus on lexical relations, i.e. whether two words are synonyms. Cruse (1986) provided a detailed account of synonymy. Cruse’s notion of ‘absolute synonymy’ corresponds to the type of synonymy that is normally associated with semantic similarity. Cruse’s absolute synonymy was described as follows (1986, p.26):

‘Two words are absolute synonyms if they can be inter-substituted in all possible contexts without changing the meaning.’

No matter how we define similarity, it is essentially related to the focus of this study, which means it should be suitable for IR tasks conducted by Translation memory systems. In general, any IR systems need to retrieve all the texts that are relevant to the user query and avoid retrieving non-relevant documents (Baeza-Yates & Ribeiro-Neto, 2011:p.4). Obviously, a translation memory system performs the task of retrieving relevant translation units in response to queries, i.e. newly inputted source texts. Therefore, the special case of TMS is that user queries and targeted documents (i.e. translation suggestions) are both sentence-length text fragments, which can be seen as a type of short text. The relevance between two sentences is judged by their similarity. Thus, the similarity discussed here is that of the two text fragments to be matched by the IR system. A similarity measure is a method of judging the similarity of two text fragments. Trujillo (1999:p.61) states that similarity measures are essentially heuristic.
and that different similarity measures may have very different results because of the different features of texts that they use. One similarity measure may outperform in one application but is not necessarily suitable for another application. Therefore, it is difficult to find the ‘best’ similarity measure.

In the context of TMS, texts from translation suggestions are processed as a string, i.e. a sequence of characters (Geroimenko, 2004). Levenshtein distance, the most popular edit distance measure adopted for translation memory systems, is more appropriately referred to as a string similarity measure that only employs surface-level features of strings. The similarity defined by the edit distance is essentially based on the degree of overlapping between two strings (Trujillo, 1999:p.63). Normally, the edit distance measure computes the similarity score for two translation suggestions based on the number of keystrokes needed to convert the one to the other in order to indicate the degree of similarity. This is called the fuzzy matching score. End users can set certain threshold values to judge if a stored translation suggestion is relevant or not (SDL, 2014a); by a process of trial and error, most TMSs have settled on a default threshold value of 70%, although this can be changed by the end user.

In fact, threshold values are very likely always be set differently, for example within a range of 30% to 99%, depending on translator’s preferences or the particular features of the TM files and the text being translated. As pointed out by Macken (2010:p.196), the threshold value in the TMS context is more properly seen as a balance between the recall and precision of translation suggestion filtering rather than as a rigid discriminator of the quality of translation suggestions. In other words, the threshold value is not mainly used to identify valid and invalid translation suggestions but is rather a filter to receive the suitable number of potential useful translation suggestions. Lower threshold values are more likely to yield large numbers of translation suggestions but it is not probable that many of them will be useful; while a higher threshold value may potential miss useful translation suggestions (Macken, 2010:p.197). For a TMS, sentences with scores higher than a certain threshold value are not necessarily useful for translators, and useful sentences could also potentially receive a lower score.

The next section introduces the Levenshtein distance, which is the most common method of computing the edit distance.
3.1.1 Levenshtein Distance

The Levenshtein distance was developed in the 1960s and was originally used for the correction of spelling errors (Damerau, 1964). The Levenshtein distance is still used today for its efficiency in computation.

The Levenshtein distance method measures similarity between strings by counting the minimum effort required to transform one string to another (Trujillo, 1999:p.66). The Levenshtein distance counts three types of operation: insertion, deletion and substitution (ibid.). Each operation is assigned a penalty score of 1. The fewer the editing operations, the more similar the two strings are. For instance, the process of calculating the similarity between two strings, ‘bag’ and ‘bug’, is as follows:

First, we need to map the two strings in a table:

<table>
<thead>
<tr>
<th></th>
<th>-</th>
<th>b</th>
<th>u</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: An illustration of the transformation of ‘bag’ to ‘bug’

We get ‘0’ for the second row in the second column from the left, as transforming ‘-’ to ‘-’ do not assign penalty score. But we get a penalty score of ‘1’ for the cell next to it, as it needs to add ‘b’ to transform ‘b’ to ‘-b’, and we have a penalty score of ‘2’ for changing ‘-’ to ‘-bu’. We can fill up this table by following the same procedure. The actual editing score, which is always located at the bottom right-hand cell, is 1. The edit distance is calculated as the edit score divided by the number of characters of the longer string. In this case, it is 1/3 = 0.33.

The default value of string similarity is 1. We need to subtract the value of the edit distance to obtain the string similarity. In this case, the string similarity between ‘bug’ and ‘bag’ is calculated as 1 – edit distance (0.33), and the resulting string similarity is 0.67. This would be displayed as a fuzzy matching score of 67% by translation memory systems.
3.1.2 The Advantages and Disadvantages of the Levenshtein Distance Method

The greatest advantage of the Levenshtein distance method is its simplicity, as it does not require an external knowledge base. This is also why most translation memory systems still use the Levenshtein distance method as a matching method. Its simplicity makes it easily extensible. To date, there are various methods designed to improve the edit distance for translation memory systems. For example, the chunk-based TMS matches shorter segments of translation suggestions, such as phrases, rather than complete sentences. Colominas (2008) and Macken (2010) suggest that the efficiency of the chunk-based TMS in using TM files is normally higher than that of the traditional TMS. Moreover, the current matching method can be also enhanced with paraphrase databases based on greedy approximation and dynamic programming techniques to improve the efficiency of retrieving translation units (Gupta &, 2014).

An obvious disadvantage of the Levenshtein distance method is that it is impossible to measure the similarity of two strings according to other important features such as the semantic meaning or rhetoric features of texts. Therefore, it is still a low-efficiency method of managing TM files. Conversely, as Mitkov and Corpas point out (2008), TMS may not be able to match sentences that are semantically equivalent but are expressed in different words. For example, the pair of sentences ‘Microsoft developed Windows XP’ and ‘Windows XP was developed by Microsoft’ cannot be matched because their similarity score is 43%, which is far below the common fuzzy matching score (70%).

Another important but less obvious drawback of the Levenshtein distance method is its performance in languages other than Western languages. The Levenshtein distance method may be less useful for logographic writing systems such as Chinese as the features of the Chinese writing system make the Levenshtein distance method less readily applicable to translation where one of the languages uses a logographic writing system. The Chinese writing system is based on Chinese characters (zi 字), which are normally morphemes, most Chinese words being formed of two or more characters. The Chinese language lacks grammatical indications, and the meanings are expressed by a combination of characters, word order and contextual information, rather than by morphological means (Zhao et al., 2011). These features may result in low efficiency when using string similarity measures. For example, ‘微软视窗’
and ‘微软窗口’ are both Chinese translations of ‘Microsoft Windows’. They have exactly the same meaning (*Microsoft Windows*), and only one character is different. But their fuzzy matching score, calculated using the Levenshtein distance method, is only 50%. There are several potential problems for Chinese NLP tasks, because of the different writing system. This will be also addressed in Section 6.2.

Measuring translation suggestions based solely on the surface level of strings makes TMS capable of retrieving only highly repetitive texts contained in long-lasting projects. The efficiency of leveraging TM files can be very low in TMS tasks. Thus, it is reasonable for researchers to focus on processing the semantic information contained in translation units.

### 3.2 Measures of Semantic Similarity

Some researchers have attempted to use different semantic process techniques to enhance the performance of translation memory systems. Mitkov and Corpas (2008) state that the third generation of TMS should be able to match translation suggestions in terms of their meaning. To enable this function, Yao (2010) employs ontology-based resources in TMS that can represent the meaning of translation suggestions. Yao (2010) proposes a localisation software model that consists of both universal ontology to describe semantics of texts and specialised ontology to represent concepts from specialised areas. Their TMS-enabling effects can be seen as measuring the semantic similarity of translation suggestions. To represent the meaning of the texts is the first step in measuring semantic similarity.

Budanitsky and Hirst (2006) distinguish semantic similarity from semantic relatedness. They consider semantic relatedness as a more general concept than semantic similarity. For example, *bank* and *trust company* are two similar terms. *Pencil* and *paper* are not similar, but they are semantically related (Budanitsky & Hirst, 2006:p.13-15). Turney (2006) also defines two similar concepts: attributional similarity (semantic similarity) and relational similarity (semantic relatedness). Two words are synonyms when they have a high degree of attributional similarity, whereas two words are analogous when they have a high degree of relational similarity. Although this study needs to focus on the measure of semantic similarity of texts, Jurafsky and Martin (2009:p.687) states that the techniques of computing both semantic
similarity and semantic relatedness can be very similar in terms of the knowledge base they employ, especially in so-called thesaurus methods.

Semantic similarity can be loosely defined as ‘likeness in terms of meaning’ (Trujillo, 1999:p.63). Semantic similarity in this study is about the similarity being measured with reference to certain resources. It is hard to come up with a straightforward definition of ‘semantic similarity’ without defining ‘meaning’. For the purpose of the study, I define 'meaning' as the information from a particular knowledge base. Hence, a semantic similarity measure is a process that represents the meaning of text fragments and then measures the similarity according to particular metrics.

The methods of measuring semantic similarity can be categorised into two general groups: knowledge-based/thesaurus methods and corpus-based methods (Mihalcea et al., 2006). In the first method, two words are semantically similar if they are assigned as synonyms in a thesaurus. In the second method, two text fragments may be seen as semantically similar if they can be used interchangeably in similar contexts (Weeds, 2003:p.13). For example, buy and purchase are semantically similar, because they can be used interchangeably, in the sentence ‘Eventually she had saved enough money to buy a small car.’

3.2.1 Resources for Measuring Semantic Similarity

Thesaurus methods or knowledge-based methods use information from knowledge bases to measure semantic similarity (Mihalcea et al., 2006:p.777). As defined in Chapter 1, knowledge bases are different kinds of resources containing different information. These include lexical knowledge bases, such as WordNet, or knowledge representation of commonsense knowledge, such as Cyc. The commonsense knowledge contains ‘facts, rules, stories, and descriptions’ about the world in order to understand the text written by people (MIT Media Lab, 2012). Although each knowledge base uses different procedures, the basic idea is the same for measuring the similarity of texts or sentences. In particular, the knowledge base is used as a well-structured thesaurus to measure the similarity of individual words from the text; hence the similarity of texts is the aggregation of word similarity (Jurafsky & Martin, 2009:p.687; Liu & Li, 2002; Li et al., 2006). Wu and Palmer propose a refined path-based method with

Some studies have used traditional thesauri, such as Roget’s Thesaurus (Curran, 2003:p.9) and 同义词词林, i.e. A Dictionary of Synonyms (Guan, 2002), to measure word similarity on a semantic level. However, these were early studies that aimed to test the feasibility of thesaurus methods. In this section, I introduce methods for measuring semantic similarity based on three state-of-the-art knowledge bases: WordNet, Open Mind Common Sense (OMCS) and HowNet. These resources are all available online and are free to use for non-commercial purposes. They are also large-scale databases that contain sufficient information for complicated NLP applications.

3.2.1.1 Using WordNet for Measuring Semantic Similarity

WordNet is one of the most popular knowledge bases for measuring semantic similarity (Jurafsky & Martin, 2009:p.680). WordNet was developed in the mid-1980s, and its latest version (version 3.1) was released in 2012(Princeton University).

WordNet is a large lexical knowledge base that groups words according to their meanings. In version 3.0, WordNet had information on 155,287 words (Princeton University, 2010). The ‘synset’, the WordNet term for a set of synonyms, is the basic unit of WordNet. Every synset represents a concept and is linked to other synsets by various relations. For example, ‘bed’ links to the more general synset ‘furniture’ by a super-subordinate relation. There are other relations such as antonymy and meronymy (Miller, 1995:p39-40). Nouns, verbs and adjectives are the main word classes presented in WordNet. It also includes a few adverbs (ibid.). All the synsets form a hierarchical structure. In WordNet, all synsets are placed in a tree structure, where synsets are nodes and each node can also have its own child node (ibid). The links connecting two nodes are called edges (ibid.). A synset may subsume multiple synsets as subordinated concepts. For convenience, I also refer to synsets as concepts in this section. The concepts in WordNet are described by the words they subsume. The more specific a concept is, the fewer subordinated concepts it has. Thus, a very specific concept can be represented as just one word.
Semantic distance is the inverse of semantic similarity (Budanitsky & Hirst, 2006). It is convenient to measure the semantic distance with the structure of WordNet. If two words have longer semantic distance, then the two have lower similarity. Some of the representative methods aim to measure semantic distance rather than directly measure semantic similarity (Jurafsky & Martin, 2009:p.686).

The WordNet-based method of measuring semantic distance can be categorised into two types. The first type of method is path length-based similarity, which measures the path between two nodes in the network of WordNet. Here, the path length is defined as the number of edges that one node travels to another (Jurafsky & Martin, 2009:p.686–687). The basic form of this type of method is the logarithmic transformation of the path length, which was defined formally by Leacock and Chodorow (1998) as follows:

\[ \text{sim}_{\text{path}}(w_1, w_2) = -\log \text{pathlen}(w_1, w_2) \]

In this equation, \( w_1 \) and \( w_2 \) are the two WordNet nodes whose semantic distance will be measured. The smaller the value of \( \text{pathlen}(w_1, w_2) \), the fewer edges are needed to transfer from one node to another, and the larger the value of \( -\log \text{pathlen}(w_1, w_2) \). Therefore, the maximum value of \( \text{sim}_{\text{path}}(w_1, w_2) \) is their semantic similarity.

The Leacock-Chodorow measure can also be further refined by considering the depth of two words in the WordNet thesaurus. The depth (D) is the distance between the top node of the WordNet thesaurus and the node that includes the words. Thus, the formula is refined as follows:

\[ \text{sim}_{\text{path}}(w_1, w_2) = -\log \left( \frac{\text{pathlen}(w_1, w_2)}{2D} \right) \]

However, the synset depth of WordNet is more complex than Leacock and Chodorow assumed. Simpson and Dao (2010) demonstrate that the lower the level of WordNet is, the more nodes this level has, as the following figure shows:
As can be seen in Figure 3.1, higher-level WordNet concepts have more general meanings than lower-level concepts, which tend to have more specific meanings. For example, the synset ‘truck’ is placed at the bottom level of the thesaurus because it has only one specific meaning. In contrast, its parent concept ‘automotive’ is a more general concept; hence it is placed at an upper level of the thesaurus. Likewise, the top-level synsets have broader meanings than their child synsets, i.e. hyponyms.

The depth of concepts in the WordNet hierarchy is also important when computing semantic distance. The semantic distances between ‘automotive vs. bike’ and ‘instrumentality vs. article’ are both 2. However, considering their depths in the WordNet hierarchy, the semantic distance between the ‘instrumentality/article’ pair should be larger than that of the ‘automotive/bike’ pair.

The second type of WordNet-based method, the information content word similarity method, was developed to take into account the depths of synsets. The information content word similarity method uses probabilistic information to reflect the hierarchy structure of WordNet (Jurafsky & Martin, 2009:p.688). This method measures the extent of information shared by two words (Resnik, 1995). The theoretical foundation of the information content similarity method is based on information theory. Information content is derived from an information-theoretic concept, self-information (Cover & Thomas, 2006:p.13). The information content (IC) of an event $x$ is the amount of information, which is quantified as follows (Cover & Thomas, 2006:p.14):
IC(x) = \log \frac{1}{p(x)} = -\log p(x)

\(p(x)\) is the probability of event \(x\) occurring.\(^2\) In this study, event \(x\) denotes a word \(x\), or more appropriately, a concept \(x\) used in WordNet. A synset can be seen as a concept that may include subordinated concepts. A word can belong to a synset. A general concept can have many words. By contrast, a specific concept can have only a few words or even just one word. Resnik (1995:p.3) computes the information content of a concept \(x\) in two steps. A fragment of the WordNet hierarchy is used to illustrate this process:

![Figure 3.2: A fragment of WordNet hierarchy (Jurafsky & Martin, 2009:p.689)](image)

If we want to obtain the information content of ‘geological-formation’, a large corpus must first be provided. We denote the number of words used in both the corpus and in WordNet as \(N\). The concept ‘geological-formation’, according to figure 3.2, has four sub-concepts: natural-elevation, hill, shore and coast. We need to count the total number of their occurrences in the corpus. Therefore, the probability of encountering the concept ‘geological-formation’ in WordNet, or \(p(c)\), can be calculated as follows (Jurafsky & Martin, 2009:p.449):

\[
p(x) = \frac{\sum_{w \in \text{words}(c)} \text{counting}(w)}{N}
\]

\(^2\) In Probability theory, a simple event is defined as ‘the outcome that is observed on a single repetition of the experiment’, and an event is a collection of simple events (Mendenhall 2009:p.129). An example of an experiment is tossing a coin and observing which side appears. There are two simple events: 1) heads denoted as \(E_1\), 2) tails denoted as \(E_2\). According to the function of IC(x), the value of information content of event \(E_1\) is 1 bit.
Where \( c \) is the concept ‘geological-formation’, and \( \sum \text{counting}(w) \) is the total number of occurrences of the concepts of geological-formation, natural-elevation, hill, shore and coast.

Resnik (1995:p.2) defines the information content (IC) as follows:

\[
\text{IC}(c) = -\log p(c)
\]

By following the same procedure, we can calculate the information content of all concepts of the selected fragment of WordNet. Each concept can be assigned a probabilistic value, as shown in Figure 3.3:

![Figure 3.3: A fragment of WordNet hierarchy with trained probabilistic values](Jurafsky & Martin, 2009:p.689)

The values of information content, i.e. probabilistic values can be seen as the inverse of the cost of identifying a word to be subsumed under a concept. General concepts with higher probabilistic values indicate that they are placed in higher positions of the WordNet hierarchy, and are ‘obviously’ to be found. Therefore, it is relatively ‘cheap’ and easy to identify a word belonging to a very general concept such as ‘entity’ (0.395). On the other hand, it is more difficult to identify a word belonging to a more specific concept such as ‘coast’ (0.0000216).

To measure the similarity between two concepts or two words, Resnik (1995) uses the concept ‘lowest common subsumer’ (LCS). The LCS is the lowest concept in the WordNet thesaurus that subsumes two subordinated concepts; therefore, it is the hypernym of two concepts (Resnik, 1995:p.1). According to the Figure 3.3, the LCS of ‘natural elevation’ and ‘shore’ is the synset ‘geological-formation’. Resnik (1995:p.3) defines the similarity between concepts \( c_1 \) and \( c_2 \) as follows:
\[ \text{sim}_{\text{res}}(c_1, c_2) = -\log p(LCS(c_1, c_2)) \]

The information content of the LCS is used to measure the similarity between two concepts or words because the LCS is the lowest level of nodes that cover the two concepts. Obviously, two different concepts would require a broader LCS located on a relatively higher level of WordNet hierarchy to cover them both. On the contrary, two concepts with closer semantic distance in the WordNet taxonomy require a lower level of LCS, which gives a relatively small value of \( p(c) \) and a larger value of \( IC(c) \), i.e. the two concepts share a bigger portion of information content.

Various methods are based on Resnik’s notion of LCS. Lin (1998) and Jiang and Conrath (1997) both propose two influential measures. Like Resnik, Lin (1998) states that the similarity between two concepts depends on the extent of information they have in common. Lin (1998:p.298) defines similarity theorem as follows:

‘The similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.’

Thus, this theorem is formalised as follows (ibid.):

\[
\text{sim}_{\text{Lin}}(A, B) = \frac{\text{common}(A, B)}{\text{description}(A, B)} = \frac{2 \times \log p(LCS(A, B))}{\log p(A) + \log p(B)}
\]

Jiang and Conrath (1997) have the same understanding of similarity as Lin does, but they have developed the following semantic distance measure:

\[
\text{sim}_{\text{JC}}(A, B) = 2 \times \log p(LCS(A, B)) - (\log p(A) + \log p(B))
\]

The different definitions were used in experiments designed to perform different tasks, as different methods were suitable for different types of task (Jurafsky & Martin, 2009:p.700; Curran, 2003).
3.2.1.2 Using HowNet for Measuring Semantic Similarity

HowNet is another online knowledge base that contains machine-readable semantics about Chinese concepts and their English equivalents. HowNet was developed in 1988 for machine translation and became available online in 2008 (Dong et al., 2010). Dong Zhendong, the main creator of HowNet, describes it as follows:

‘On-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalent’ (Dong & Dong, 2002)

The structure of HowNet is different from that of WordNet. As discussed in Section 3.1.2 above, the Chinese language uses Chinese characters rather than an alphabet as the basic unit of writing. HowNet needs a different structure to provide formalised information to describe each concept, which is represented by a Chinese word. HowNet employs ‘sememe’ as its basic component to construct its knowledge base (Dong et al., 2010:p.53). A sememe is defined as the smallest semantic unit, which cannot be decomposed further (ibid.). Every Chinese word, i.e. concept, in HowNet is described by a set of sememes. The sememes are manually selected by HowNet developers and defined based on the meanings of 6,500 Chinese words. The current version of HowNet has 700 sememes, which are used to describe about 30,000 concepts in Chinese words.

HowNet not only defines lexical relations between different sememes, but also defines more than ten semantic relations, which are annotated with special mark-up symbols. For example, ‘*’ stands for instrument-event relations, ‘%’ stands for ‘part-whole’ relations and ‘#’ stands for concept co-relations (Liu & Li, 2002:p.64-65). The semantics of the Chinese word 手表 (watch) can hence be represented as follows:
Table 3.2: A HowNet entry of ‘手表/watch’ (Dong & Dong, p.2002: p.24)

Table 3.2 shows that ‘NO.’ is the ID number of the word ‘手表’. ‘W_C’ is the Chinese word and ‘W_E’ is its English translation. ‘DEF’ is the semantics of ‘手表’, which is tagged by three sememes. They indicate that ‘手表’ is a tool with the instrumental function of ‘tell’ and is about ‘time’.

There are different methods of using HowNet to measure semantic similarity (Dai et al., 2008; Xia et al., 2011; Wu et al., 2012). Obviously, the most direct intuitive method is to count the number of sememes that two HowNet concept words have in common. However, Liu and Li (2002) criticise this intuitive method for not fully employing the semantics presented in HowNet. Some other researchers have proposed their own HowNet-based similarity measures, but most of them are similar to Liu and Li’s method. Therefore, I focus on the method developed by Liu and Li (2002). They state that four aspects of a HowNet entry can be used:

- **Primary sememe:** It is the first sememe used to describe a HowNet concept. For example, ‘tool’ is the primary sememe for the word ‘手表’. The primary sememe is extracted from other sememes because the primary sememe is the most important sememe to define the semantic meaning of a HowNet concept.

- **Other independent sememes:** These are the sememes other than the primary sememe. Most HowNet concepts have more than one sememe.

- **Relational sememes:** They have two classes of sememes, which are ‘EventRole|动态角色’ and ‘EventFeatures|动态属性’. They are mainly used to indicate the grammatical case relations between different sememes and concepts.

- **Symbols:** Liu and Li also take the marked up symbols of sememes as indicators of the semantic meaning.

The measures of Liu and Li (ibid.) are only applicable to the semantic similarity between content words including nouns, verbs and adjectives. Although functional words are also included in HowNet, they are only tagged by relational sememes and have less semantic
meaning in real-world NLP applications (Dong & Dong, 2002). Therefore, Liu and Li’s method excludes the semantic similarity of function words.

Liu and Li (2002:p.72) hence defines the semantic similarity as follows:

\[
\text{sim}_{\text{HowNet}}(C1, C2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{l} \text{sim}_j(C1, C2)
\]

The above formula shows that \( C1 \) and \( C2 \) are two HowNet concepts. \( \text{Sim}_1 \) stands for the similarity of the primary sememe. \( \text{Sim}_2 \) is the similarity of other independent sememes, \( \text{Sim}_3 \) is the similarity of relational sememes and \( \text{Sim}_4 \) is the similarity of symbols. Each \( \text{Sim}_i \) is calculated by the following function (Liu & Li, 2002:p.69):

\[
\text{sim}(p_1, p_2) = \frac{\alpha}{d + \alpha}
\]

where \( p_1 \) and \( p_2 \) are both sememes, \( d \) is the semantic distance between two sememes in the HowNet structure and \( \alpha \) is an adjustable constant.

\( \beta \), as Liu explained, is also an adjustable constant, which should satisfy two conditions:

1) \( \beta_1 + \beta_2 + \beta_3 + \beta_4 = 1 \)

2) \( \beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4 \)

This reflects the different importance of \( \text{Sim}_i \). The weighting scheme demonstrates that \( \text{Sim}_1 \) has the heaviest weight, but the other three aspects of a HowNet concept are also considered.
3.2.1.3 Using ConceptNet for Measuring Semantic Similarity

ConceptNet is a free online crowdsourcing commonsense knowledge base developed by the MIT Media Lab. ConceptNet is described by its developers as a ‘large-scale commonsense knowledge base with an integrated natural-language-processing tool-kit’ (Liu & Singh, 2004a:p.212). The commonsense knowledge stored in ConceptNet is defined as ‘a huge portion of human experience, encompassing knowledge about the spatial, physical, social, temporal, and psychological aspects of typical everyday life (Liu & Singh, 2004a:p.211)’. Examples of commonsense knowledge are our experiences such as ‘lemon is sour’ and ‘if you forget someone’s birthday, they may be unhappy about it’. Unlike some knowledge bases that are maintained by knowledge engineers, the information of ConceptNet, called the Open Mind Common Sense (OMCS) corpus, is mainly contributed by online volunteers. The OMCS corpus is populated by various sources such as online digital pet games. Game players may contribute their commonsense knowledge to the OMCS corpus simultaneously by teaching their pets (MIT Media Lab, 2012). The latest version is ConceptNet 5, which has more than one million pieces of commonsense knowledge. ConceptNet became multilingual in its third generation (Havasi, 2007). Now, ConceptNet 5 supports languages such as Japanese, Dutch, Portuguese, Chinese, French, Arabic, German (MIT Media Lab, 2012.). The Chinese ConceptNet is built by a research group from National Taiwan University. It is not a translation of the English ConceptNet, but builds its own database following the same procedure of the original ConceptNet (Hsu, 2013).

In ConceptNet, commonsense knowledge is formed as a statement or description of a fact in one sentence. Templates can be filled with statements, which are formalised as different sentence patterns. Developers manually summarise about 90 sentence patterns based on texts submitted from contributors (Havasi et al., 2009:p.23). The sentence patterns are transformed into about 20 semantic relations in machine-readable formats, as can be seen in Figure 3.4 below.
Similar to WordNet, ConceptNet is a semantic network with a graph structure (Speer, 2012). Nodes, the statements in ConceptNet, are the basic elements of ConceptNet. The nodes are linked via edges to reflect their semantic relations. Below is an example of a fragment of ConceptNet (cf. Figure 3.5):

Figure 3.4: Semantic relations in ConceptNet\(^3\) (Havasi et al., 2009:p.24)

<table>
<thead>
<tr>
<th>Relation</th>
<th>Example sentence pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>IsA</td>
<td>NP is a kind of NP.</td>
</tr>
<tr>
<td>UsedFor</td>
<td>NP is used for VP.</td>
</tr>
<tr>
<td>HasA</td>
<td>NP has NP.</td>
</tr>
<tr>
<td>CapableOf</td>
<td>NP can VP.</td>
</tr>
<tr>
<td>Desires</td>
<td>NP wants to VP.</td>
</tr>
<tr>
<td>CreatedBy</td>
<td>You make NP by VP.</td>
</tr>
<tr>
<td>PartOf</td>
<td>NP is part of NP.</td>
</tr>
<tr>
<td>HasProperty</td>
<td>NP is AP.</td>
</tr>
<tr>
<td>Causes</td>
<td>The effect of NP</td>
</tr>
<tr>
<td>MadeOf</td>
<td>NP is made of NP.</td>
</tr>
<tr>
<td>AtLocation</td>
<td>Somewhere NP can be is NP.</td>
</tr>
<tr>
<td>DefinedAs</td>
<td>NP is defined as NP.</td>
</tr>
<tr>
<td>SymbolOf</td>
<td>NP represents NP.</td>
</tr>
<tr>
<td>ReceivesAction</td>
<td>NP can be VP (passive).</td>
</tr>
<tr>
<td>HasPrerequisite</td>
<td>Before you VP, you must VP.</td>
</tr>
<tr>
<td>MotivatedByGoal</td>
<td>You would VP because you want to VP.</td>
</tr>
<tr>
<td>CausesDesire</td>
<td>NP would make you want to VP.</td>
</tr>
<tr>
<td>HasSubevent</td>
<td>One of the things you do when you VP is NP</td>
</tr>
<tr>
<td>HasFirstSubevent</td>
<td>The first thing you do when you VP is NP</td>
</tr>
<tr>
<td>HasLastSubevent</td>
<td>The last thing you do when you VP is NP</td>
</tr>
</tbody>
</table>

---

\(^3\) where NP stands for noun phrase and VP for verb phrases.
ConceptNet has more diverse semantic relations than WordNet, as ConceptNet specifically focuses on the representation of commonsense knowledge rather than knowledge about language. ConceptNet should be considered a reasoning tool kit rather than a lexical knowledge base. WordNet essentially measures similarity by the lexical relations suggested by its thesaurus structure, while the semantic similarity measure is a form of reasoning in the use of ConceptNet (Liu & Singh, 2004a:p.220). There is more than one way to use ConceptNet to measure semantic similarities between text fragments, but the AnalogySpace technique is currently the most effective method recognised by ConceptNet’s developers because of its computational effectiveness in using a large-scale knowledge base. As a reasoning technique, AnalogySpace mainly relies on singular value decomposition, which is widely used in other corpus-based semantic similarity measures such as latent semantic analysis (LSA).
In contrast to the methods of using WordNet and HowNet, AnalogySpace needs to first present two concepts in a matrix that indicates the features, i.e. their semantic relations to other concepts. For example, the concepts ‘cat’ and ‘dog’ are presented as follows:

<table>
<thead>
<tr>
<th></th>
<th>IsA(_, pet)</th>
<th>AtLocation(_, home)</th>
<th>CapableOf(_, fly)</th>
<th>MadeOf(_, metal)</th>
<th>PartOf(fur, _)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>9</td>
<td>5</td>
<td>-3</td>
<td>-6</td>
<td>5</td>
</tr>
<tr>
<td>dog</td>
<td>12</td>
<td>?</td>
<td>-5</td>
<td>?</td>
<td>8</td>
</tr>
<tr>
<td>airplane</td>
<td>4</td>
<td>-5</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>toaster</td>
<td>?</td>
<td>6</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 3.3: Matrix presentation of ConceptNet concepts (Hsu, 2013)

The columns in Table 3.3 are the occurrences of the assertions for that concept in ConceptNet. The value of ‘IsA (\(\rightarrow\), pet)’ is ‘9’ for ‘cat’ and ‘12’ for ‘dog’, which means that the assertion ‘cat is pet’ is stated nine times and the assertion ‘dog is pet’ is stated 12 times. In general, a higher value means the assertion is more likely to be true. It is also positive to have a negative value. The ‘CapableOf(\(\rightarrow\),fly)’ column assigns ‘-3’ for cat, meaning the assertion ‘CapableOf(\(\rightarrow\),fly)’ is negated thrice for cat. For example, ConceptNet may receive statements, such as 'cat can't fly', 'I don't think cat can fly', 'I can't believe a cat can fly', and also categories them as incidents of assertion ‘CapableOf(_,fly)’, although the incidents negated the assertion. Most ConceptNet relations are affirmative, so ConceptNet has to assign negative value to reduce the reliability of this assertion (Hsu, 2013). Moreover, the ‘?’ mark means no record for that feature.

The basic idea behind measuring the semantic similarity between two concept vectors is to calculate their common features.\(^4\) The table shown above forms an AnalogySpace matrix, which defines ‘cat’ and ‘dog’ as concept vectors:

\[
\overrightarrow{\text{cat}} = (9, 5, -3, -6, 5) \\
\overrightarrow{\text{dog}} = (12, 0, -5, 0, 8)
\]

\(^4\) Vector is a type of measurable quantity that has magnitude and direction. A vector is presented by a collection of numbers (Baeza-Yates, 2011:p.71).
The vector \textit{cat} has five elements. In linear algebra, we call it a 5-dimension vector. These two concept vectors can be mapped in the 5-dimension vector space (Speer et al., 2009). However, ConceptNet is large and sparse. This means a vector normally contains many features, i.e. many elements, so it can be a high-dimensional vector (Speer et al., 2009). The sparseness means the vector has many nulls. In short, AnalogySpace measures the similarity between two concepts by measuring their representation in a vector space. Hence, the semantic similarity measures are transformed as a linear algebra problem that can be solved by many mathematical techniques. For example, singular value decomposition is employed for a more effective computation by extracting feature values (Speer et al., 2009). A more detailed discussion of these mathematical methods will be provided in Section 3.2.2 on corpus-based methods, which rely more on a statistical approach to capture the information from a large corpus.

3.2.1.4 Limitations of Thesaurus Methods

This section discusses three semantic resources and their corresponding semantic similarity measure methods. However, these methods have not been used in real world applications, and hence few researchers have compared performances of different methods for TMS tasks. A similarity measure may outperform others in one application but may be unsuitable for another application. Thus, it is difficult to have comparable evaluations of these semantic similarity measures without testing them in the same application. However, we can still predict some limitations of using thesaurus as a semantic resource.

First, the information of a thesaurus is not always efficient for real-world applications. Although all semantic resources are large, they often lack domain-specific knowledge. WordNet and HowNet cannot be frequently updated. Although it is possible to merge ConceptNet with other domain-specific knowledge bases, the process would be complex and time-consuming (Kuo & Hsu, 2012). This drawback indicates that these resources may not have sufficient knowledge for processing texts from specialised domains. Therefore, the uses in TMS tasks are limited.

Second, although thesaurus methods share a more or less similar paradigm, the actual implementation is largely determined by the structure of these resources. For example, WordNet can only compute the similarity of words in the same word class. This feature creates
a huge problem for processing foreign languages such as Chinese, which has a different system of word classes. For example, the tagging of words in a phrase according to their part of speech can be particularly ambiguous in the case of Chinese. In Chinese, the word class is largely dependent on context, and virtually none of the words have morphological changes. The language property of Chinese raises serious unresolved changes for a lexical-based approach (Zhan, 2010). The problems of processing the Chinese language will be discussed in Chapter 6.

The limitations of thesaurus methods suggest the need for more flexible methods that can be extended and are not constrained by the structure of the semantic resources. We may review another type of method in the next section, because of these limitations.

### 3.2.2 Corpus-Based Methods

Corpus-based methods do not require either well-structured ontologies or a semantic thesaurus to measure semantic similarity; rather, they employ the information derived from corpora and then estimate the semantic similarity of text fragments (Islam & Inkpen, 2008). Although most corpora do not provide information on semantics directly, they can provide the contextual information or distributional information of words. For example, apple and pear can be considered as having similar meanings, as they may both be associated with other words like 'eat', 'sweet', 'crunchy'. The relation between 'apple' and 'sweet', and 'crunchy' is called co-occurrence. The similarity between ‘apple’ and ‘pear’ is based on their co-occurrence features within a corpus. In general, these methods follow the distributional hypothesis suggested by Firth (1957:p.11): *You shall know a word by the company it keeps*. Some methods, such as LSA (Latent Semantic Analysis), may use different models to define how semantic concepts are related to words based on the distributional hypothesis as well.

Obviously, corpus-based methods depend on the size of the corpora. Nevertheless, co-occurrence can also be defined on the basis of many different aspects which are also derived from different methods. When the size of a corpus becomes very large, some necessary computing techniques are also applied to improve the efficiency of using the language resources.
3.2.2.1 Point-wise Mutual Information

Turney (2001) proposes the point-wise mutual information method which measures semantic similarity based on the degree of association between two words in terms of the possibility of co-occurrence. Point-wise mutual information is an information-theoretic concept which refers to the certainty of event $x$ occurring if event $y$ also occurs (Cover & Thomas, 2006:p16). The point-wise mutual information value $I$ of event $X$ and $Y$ is then defined as:

$$I(X;Y) = \sum_{x,y} (x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

We can take events $x$ and $y$ as words that co-occur. This formula can be applied to measuring the semantic similarity of two words on the basis of the possibility of the two words co-occurring in the same context. Turney applies this point-wise mutual information formula in the context of semantic similarity when a large corpus $N$ is given. Turney (2001:p.4) defines the point-wise mutual information method as follows:

$$assocPMI(x, y) = \log \frac{p(x,y) \times N}{p(x)p(y)}$$

Turney’s point-wise mutual information method is evaluated to answer the TOEFL (Test of English as a Foreign Language) Synonym Test. TOEFL is an English-language test for non-native English language speakers to qualify to enrol in North American universities. The use of this method yielded 72.5% accuracy (Turney, 2001:p.8). Although the point-wise mutual information method generally produces good results in synonym tests, this method is not found in other applications. The reason for this is that the point-wise mutual information method only considers the probabilistic co-occurrence of two words, rather than employing semantic information. The next section will focus on some more popular methods.

3.2.2.2 Latent Semantic Analysis and its Variants

Latent Semantic Analysis (LSA), sometimes called Latent Semantic Indexing (LSI), is the predominant method for representing texts as vectors from a semantic space, which is a mathematical representation of concepts constructed from any selected large corpus
(Deerwester, 1990). LSA was introduced by Deerwester to resolve the problem of using synonyms as query terms for IR tasks (1990). Recently, LSA has been applied in various different areas, from cross-language IR (Berry & Young 1995; Celebi et al., 2009) to cognitive psychology and cognitive linguistics (Gui, 2003; Landauer & Dumais, 2007). Although LSA is not a state-of-the-art method for measuring semantic similarity, its importance lies in that it has inspired many other methods, such as probabilistic latent semantic analysis, hyperspace analogue to language, and latent dirichlet allocation. The application of LSA in TMS tasks has also been patented for commercial use (Chen et al., 2006).

Implementations of LSA are rather complicated and vary from application to application. However, the basic notion of LSA is consistent and can be summarised as follows: LSA hypothesises that a concept is always described by a similar set of words (Deerwester, 1990). The words in this set may be semantically similar, but their semantic relations with regard to a concept are usually unseen, because people cannot be aware of every incident in which the relevant words appear, and these words are not always used in the same context (for example, in the same document). Such relationships can only be formed when a large amount of information is provided by a corpus. LSA fundamentally aims at clustering the words about the same concepts contained in a large corpus. These concepts are ‘latent’ in terms of two aspects: (1) they are unseen by humans and (2) they are not concepts which have been defined by humans, in contrast to explicit concepts (for example, entries in an encyclopaedia). The ‘latent’ concepts are generated ad hoc from the chosen corpus (Cimiano et al., 2009). This also means that the ‘latent’ concepts can be changed with those in a different corpus.

Implementations of LSA rely on singular value decomposition, a mathematical technique for accelerating the process of identifying unseen semantic relations. The corpus should be very large in order to have sufficient features to find the semantic relations between words. LSA first needs to transfer a given corpus into a term-document matrix $T$.

Inevitably, we may have a high-dimensional and spare matrix. This means that many cells of the matrix are null. The reason for this is simply that one document contains a limited number of terms, and these terms are unlikely to appear in all other documents. It is difficult to find a conceptually related term from a high-dimensional and spare matrix. However, singular value decomposition can transfer the matrix $T$ to a relatively smaller and more informative matrix by decomposing it into three independent matrices---$U$, $S$, $D^T$---as follows:
\[ T = U \cdot S \cdot D^T \]

Singular value decomposition reconstructs matrix \( T \) after a series of complex calculations are performed. We can get an approximated matrix \( T' \) which is more informative for finding latent semantic relations. This matrix \( T' \) thus forms a vector space, also called semantic space, which represents all terms and their semantic relations as vectors (Deerwester, 1990:p.400). Words or documents can first be transformed into vectors of the semantic space formed by matrix \( T' \) and their similarity can be measured to different metrics, such as the cosine similarity metric.

Nevertheless, various methods are influenced by LSA. Similar to LSA, the Hyperspace Analogue to Language model employs the co-occurrence information from a large corpus and represents the word co-occurrences in a vector space (Lund & Burgess, 1996). The similarity of each pair of words then becomes the probability that appears in an \( N \) sized window, and can also be measured by certain vector similarity metrics (Lund & Burgess, 1996:p204).

Probabilistic Latent Semantic Analysis by Hofmann (1999) and its variation, latent dirichlet allocation by Blei et al. (2003), can be seen as the improved versions of LSA. Probabilistic latent semantic analysis simplifies the computing processes of LSA by involving various probabilistic models that describe the mutual relations between words, concepts, and documents (Hofmann, 1999:p.1-2). Consequently, probabilistic latent semantic analysis can identify how many latent concepts are related to a word through the word’s documents, which are also related to those of latent concepts. Thus, probabilistic latent semantic analysis can recognise polysemy words, that is, words that have two or more meanings, such as ‘bank’ can means 1) an organisation where people can save and borrow money, 2) “sloping raised land, especially along the sides of a river”(Hofmann, 1999:p.4). Latent dirichlet allocation improves probabilistic latent semantic analysis by employing a Bayesian model (Blei et al., 2003). Latent dirichlet allocation models document-latent semantics and latent semantics-term relations, and also employs certain techniques to reduce the complexity of computation (ibid.).
3.2.2.3 Vector Similarity Metrics

Many corpus-based methods may represent documents as vectors, and the similarity between two documents is defined by their distances in the vector space (Jurafsky & Martin, 2010:p.697). There are various metrics for measuring the distances.

Manhattan distance and Euclidean distance metrics are the two simplest metrics for measuring the lengths of vectors in 2-dimensional vector spaces (Jurafsky & Martin, 2009:p.698). Consequently, this means that two metrics are only useful for measuring the similarity between two words, and they ignore another important property of vectors, namely, the direction in the vector space. Curran (2003, p.73) suggests three vector similarity metrics: cosine, Dice, and Jaccard for measuring vectors that are applicable in high dimensional vector spaces:

\[
\text{sim}_{\text{Dice}}(\overline{A}, \overline{B}) = \frac{\sum_{i=1}^{N} \min(A_i, B_i)}{\sum_{i=1}^{N} \max(A_i + B_i)}
\]

\[
\text{sim}_{\text{Jaccard}}(\overline{A}, \overline{B}) = \frac{\sum_{i=1}^{N} \min(A_i, B_i)}{\sum_{i=1}^{N} \max(A_i, B_i)}
\]

Accordingly, the cosine metric is the most popular metric for defining the distance between vectors. A cosine metric, as its name suggests, measures the cosine values between vectors. Calculating a cosine metric in a high-dimensional vector is based on the dot product of two vectors, defined as follows:

\[
\overrightarrow{A} \cdot \overrightarrow{B} = |\overrightarrow{A}| \cdot |\overrightarrow{B}| \cos \theta
\]

Thus the cosine value is:

\[
\cos \theta = \frac{\overrightarrow{A} \cdot \overrightarrow{B}}{|\overrightarrow{A}| \cdot |\overrightarrow{B}|}
\]

The value of \(\cos \theta\) is the similarity between two vectors, and is calculated as follows (Baeza-Yates, 2011:p.78-79):
If two vectors have the same direction, their cosine value is 1, indicating that two documents are identical. If two vectors are orthogonal (i.e. the angle between the two vectors is 90°), their cosine value is 0, meaning that the two documents are completely different. Therefore, using the cosine metric assigns the range of semantic similarity from 0 to 1.

### 3.2.2.4 Limitations of Corpus-Based Methods

The corpus-based methods discussed in Section 3.2.2 have not been tested in translation memory systems tasks; therefore, it is difficult to carry out a comparable evaluation of their performance. The accuracy of LSA is 68.4% in paragraph-length similarity tests, while achieving 69.7% for precision rate and 95.2% for recall rate (Mihalcea et al., 2006). However, we can still outline limitations of corpus-based methods from their performance in some similar experiments which aim at measuring semantic similarity.

Although LSA and its relevant methods are able to identify both synonyms and polysemy, the corpus-based methods do not employ the meaning of text fragments, but do employ their contextual information from a large corpus. In other words, with LSA, the knowledge behind terms in texts is not fully used. The semantic similarity is solely based on contextual relations, rather than the semantics of the text.

### 3.2.3 Evaluation of Semantic Similarity Measures

Many methods are used to evaluate semantic similarity measures. Most of these are technology-based evaluations of 'the performance of the underlying technology components' (Hirschman & Mani, 2005: p. 414). As demonstrated in Sections 3.2.1 and 3.2.2, semantic similarity measures are rarely used as independent functions. For example, semantic similarity
measures can be used as matching components of translation memory systems. It is difficult to access end-users to directly evaluate output of semantic similarity measures in a completed system. The technology-based evaluations compare the results of similarity measures against human performance or other resources. Usually, these approaches can be categorised into three types: evaluation against human judgments, evaluation against gold standards and application-based evaluations.

The most direct method to evaluate a semantic similarity measure is to compare its results with that of human judgement. In this procedure, a collection that contains different pairs of words is prepared. For example, Rubenstein and Goodenough's (1965) collection contained 65 word pairs. The similarities of the word pairs were ascertained by similarity and human judgments, and the results were then correlated. Evaluation against human judgement is used in the research of many similarity methods (Resnik, 1995). However, the use of this evaluation method must include careful consideration of the subjectivity of human judgement. Obviously, different people may rate the semantic similarities of words differently. Therefore, the use of this method must be specified in the context of the application of a semantic similarity measure.

Another method is evaluation against 'gold standards', which are collections of resources that present 'correct' lexical relations. Most of these resources are manually created. In the application of semantic similarity measure, WordNet and Roget's Thesaurus are usually considered the gold standards because of their good reputations. The implementation of this method is basically the same as evaluation against human judgement. The results of a similarity measure are compared with content of a gold standard; for instance, sets of synonyms from Roget's Thesaurus can be used to cross check collections of semantically similar word pairs (Grefenstette, 1994). However, the use of gold standards is often limited, because it is generally difficult to obtain appropriate standards for specific tasks.

The third method includes the evaluation of semantic similarities in specific applications rather than the direct evaluation of semantic similarity measures (Weeds, 2003: p.40). As explained, the semantic similarity measure is unlikely to become an independent application but usually remains a component of a larger system. For measuring the semantic similarity of words, applications such as word-sense disambiguation and vocabulary tests are often used to determine whether a similarity measure improves the performance of an application. In this study, application-based evaluation was used because ESA is seen as a semantic similarity to improve TMS tasks. A software platform was included to evaluate if ESA could profitably be
used in TMS tasks. Additional details, particularly concerning the placement of ESA in a translation memory system, will be discussed in Chapter 4.

An understanding of ESA is necessary before discussing how to evaluate its use in TMS tasks. Section 3.3 comprehensively introduces ESA for this purpose.

### 3.3 Explicit Semantic Analysis

Hitherto, two kinds of semantic similarity methods have been reviewed. This section reviews different aspects of ESA. Each one has its own advantages and disadvantages. ESA is a corpus-based method which inherits many features from LSA. However, ESA more closely focuses on the use of knowledge recognised by humans. Coming back to ESA, ESA represents any text fragments in terms of Wikipedia articles, which is knowledge contributed and edited by online volunteers. The main notion of ESA can be described as follows: a text fragment corresponds to one concept in Wikipedia. Each concept is described by a Wikipedia page. In other words, the content of a Wikipedia page defines the concept. Those concepts are ‘explicit’ because they have been defined in a human generated encyclopaedia (Gabrilovich, 2009:p.444). By contrast, the ‘latent’ concepts generated by LSA are collections of semantically related words clustered from a large scale corpus (ibid.). ESA can be seen as a method which is supported by both mathematical notions from corpus-based methods and by a semi-structured knowledge base.

Hence, many Wikipedia pages constitute a collection of many concepts. These Wikipedia pages can then be transformed into a matrix, where each column indicates a concept and each row corresponds to a word used in the set of Wikipedia pages. Therefore, the size of the matrix is determined by the number of Wikipedia pages. Inevitably, the matrix is usually large for real world applications. For example, Chen and Long (2011) employ ESA in an experimental application of automatic abstraction for medical articles. The matrix they used is 5847×9612 in size (Chen & Long, 2011:p.184). Currently, English Wikipedia has more than 4 million articles (Wikipedia, 2013). Using the full size of Wikipedia would create a very large matrix. After applying a particular weighting scheme, e.g. TF-IDF, this matrix is the semantic space of Wikipedia concepts (Gabrilovich, 2009:p.445). Similarly to LSA, text fragments can therefore be mapped into this semantic space as weighted vectors. The semantic similarity between two
text fragments can then be measured according to similarity metrics, such as the cosine metric as mentioned in Section 3.2.2.3 (Gabrilovich, 2009:p.446).

ESA was developed by Gabrilovich, who aims at employing knowledge of the world for information retrieval or text categorisation tasks (2006, 2007, and 2009). Gabrilovich realises the feasibility of employing Wikipedia—arguably the largest collaboratively edited and multilingual online encyclopaedia—as a knowledge base for NLP applications. ESA has become a very popular method which is used in many information retrieval-related applications, such as cross-language IR (Sorg & Cimiano, 2012; Cimiano et al., 2009; Potthast et al., 2008; Sorg & Cimiano, 2010), query expansions (Luo et al., 2012), knowledge discovery (Yan & Jin, 2012), word sense disambiguation (Turdakov, & Velikhov, 2008), and automatic abstraction (Chen & Long, 2011). ESA in this study is expected to be the component for measuring the semantic similarity of translation suggestions of translated memory files which are fostered by several test collections.

3.3.1 Wikipedia

ESA uses Wikipedia as the knowledge base for measuring semantic similarity. The following three sections briefly review its history, presenting some relevant research about Wikipedia and introducing the technical structure of this collaboratively edited online encyclopaedia.

3.3.1.1 Brief History of Wikipedia

Wikipedia is a wiki-based website. A wiki allows users to create and collaboratively edit interlinked web pages using a simplified mark-up language via a web browser (Wikipedia, 2014b); it is usually maintained and updated by users and does not require registration (Albors, 2008). The technological component of Wikipedia is powered by a wiki engine (or wiki software).

A wiki engine can be regarded as a content management system which has many applications, such as building a KMS or note services. Wikis were created and named after a Hawaiian word for ‘fast’ by Ward Cunningham in 1994 (Stvillia, Smith & Gasser, 2008). According to

---

5 Section 3.3.1 is based on the author’s MSc Dissertation (Wang, 2011).
Cunningham, the idea of a wiki was originally from HyperCard which is an application that combines the functions of databases with user-friendly interfaces and was released in 1987 for Macintosh computers (ibid.). Users can record the information as a form of ‘cards’ in their local files. HyperCard became one of the first hypertext systems used to manage the data stored in local drives (in the form of a card, which is the file page that constituted the database). Cunningham identified a stack of three functions of the HyperCard—‘cards for ideas; cards for people who hold ideas, cards for projects where people share ideas’—and then applied these notions to the first online wiki—WikiWikiWeb—in 1994. ‘[A] community-oriented hypertext system to support collaborative writing’ has been the central concept of wiki-based websites (Vigneault, 2008). The concept of wikis gained support from a small research community in information technology and software development, but the real popularity of wikis was gained through Wikipedia.

After the first unsuccessful attempt of Nupedia which adopted some commercial models (such as having paid-employees), James Wales launched the Wikipedia project in 2001. The influence of Wikipedia was initially limited to the English-speaking community, with 20,000 entries in the first year of creation. The first Chinese entry was produced in October 2002 (Zhang, 2006: p.13). But three years later, English entries increased by ten times and entries in 50 languages became available (Lih, 2004:p.2-4).

Wikipedia has already established several mechanisms and policies to ensure the quality of its contents and to avoid vandalism. The fundamental principles of the operations of Wikipedia are summarised as ‘Five Pillars’, which state that Wikipedia as an online encyclopaedia aims at presenting structured information written from a neutral point of view and edited collaboratively (Wikipedia, 2014f; Wikipedia, 2014g). Wikipedia has inherited Cunningham’s original idea of a collaborative writing model. However, it is not merely a technological application of a wiki, but it also features its own socio-technical system. This is probably the reason why it has gained popularity across the world. Wikipedia is used by ESA because it is not only an encyclopaedia, but also a knowledge base with an accessible and simple structure. The next section introduces the structure of the Wikipedia website.
3.3.1.2 The Structure of Wikipedia

Each Wikipedia entry can be regarded as a statement or explanation of a concept or proposition, especially in the fields of the natural sciences, which usually possess rigid predetermined knowledge structure.

An entry based on a highly general concept can be very long and contain a great deal of information that must be divided into several entries. As a result, the Wikipedia community has reached a consensus to limit the length of entries (Soinellis & Lourids, 2008:p.70). Some entries are well structured (or have much longer textual length) if more users are interested in the topics. Such entries become directories of relevant information. For example, in the relevant entry 'oxygen' can be described from many points of view, its allotropes, physical properties, or in relation to living organisms.
Figure 3.6 Extract from the Wikipedia article on 'oxygen'

Figure 3.6 shows that the entry 'oxygen' offers hyperlinks to many relevant entries, such as those on allotropes, liquid oxygen, solid oxygen, isotopes of oxygen, silicate minerals and oxide minerals. The article contains many hyperlinks to relevant articles; the article would be extremely long if it contained all of this information.

However, users do not normally use the 'category' page, which serves as an index of the entry's domain knowledge and can quickly reveal the knowledge structure of the entry's domains. There are two types of category pages: one is 'subcategories', which presents the structure of the entry's knowledge domain by showing the branches of information derived from the entry;
the other is pages in 'category'. As the name implies, this page lists the article's hyperlinks. A general entry with many related concepts may have a great number of subcategories and 'pages in category' (Figure 3.7).

Figure 3.7: Category page for 'oxygen'
However, a highly specific concept falls at the bottom of this categorical hierarchy, and has few or no other pages in its category or subcategory.

Figure 3.8: A basic proposition of 'ozone': It may have fewer pages and no subcategory of its own.
Every page shown above has its own uniform resource locator (URL). Notably, the syntax of URLs is highly regulated. For example, the URL of the entry ‘XML’ in English is ‘http://en.wikipedia.org/wiki/XML’, while the Chinese version is ‘http://zh.wikipedia.org/wiki/XML’.

This regularity is also applied in category pages. The URL of category ‘XML’ in English is ‘http://en.wikipedia.org/wiki/Category:XML’, while the Chinese version is ‘http://zh.wikipedia.org/wiki/Category:XML’. Hence, the regularity of syntax variation is very easy to be tracked.

If we go behind the interface of the Wikipedia website, we can find that Wikipedia uses a standard template to present the information for each entry. The features of its HTML (hypertext markup language) code are also highly uniform. For example, the number of related entries (i.e., hyperlinks for relevant entries) is featured in syntax as:

```html
<a href="http://en.wikipedia.org/wiki/[name of entry]" title="[name of entry]" >[name of entry]</a>
```

So, it is possible to find and record the number of the HTML code that corresponds to the syntax feature above, to show the relatedness of one entry in a particular domain. This provides a reliable application programming interface for machine processing if a researcher wants to know the entire structure of Wikipedia.

Wikipedia is also structured to be multilingual. Most entries are collaboratively contributed by users from various language communities, while there is occasionally a translated version of another language entry. Users can use the column at the left-hand side of the page showing the language availability, which directs them to other language versions.
The language bar links are also called inter-language links, interwiki links or interwikis (Wikipedia, 2014a).

3.3.1.3 Research on Wikipedia

The quality of Wikipedia articles is one of the most popular issues in the academic study of Wikipedia. The most-cited study on measuring Wikipedia is undoubtedly Nature’s manual evaluation of its content quality when comparing Wikipedia with Britannica Encyclopaedia (Giles, 2005a). Forty two entries were examined by relevant experts to check for three types of mistake: factual errors, critical omissions and misleading statements. The result was 162 mistakes found in Wikipedia and 123 mistakes in Britannica, and each of them had four serious mistakes. This suggests that the quality of traditionally well high-quality reference books may not be so much better than the new collaborative Internet encyclopaedia.

However, the evidence that Nature presents does have a few drawbacks. First, the whole evaluation was conducted manually. With fewer than 50 articles examined, it is not a convincing statement, when the total number of entries in the English version by that time was 751,666; only 0.00588% of these entries were examined (Giles, 2005a). Second, a strong deductive trend dominates the whole of this research. Nature selected the entries in a totally random manner, but it does not set up a proper taxonomy of the entries selected, and
consequently makes it difficult for others to build further analyses on this survey (Giles, 2005b:p.2). For example, we cannot discover the distribution of errors among different topics, or any collocation of factors with the topics, etc.

A more careful study is Stvillia, Smith and Gasser’s (2008) attempt to examine Wikipedia’s organizational structure for information quality assurance. Stvillia et al. (2008:p.983) are aware that Wikipedia is a ‘large-scale, continuously evolving, open collaborative content creating system’, so its content is strongly linked to how users ensure Wikipedia’s quality. Stvillia shows that researchers would do well to look beyond Wikipedia’s contents, to also focus on other factors that can influence content. After an empirical survey of Wikipedia’s managerial architecture, it is shown that the articles in Wikipedia are only part of the focus. In fact, many other aspects, such as the discussion page and the system of deletion, can also have great impact on Wikipedia (2008:p.987). However, Stvillia’s study does not have clearly defined variables, and it also suffers from of the limited nature of its data, as approximately 120 articles were selected (2008: p.996). As a result, we may only discover a general overview of the content quality of Wikipedia’s articles. This suggests that an automated process based on a reproducible scientific method is essential for studying Wikipedia.

There are other Wikipedia-focused studies based on experimental methods, e.g., Halavais and Lackaff’s (2009) analysis of topic coverage in Wikipedia, and Soinellis and Lourids’s (2008) study of why a sustainable growth of Wikipedia has been achieved.

Halavais and Lackaff propose two methods for measurement: 1) Present the distribution of topics in Wikipedia according to the distribution of books published and the classification system employed by the Library of Congress (LC); and 2) Compare the distribution of topics in reputed encyclopaedias to Wikipedia’s entries. Due to having a larger scale of samples (around 3000), several significant findings are revealed: Wikipedia arguably reflects the general interests of its contributors, such as that Wikipedia has the advantage a lot of information on naval, music and military topics; whereas books in concentrate more on literature, and are particularly strong in expert-areas, i.e., law and medicine (2008:p.432). While Halavais and Lackaff find that Wikipedia’s coverage of topics is also driven by the interests of contributors, articles of Wikipedia in any particular domain still outnumbered those in encyclopaedia that are written by subject-matter experts (2008:p.436). However, the limitations of Halavais’ research are two-pronged: first, Halavais and Lackaff do not provide any evidence showing which topics especially interested contributors, as this is hard to measure. This can be
especially problematic when researchers try to narrow their work down to more specific domains of topics. For example, in the realm of linguistics, how would we judge whether semantics or pragmatics is more popular? Second, the classification system (i.e., Library of Congress system) that Halavais and Lackaff used remains questionable. Halavais and Lackaff do not show that the comparison with printed materials cannot solve many problems of online knowledge processing, such as multi-category topics.

Another direction of research is to use Wikipedia as a source for particular purposes. The process common to these kinds of study is first to analyse the potential for achieving the stated purpose, then to break the possibilities down into two pairs of measurable dimensions, and then convert these measurable dimensions into statistical data that can be computed quantitatively. The results are analysed and discussed within a particular framework. Lih (2004) evaluates Wikipedia as a reliable journalistic source. In Lih’s research, the reliability of this journalistic source is considered in the form of ‘reputation’, which is based on the metadata, i.e., edit history. Consequently, ‘reputation’ is evaluated by two metrics: Rigour (total number of edits for an article) and Diversity (total number of unique users). Nguyen et al (2009) investigate Wikipedia’s potential as a translation resource for Cross-Linguistic Information Retrieval (CLIR). Nguyen’s study regards Wikipedia articles as representations of concepts (i.e. units of knowledge) and the translation query is to retrieve a translation from another language through cross-linguistic links (2009:p.58). Its performance is judged by averages of query terms successfully mapped to the database of Wikipedia (2009:p.62). Despite having different focuses, the ideas of rationalising the research methods of using Wikipedia can be universally applied.

Wikipedia is sometimes used to complement other knowledge bases. The BabelNet project employs Wikipedia content to annotate its synsets (Navigli & Ponzetto, 2010). As mentioned in Section 3.3.1, Wikipedia is multilingual, so one Wikipedia page may be available in many languages. Navigli and Ponzetto develop a complex technique to map the inter-language links of Wikipedia and particular WordNet concepts (Navigli & Ponzetto, 2010). Consequently, BabelNet is a semantic network with linguistic information for words, but also encyclopaedic information (i.e., brief explanations of a word according to Wikipedia). BabelNet can be a useful aid to MT. For example, the query 'oxygen' in BabelNet shows multiple meanings of the word, ranging from a chemical element to an album by Avalon, a film by Richard Shepard, and a Linux software project. Each meaning provides information, such as POS (part of speech),
related concepts, brief explanations, and categorical information, from Wikipedia. Some of the information is multilingual if available in this form on Wikipedia or other sources. Some of this is available as multilingual information in this form on Wikipedia.

Figure 3.10: An example entry for 'oxygen' in BabelNet
Another of the BabelNet project's applications is 'named entity recognition' (NER), which is an application that '…labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names' (The Stanford Natural Language Processing Group, 2013). With the aid of BabelNet, MT systems may identify named entities within a sentence. For example, if a sentence contains *Oxygen* as a film tile, it should not be translated in the usual sense (氧气 in Chinese) but as the title (极速杀阵).

### 3.3.1.4 Wikipedia Dumps

Wikipedia allows users to read its articles as well as to download its content. Although the structure of Wikipedia is open, the encyclopaedia does not encourage users to copy its content using web-crawlers because these may slow down the site's speed (Wikipedia, 2014b)\(^6\). Alternatively, Wikipedia makes its database dump available for downloading. A database

---

\(^6\) The Wikipedia articles referenced in this thesis were based on information available on February 2014, unless specified otherwise.
dump is '…a major output of data that can help users to either back up or duplicate a database' (Janssen, 2014). There are many types of dumps, as Wikipedia includes article pages, images, history revisions, user comments and conversations and so on. Most of these dumps are saved in compressed XML files for the ease of importing to the MySQL database (Wikipedia, 2014b).

The complete Wikipedia dump is extremely large. By February 2014, the site's article dump was about 9.8 GB compressed (44 GB uncompressed) (Wikipedia, 2014b). It exceeded the ability of processing for personal computers; therefore, the author accessed http://dumps.wikimedia.org/ to locate smaller files. Detailed information is provided in Section 4.2.

3.3.2 Using Wikipedia as a Knowledge Base to Measure Semantic Similarity

ESA is not the first attempt to employ Wikipedia to measure semantic similarity. Adfre and Rijke have conducted research on identifying similar sentence pairs across different language versions of Wikipedia (2005). Their research method was to machine translate any given Wikipedia page in one language and then count the number of words that overlapped in the same Wikipedia page in the other language (Adfre & Rijke, 2005:p.64). This over-simplistic identification method nevertheless offers fast processing, but does not precisely judge the similarity between two sentences in different languages from in Wikipedia. Their method did not consider the actual quality of the machine translation, and also omitted a possible fact that similar sentences could be found on another Wikipedia page. Adfre and Rijke’s method is essentially a ‘bag of words’ method which represents text fragments as collections of unordered words (Sorg, 2009:p.37). Obviously, Adfre and Rijke’s method does not directly use any semantic information stored in Wikipedia but rely on overlap of words on two pages. Another more reliable method is WikiRate! developed by Strube and Ponzetto (2006). WikiRate! measures two words $w_1$ and $w_2$ on the basis of the similarity between two Wikipedia pages $p_1$ and $p_2$. $w_1$ and $w_2$ must appear in the titles of $p_1$ and $p_2$, respectively. WikiRate! is designed only to measure the semantic similarity of words, and the semantic similarity is associated with path distance in Wikipedia’s hierarchy (Strube & Ponzetto, 2006: p.1419-1420). This method considers the structure of Wikipedia, and is therefore superior to Adfre and Rijke’s method. However, it does not fully employ the content of the Wikipedia page.
3.3.3 Formalisation of Explicit Semantic Analysis

ESA is a method for representing the meaning of any text fragments with respect to an external knowledge base, and the similarity between two text fragments is measured according to their associations with the content of that external knowledge base. This notion was first implemented by Gabrilovich, who used Wikipedia as the external knowledge base (Gabrilovich, 2006 and 2009). Gabrilovich's implementation can be formalised to illustrate how the ESA algorithm works:

We select a knowledge domain of Wikipedia $Wiki_k$, where $|Wiki_k|$ is the number of articles in that domain. Each Wikipedia article $a$ from $Wiki_k$ is defined as an indexing vector, and the number of indexing vectors is $m$. $Wiki_k$ is therefore a vector space, which is a collection of indexing vectors. Accordingly, $|Wiki_k| = m$. The vector space $Wiki_k$ can be seen as the semantic space of the domain $K$.

The terms in $Wiki_k$ constitute a set $T_n = \{t_1, t_2, ..., t_n\}$. The size of the vector space is $n \times m$. As suggested in Section 3.3.1, it is conventional to choose a large domain of knowledge for the effectiveness of its semantic measures; thus, the value $m$ is expected to be very large. On this basis, $Wiki_k$ is normally a high-dimensional matrix of the vector representation of domain $K$.

The matrix is defined as follows:

$$\mathbf{R}^{[Wiki_k]}$$

The relations between a given text $D$ and knowledge domain $K$, is the text mapped into this vector space $\mathbf{R}^{[Wiki_k]}$. It is a transformation, processed by function $\Phi$ as:

$$\Phi_k : D \rightarrow \mathbf{R}^{[Wiki_k]}$$
However, we must carefully define the association between the terms and the Wikipedia articles. In order to obtain accurate information, the TF.IDF weighting scheme is applied to quantify the association between terms $t_i$ and $a_j$ as function $as(t_i, a_j)$:

$$as(t_i, a_j) = tf.idf_{wi}(t_i)$$

The vectors of the Wikipedia article in $Wiki_k$ are defined as index vectors $V_m = \{v_1, v_2, ..., v_m\}$. Gabrilovich defines the value of $V$ as the aggregate value of its term association, as defined below:

$$v_i = \sum_{t_i \in T} as(t_i, a_j)$$

Thus, any text fragment—whether it is a word or a text—can be mapped as a vector by $\Phi_k$ from the semantic space $Wiki_k$. The mapped vectors are called ESA vectors. The semantic similarity of two text fragments can be measured as the cosine value between their two ESA vectors, as discussed in Section 3.2.2.3.

### 3.3.4 ESA as a Generalised Vector Space Model

ESA is not an entirely new method, but is linked to the vector space model. Gottron argues that the original ESA method proposed by Gabrilovich ignores a property of the Wikipedia index vectors, i.e. $V_m$ (2011:p.1962). Gabrilovich and Markovitch recognise that each Wikipedia page focuses on one concept; hence, this is in fact based on the hypothesis that all Wikipedia pages are not semantically related to each other. Mathematically, Gabrilovich and Markovitch do
facto suggest that vectors from $V_m$ are pairwise orthogonal. Gottron et al. (2011) refer to it as a concept hypothesis.

My interpretation is that a concept hypothesis is a simplification of our understanding of an encyclopaedia or any knowledge database. For example, using the Routledge Encyclopedia of Translation Studies, we may find that the entry computer-aided translation is very similar to machine translation, but totally different from poetry. It is not very difficult to see that the concept hypothesis does not hold in the case of the Routledge Encyclopedia of Translation Studies, unless the editor would like to make every entry very distinctive. This condition is also applicable to the use of Wikipedia, as the Wikipedia articles in a certain domain are also correlated. In order to reflect the most realistic condition for using domain knowledge, Gottron reforms the ESA according to the generalised vector space model (GVSM) proposed by Wong et al. (1985) to resolve the issue of the correlation of index vectors.

The essence of GVSM is to consider the correlation of index vectors as a factor of building the semantic space (1985). Wong et al. achieve this to present the index vectors $V_m$ in a $2^m$ dimensional matrix which indicates the full possibility of index vectors in the case where they are omitted because two of them are not pairwise orthogonal. On that basis, Gottron et al. (2011) argue that ESA can be a variation of the GVSM model, and define the similarity measure as follows:

$$\text{sim}_{ESA}(\vec{a}, \vec{b}) = \frac{\sum_{j=1}^{m} \sum_{k=1}^{m} as(t_j, \vec{a}) \cdot as(t_k, \vec{b}) \cdot R(t_j, t_k)}{\sum_{i=1}^{n} (a_i)^2 \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

The function $R$ indicates the correlation of term sets $t_j$ and $t_k$ which belong to Wikipedia index vectors $\vec{a}$ and $\vec{b}$, respectively. The computing function $R$ is defined differently according to the understanding of term interdependence (Gottron et al., 2011). However, for some practical reasons, ESA is not normally implemented as GVSM suggested.

When two vectors are pairwise orthogonal, the angle between two vectors is 90° in a multi-dimensional vector space), meaning $V_{jm} \cdot V_{jm} = 0$, which is proved by linear algebra (Baeza-Yates, 2011:p.70; Anderka, 2009).
Although GVSM is a well-recognised theoretical model for dealing with the correlation of index vectors, it is also a highly complex computing process, as its function, \( \text{sim}_{\text{ESA}}(\vec{a}, \vec{b}) \), reveals. Baeza-Yates reminds us that the use of GVSM can have a very low level of efficiency. Gottron’s representation of ESA vectors requires a set of index vectors which can be a significant proportion of a whole matrix, because a concept is related to a set of semantically correlated vectors (Baeza-Yates, 2011:p.99-101).

In fact, the performance of ESA is not affected by holding a concept hypothesis. ESA is tested to compute the semantic relatedness of texts and words against human judgement. The performance is assessed by Spearman’s rank-order correlation coefficient (denoted as \( r_s \)): \( r_s = 0.56 \) to 0.75 in individual words tasks, and \( r_s = 0.6 \) to 0.72 in text similarity tasks (Gabrilovich & Markovitch, 2007:p.1610). Anderka and Stein (2009) also suggest that the correlation is significantly improved by enlarging the document collection for indexing. When the document collection is enlarged, for example, to more than 10,000 documents, the correlation between ESA and human judgements of text similarity tasks is 0.784 to 0.802 in Pearson’s correlation coefficient (2009:p.670). Applying the notion of GVSM, the performance of ESA is not significantly improved under the same index collection size. The Pearson’s correlation coefficient of GVSM-refined ESA achieves 0.798 in the tasks measuring text similarity (Gottron et al., 2011:p.1964). In addition, Gottron also reports the performances of ESA vary according to different topic domains. In general, the performance of ESA in technical and scientific domains is better than that in arts and humanities domains (ibid.). Thus, the experimental results demonstrate that ESA is a practical and usable method for measuring semantic similarity, and performance associates with the size and content of the index collection.

### 3.3.5 Implementations of Explicit Semantic Analysis

Gabrilovich has not released the source code for his implementation of ESA, written in Perl (2007). However, he encourages developers to implement ESA. Currently, the ESA algorithm

---

8 Both Spearman’s rank-order correlation and Pearson’s correlation coefficient measure the strength of the correlation between two variables by giving values between 1 and -1. 1 indicates that two variables are positive correlation, and -1 means that two variables are negative correlation. 0 means two variables are not correlated. Spearman’s rank-order correlation calculates ranked valued for Pearson’s correlation coefficient (Mendenhall 2009: p.533-6 and p.660-4).
is written and available for different programming languages, including C++ (Jacobs, 2009), Python (Calli, 2007), and Java (Sorg, 2010; Zilka, 2012). The flexibility of using the codes varies. Some developers only release pre-built packages allowing us to perform other tasks based on the semantic space created by the developers.

For the purpose of this study, Sorg’s implementation is selected because of its flexibility in modifying the ESA algorithm. It is based on Terrier, a Java-based information retrieval platform. Sorg’s implementation will be modified to simulate TMS tasks for the purpose of the study. The details of the implementation for this study are presented in the next chapter.

3.4 Conclusion

This chapter reviewed semantic similarity and different semantic similarity measures. The understanding of semantic similarity depends on the requirement of the application. Measuring the semantic similarity of texts means finding a way to represent the knowledge from text fragments, so that the knowledge representations can be measured in different metrics. Both Thesaurus-based methods and corpus-based methods require a knowledge base to perform. Thesaurus/ knowledge-based methods represent the meaning of the texts with respect to highly formalised knowledge bases. However, thesaurus-based methods are always limited by their pre-existing structures and contents. A corpus-based method may not require formalised knowledge bases to represent the meaning of text fragments, but are able to estimate the relations between texts and their semantics according to various statistical techniques. Considering the advantages and disadvantages of both types of methods, I propose ESA as a semantic measure of translation memory systems. However, we can only know how the method performs in its actual application. The next chapter will provide my implementation of the research to evaluate the use of ESA in TMS tasks.
Chapter 4 Methodology

This chapter presents the methodology for evaluating ESA in TMS tasks. The first part of the chapter describes the experiments and poses the research questions in the broader context of the evaluation methods. A potential benefit of ESA is to enable translation memory systems to output a wider variety of translation suggestions. ESA, as a semantic similarity measure, can be used in the common string similarity measures used by commercial translation memory systems and enable translation memory systems to retrieve translation suggestions on the basis of semantic similarity instead of string similarity. Because of the relatively complex nature of the methodology, considerable detail concerning tools, test collections and experimental procedures for the evaluation, is given in the second part of the chapter.

4.1 Overview of the Evaluation Methodology

As indicated in Chapter 1, this study aims to carry out a comprehensive evaluation of the possibility of using ESA in TMS tasks. It is hoped that the evaluation will demonstrate both the potential benefits, and the limitations of using ESA in TMS tasks. Section 4.1.1 provides a description of the experiments that have been carried out, including how an IR platform can simulate the function of TMS and the basic procedures of the experiments. It also presents information about the test collections used. Section 4.1.2 poses the research questions. Lastly, Section 4.1.3 summarises the evaluation method.

4.1.1 Description of the Evaluation Method

The methodology in this study uses a range of experiments to evaluate the performance of ESA in TMS tasks. The experiments are based on various research questions that should reveal how ESA performs in the case of different genres and how textual factors related to a particular genre may affect ESA performance. For the purposes of this study, genre is defined as a style of texts that have different purposes, contents and levels of readability (Perelman et al., 1996).
For example, news articles, technical reports and popular scientific articles are considered separate genres in this study. As defined in Chapter 1, textual factors are linguistic features that can be directly identified by readers; for example, the number of words in a sentence.

The experiments require an ESA IR platform that can simulate ESA-enhanced TMS processes for different translation tasks and several genres. Consequently, the results of the experiments are twofold:

1) An assessment of ESA performance in different genre groups and how the results compare with the Levenshtein distance method is conducted.

2) An analysis of how textual factors within different genres affect ESA performance is carried out.

An ESA IR platform developed by Sorg's team (2010) was used to simulate the pseudo-TMS task in which a query is matched with sentences that have not been translated. To this end, test collections of different genre groups will be used. Test collections usually consist of many documents and are defined as 'standard data sets used to measure the effectiveness of information retrieval systems' (Lewis, 2004). Their relevance to this study is described below.

In this study, the pseudo-TMS task comprises the process of matching a query and sentences from a document collection (i.e., source text segments of translation suggestions), while ignoring whether the texts are actually translated in the target language as the act of matching source and target sentences represents an entirely separate operation. In the TMS scenario, the ESA IR platform is able to retrieve information from a given test collection on the basis of the inputted queries, and use the sentences retrieved as translation suggestions. Thus, the test collection is conceptually parallel to TM files that offer translation suggestions, and queries serve as the texts for translation. Every computation of an ESA similarity score is the same as the TMS process, whereby a query is matched to potential translation suggestions from TM files. The ESA IR platform was adjusted to implement the traversal method of computing ESA similarity scores for a given query with all the sentences in a selected test collection. In this study, the potential queries were manually selected from selected sources similar to corresponding genre groups. Although it can be a time-consuming process, it is more realistic to use texts from the real world rather than invented texts to simulate how translation memory
systems process text that translators may need assistance with. More details on the selection of potential queries are given in Section 4.2.2.

An ESA similarity score is assigned by the ESA IR platform to identify whether a sentence can be used as a translation suggestion, since sentences with very low ESA similarity scores are not considered. One of the aims is to identify the threshold value of ESA similarity scores for valid translation suggestions. A result consists of a pair comprising a query and a sentence matched from the test collection. A valid translation suggestion will be one that includes a sentence, clause or phrase that has at least one component (e.g. clause or phrase) that is similar or identical to one contained in the corresponding query or is potentially useful to the translation process of the corresponding query.

In the experiments the ESA similarity scores of two English sentences are computed. Although it is technically possible to compute the semantic similarity of bilingually aligned texts (Sorg & Cimiano, 2012), monolingual computation is more relevant to TMS tasks. It is clear that the TMS task is a form of the monolingual IR task, and more specifically, short text IR, because most queries are likely to be sentences, not paragraphs. A translation suggestion has a text segment in both the source and target texts. A translation memory system matches only the information in the source language, while its corresponding target text can be used as a translation suggestion. Therefore, the use of either a bilingual ESA implementation or a bilingually aligned corpus is not necessary.

The contents of test collections also need to be representative to cover a wide range of topics from a genre group. Three large test collections, consisting of sets of aircraft accident reports, *Scientific American* articles and Reuters news agency articles, is used as sources in this evaluation. Genre rather than subject domain (i.e., topic) is used to categorise test collections. It is more meaningful to use genre as the category for defining test collections. Texts related to two different topics (e.g., IT technology and electronic engineering) can still be similar from a statistical perspective, as they have similar sentence lengths and vocabulary sets. It would be more difficult to find three suitable test collections from completely different topics. Therefore, collections from three distinct genres that can be used to ensure the representativeness of the experiments were tested. The three test collections include popular scientific articles, technical texts and news articles. The intention is to reflect a wide range of texts that are commonly translated in the real world. Sentences from sources similar to those texts are manually selected as queries for computing the ESA similarity scores for the corresponding test collections.
The evaluation methodology used in this study is significantly different from standard IR or CAT tools evaluations. The key differences centre on the tendency of these evaluations to focus on retrieving translation suggestions rather than on aspects such as accuracy or quality of translation suggestions or the usability of TMS. As explained in Chapter 1, defining the quality/accuracy of a translation suggestion is a different area of research which should study the impact of translation suggestions on translators; and is not addressed in this thesis. The usability of translation memory systems is a non-functional property of translation memory systems that relates more to the design of the interaction between end users and systems (Lagoudaki, 2009: p.178). The differences between the method that are be used in this study and other existing evaluation methods are discussed in Section 4.2.3.

Although potential ESA involvement does not aim to change the fundamental nature of translation memory systems, it may be useful to remember that there are some problems with ESA in the evaluation of translation memory systems. First, different measures are required to evaluate the performance of ESA and the impact of textual factors. As is discussed in Section 4.1.3, common IR metrics (e.g., precision and recall) are not applicable in the evaluation of these experiments. For the purpose of this study, some new measures to evaluate the performance of ESA in TMS tasks must be explained. Moreover, it is important to investigate the meaningfulness of ESA similarity scores with respect to translation tasks. Some new measures are also used to quantify the performance of ESA, possibly allowing interesting patterns of ESA similarity scores to be revealed.

Second, unlike for traditional IR evaluations, large research organisations are usually willing to provide a significant quantity of documents for test collections. Due to human resource limitations, it is impossible to have existing ad hoc test collections for evaluating the performance of ESA in TMS tasks. In view of the need to minimise costs, the focus of any research should be to find other test collections that are more appropriate for the experiment requirements. In this study, new test collections are created from three corpora of documents assembled in the context of other research projects.

Third, as a semantic measure, the value of ESA is calculated differently to string similarity measures, such as Levenshtein distance method. ESA similarity scores are calculated as cosine values rather than percentages. It is reasonable to attempt a direct parallel comparison between the two different scores, although in the same way that weight and volume are different measurement scales that they cannot be compared directly. As is shown in Section 5.2.2.1, an
original cosine value of 0.98 is not equal to a Levenshtein distance score of 98%. Therefore, it is necessary to find new methods to make default cosine values more indicative of a human identification of semantic similarity. More details are given in Section 5.2.2.1.

Fourth, as a type of IR system, translation memory systems are mainly designed to retrieve short text segments rather than entire paragraphs or documents. We can treat a translation suggestion as a short text segment so that TMS tasks are short text IR tasks (Shrestha, 2011). Experiments must reflect the challenge that ESA faces in matching pairs of sentences rather than identifying the similarity between terms or documents.

The fifth problem is that it is be the responsibility of this research to identify the textual factors that can affect the performance of ESA. Researchers in other fields already have a profound understanding of ESA performance. Machine learning studies on ESA have confirmed that the ability of ESA is related to the size of its training sets (Anderka & Stein, 2009). In this study, texts serve as the primary objects processed by translation memory systems. Therefore, it is necessary to examine the impact of textual factors on translation memory systems.

Due to the above factors, the experiments in this study are designed differently for the evaluation of ESA. Overall, this is reflected in the proposed research questions. As the main purpose of the study is to demonstrate the possibility of using ESA for performing TMS tasks, some aspects of the ESA IR ‘reflect a number of inherent limitations of this particular ESA implementation (Sorg, 2010). As explained in Chapter 1, it is beyond the scope of the thesis to attempt to enhance the performance of ESA from a technical perspective. Therefore, the following aspects were excluded in reaching a conclusion as to the possibility of using ESA in TMS tasks:

1) This study does not consider factors that may affect the practical implementation of ESA, such as the speed of the ESA implementation and the disk space required for generating a large Wikipedia matrix.

2) The overall error rate of the ESA IR platform cannot be reflected in this study. During the experiments, the ESA IR platform may have missed valid results from the databases or may have failed to respond to certain queries.

3) Not all the valid results retrieved by the ESA IR platform appear to be of obvious use to translators, largely as a result of the limited size of the test collections available for the study.
Given the fact that my research is still at a stage of development, the study focuses on investigating the possibility and potential for involving ESA techniques in translation memory systems, rather than on offering a comprehensive analysis of the actual performance of ESA techniques as a functioning component within a fully implemented scenario. The next section explains the research questions and their relevance to translation memory systems.

### 4.1.2 Research Questions

This evaluation will seek to evaluate the possibility of using ESA, as a potentially useful semantic measure for translation memory systems. As has been discussed in the last section, the involvement of any new matching method is not intended to change the nature of translation memory systems, but to improve them: namely, to improve the efficiency of using TM files. Thus, the evaluation of ESA in the TMS tasks should include the following:

1) Compare the performance to other matching methods.
2) Focus on what kinds of result the matching methods can provide.
3) Understand the technical and textual factors that affect the matching methods.
4) Build the evaluation on the capacities related to the core function of TMS.
5) Base the evaluation on monolingual tasks.

The evaluation and the design of the experiment should be distinguished from the evaluation of document collections as resources for TM files. This would focus on whether certain translation resources can provide sufficient information for translators or be of better quality. For evaluations of TM files, the matching method is not of the utmost importance. Evaluations of TM files may instead rely on particular statistical measures, especially recall rate, to determine if a given set of queries corresponds to the selected documents. Fundamentally, the evaluation in this study is not just a matching method to determine how many queries can be matched, because the potential queries selected will not necessarily have matches. Thus, it is not applicable to use statistical measures, such as recall. Accordingly, these concerns can be divided in the following two lines of inquiry:

- A: To what extent is the ESA similarity score useful for translation memory tasks?
- B: How do textual factors affect the performance of ESA?
Three research questions are therefore proposed. These are divided into two groups (A and B) one of which is further subdivided (A.1 and A.2). They will first ensure the possibility of using ESA in TMS tasks by comparing the performance of Levenshtein distance method, and examine if any statistical pattern appears in the ESA similarity score. Hence, efforts will also be made to examine the impact of textual factors (both related to test collections and queries) on the ESA performance. Each research question is introduced in the following sections.

4.1.2.1 Research Question A.1

The first part of the evaluation tests the extent to which ESA can be applied to TMS tasks. To show this, we need to examine translation suggestions retrieved by the ESA IR platform and compare them with the Levenshtein distance method. Research question A.1 is presented below.

**What is the possible range of ESA similarity scores that can give rise to potential translation suggestions?**

Research question A.1 provides the overall statistics for the ESA similarity scores relating to valid translation suggestions. Forty results are gathered from each test collection, for a total of 120 results. Ideally, a greater number of results enhances the performance of ESA. Evaluation is meaningful when it has sufficient results that reflect the performance of matching methods. However, the numbers of queries are determined by both the size of the research team and the matching methods. The number of results is relatively small in this study, when comparing it to two very different matching methods, i.e., ESA and the Levenshtein distance method, as the difference between methods would be fairly obvious even from a relatively small-scale experiment. In this study, the number of results ensures that valid and accurate statistics—for example, mean and SD—should be based on a large number of examples. The resulting scores demonstrate the performance of ESA for different test collections, and are compared with the scores of the Levenshtein distance method.
The purpose of research question A.1 is to examine the possibility of using ESA in TMS tasks and to determine threshold value of ESA similarity scores as baselines for the following discussion.

Because the data set is large, both the mean and median were used to show the average values of the ESA similarity scores; more specifically, the value of the arithmetic mean is used. The standard deviation (SD) is also used to indicate the spread of the ESA similarity scores.

Accordingly, mean and median are defined as follows:

1) The arithmetic mean of a set of \( n \) measurements is equal to the sum of the measurements divided by \( n \). (Mendenhall, 2009: p.54)

\[
\bar{x} = \frac{\sum x_i}{n}
\]

Where:

\( \bar{x} \) is the mean of the ESA similarity scores.

\( n \) is the number of results.

\( x \) is each individual ESA similarity score.

2) The median \( m \) of a set of \( n \) measurements is ‘the value of \( x \) that falls in the middle position when the measurements are ordered from smallest to largest’ (Mendenhall, 2009: p.55).

3) SD is defined as follows:

The SD of a set of measurement is equal to the positive square root of the variation. (Mendenhall, 2009: p.62)

The SD of the ESA similarity scores is thus calculated as follows:

\[
\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}
\]
where:

\( \sigma \) is the SD of the ESA similarity scores.

\( \bar{x} \) is the mean of the ESA similarity scores.

\( n \) is the number of results.

\( x \) is each individual ESA similarity score.

### 4.1.2.2 Research Question A.2

Research question A.2 focuses on the translation suggestions that were retrieved by ESA IR platform. Research question A.2 is presented below.

**What generalisations can be made about the kinds of translation suggestion that are typically retrieved?**

Research question A.2 serves two purposes. First, it explores any distinguishing features of the translation suggestions that have the semantic motivation. The importance of observing these potential differences between different kinds of translation suggestions is examined.

The second purpose of the question is to identity any statistical patterns within the ESA similarity scores of valid translation suggestions. If the ESA similarity scores feature a certain statistical pattern, this could help translation memory systems provide translation suggestions accordingly.

### 4.1.2.3 Research Question B

The second part of the evaluation is designed to examine how textual factors affect the performance of ESA. The performance of ESA is measured by the ESA similarity scores and the lengths of valid translation suggestions. This serves two purposes. First, the question is necessary to estimate how textual factors affect ESA performance, especially the threshold values for retrieving potentially valid translation suggestions. Second, the investigation of the
performance of ESA shows whether ESA prefers particular ranges of query lengths to produce more translation suggestions.

The lengths of valid translation suggestions are important indicators of how much information translators receive from translation memory systems. The queries/translation suggestions ratio (QTR), rather than the number of words in a translation suggestion, is developed to represent the lengths of valid translation suggestions. QTR is defined specially for this study in the following way:

\[ QTR = \frac{\text{word number of a query}}{\text{word number of a translation suggestion}} \]

The QTR indicates how much information translators receive from a translation memory system and represents the workload of using translation suggestions. The QTR is more useful than the number of words in a translation suggestion. This is because, the formula reveals, the QTR ascertains a correlation between the query lengths and the translation suggestions. As a result, the QTR comprises both queries and translation suggestions. Ideally, if the QTR is equal to one, the query length is the same as the length of the sentence in the test collection. In this circumstance, the translators receive the same amount of text from a translation suggestion. If the QTR is > 1, the query length is greater than the length of the translation suggestions from the test collection. In these circumstances, the translators may receive an insufficient amount of text from the translation suggestion. Consequently, the translators do not have all the information from the translation suggestions with a high QTR. If the QTR is < 1, translation suggestions are longer than queries. For example, if an 8-word query has a QTR of 0.5, its translation suggestion is 16 words. This suggests that translation memory systems could cause information overload as translators are forced to read longer translation suggestions. Consequently, the TMS workload is increased.

Several textual factors relating to the genres, including the complexity of vocabulary, sentence length and changes to sentences, are employed to test how textual factors can affect the performance of ESA. These textual factors are quantified using different measures. Research question B is presented below.
How do the following textual factors affect the performance of ESA: query length, type/token ratio (TTR), size, and average sentence length (ASL) of the genre group?

These are being studied as examples of relevant factors. It should be noted that the selection of the textual factors above is not intended to reflect all the linguistic features of a test collection, but is based on the factors that are considered to be most relevant to the work of translation professionals. Research question B is based on the data from the ESA matching experiment, but focuses on the impact of different textual factors as follows:

1) Query length, that is, the number of words in a query, is selected to determine if there is an ideal length range that enables ESA to outperform other ranges.

2) The TTR (type-token ratio) is the ratio between types (unique words) and tokens (total words), as shown below (Covington, 2008):

\[ TTR = \frac{\text{type}}{\text{token}} \]

The TTRs of the test collections show their vocabulary complexity (Covington, 2008). A higher TTR means that a test collection that contains proportionately more types (i.e., more unique words). This may have affected ESA performance. Translators might have difficulty working with articles with many unique words or terms. Therefore, it is important to see if TTR could also affect ESA. The links between TTR and the ESA similarity scores of different test collections are examined.

3) Average sentence length (ASL) is also a textual factor that could affect ESA performance. ASLs of different genres can be varied, and translators may perform differently when translating long sentences. Therefore, it is important to determine whether there is a link between the ASL of the test collection and other textual factors and the ESA similarity scores. The ASL of a test collection is defined as follows:

\[ \text{ASL} = \frac{\text{number of sentences}}{\text{tokens of a test collection}} \]

The number of sentences in each test collection is obtained in the preparation phase of the experimental procedure.
4) The size of a test collection refers to its number of words. As is the case with the factors mentioned above, the size of the test collections might have affected ESA performance. The size of each test collection was obtained in the experimental procedure.

**4.1.2.4 Summary of Research Questions**

The evaluation method used in this study involves a set of research questions which have the aim of examining the possibility of using ESA in TMS tasks.

As a first step, the statistics of the ESA similarity scores for valid translation suggestions will be compared with the Levenshtein distance scores. Then, the level of generality of the translation suggestions that were retrieved by ESA and any statistical patterns of their ESA similarity scores will be explored. The impact on the ESA performance of different textual factors that relate to genres will be also examined, as they will indicate the types of genre ESA could process. There are many evaluation methods that could have potentials similar to the scope of this study, particularly in the study of IR and MT systems. However, researchers need to modify previous evaluation methodologies that possibly place a different emphasis on ESA. After reviewing the details of the experiments and research problems, a brief overview of other similar attempts will be provided in Section 4.2.3.

**4.1.3 Summary of the Evaluation Method**

In this study, the 'possibility' of ESA in TMS refers to ESA's ability to retrieve translation suggestions, because it enables translation memory systems to match translation suggestions on a semantic level. As shown in Section 4.1.1, 'efficiency' is the key aspect for evaluating ESA in TMS tasks. The involvement of the semantic processing technique does not change the fact that TMS is a type of CAT tool. Translation memory systems are used to minimise the workload of translators by offering translation suggestions that are similar to segments of the text the translators must process. Translators are still required to edit translation suggestions according to the context of the queried text. In these circumstances, translation memory systems should aim to provide valid translation suggestions that are potentially useful to
translators. Therefore, how ESA affects the efficiency of translation memory systems is the most important aspect of the evaluation.

4.2 Experiment Design

This section contains details of the tools, test collections and the experiments used in this methodology.

4.2.1 Tools

The tools used for the experiments are employed for three tasks:

- Task 1: Pre-processing the test collections
- Task 2: Retrieving and computing the ESA similarity scores
- Task 3: Analysing the data produced by the experiments

The implementation of the experiment is conducted in a Windows 7 system.

4.2.1.1 Tools for Pre-processing the Test Collections

Before conducting the experiments, the test collections had to be pre-processed. The objective is to make the test collections ready for experiments and to obtain the necessary information about the test collection. As stated in Section 4.1.1, the test collections were not specific to this study. The preparation of tools for this task is time-consuming, because the researchers need to produce different test collections from scratch. There were three sub-tasks involved in the process, Task 1.1, Task 1.2 and Task 1.3.

For Task 1.1, test collections had to be saved in an appropriate format to enable them to be processed by the software platform. Task 1.1 involved sub tasks such as removing unnecessary tags and merging multiple text files into one large file.
For Task 1.2, the textual statistics of the test collections also had to be acquired, as all three test collections become *ad hoc* test collections. The topic modelling technique was also used. Topic modelling is a technique whereby ‘topics’ can be extracted from given collections of texts. Each topic is presented as a term and can be used to indicate a subject contained in document collection (Hagedorn, 2013). Topic modelling of test collections offers an initial overview of the content of the test collection.

For Task 1.3, an ESA platform also had to be installed on a local computer, as instructed by the original program developers.

Tools for Task 1.1:

- Java program for removing tags: P1-1rt.9
- Java program for sentence segmentation: P1-1ss.
- MS Dos batch task for merging notepad files: copy.

Tools for Task 1.2:

- Concordance tool: Textstats.
- Java program for topic model: TopicModelingTool-google
- Java program for counting words: P1-2wc.

Tools for Task 1.3:

- ESA source package
- Two Integrated Development Environment workbenches: Eclipse and NetBean
- Perl program for converting Wikipedia dump: xml2sql
- Database for training data: Mysql 5.7.

---

9 It is useful to remind readers that the programs written by the researcher himself were named in this fashion.
4.2.1.2 Tools for Computing Similarity Scores of Text Segments

For Task 2 a software platform was used to simulate the monolingual information retrieval system operated by translation memory systems to match texts being translated with translation units from the TM database. The objective of the task is to generate the data for the experiments. An ESA implementation based on the Terrier information retrieval framework is the core of this platform and can be modified to measure the semantic similarity of pairs of queries and sentences from the test collections according to different research questions. Task 2 consisted of three sub-tasks:

- Task 2.1 aims at computing the ESA similarity scores between queries and sentences from the test collections;
- Task 2.2 aims at computing the ESA similarity scores between retrieved translation suggestions and their paraphrased versions;
- Task 2.3 aims at computing the Levenshtein distance scores between queries and sentences from the test collections.

Most programs in Task 2 needed to be customised for the purpose of the study. They have been named as follows:

- Programs for Task 2.1: ComputeESASimilarity15Results.java.
- Programs for Task 2.2: SimilarityComputeTwoUnits.java.
- Programs for Task 2.3: OmegaT.

4.2.1.3 Excel Functions for Analysing Data

The objective of this task is to analyse the results produced by the experiments. Once all data were obtained from Task 1 and 2, the next task was to analyse them and present them in a suitable way. MS Excel 2010 was the tool used to complete task 3. The evaluation processes many kinds of data and raw information, including the scores of matched queries, the texts of queries, the sources of queries, and a range of different statistics. Several Excel functions were used to calculate the final results of the experiments. It would be unrealistic to record the data
without the appropriate software, as most patterns of the results can only be observed from a statistical perspective.

4.2.2 Sources of Test Collections and Potential Queries

For materials used in the evaluation of semantic processing techniques in matching translation suggestions for translation memory systems, researchers should have two sets of texts: one for queries and another for test collections. The requirements for test collections are the same as for TM files that are sufficiently large thereby providing a high possibility of sentences to be matched, and ideally, the test collection should be as large as possible. As discussed, *ad hoc* test collections were required for the purpose of experimentation. Three test collections of different genres were used to test the ESA implementation in the performance of TMS tasks:

A) Air Accidents Investigation Branch (AAIB) aircraft accident reports (ACA)

B) *Scientific American* articles (SciAm)

C) Reuters-21578 Collection (RN)

The first two text collections were obtained manually online and saved in text files. The RN is a classic test collection that contains 21,578 newswire articles, all of which are tagged. A sample text is as follows:

```
<REUTERS TOPICS=""798"" DEPSSPLIT=""TRAIN"
CGISSPLIT=""TRAINING-SET"" OLDID=""12981"" NEWID=""798">
<Date>2-MAR-1987 16:51:43.42</Date>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
<DateLine>CHICAGO, March 2 - </DateLine><Body>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC. Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said. A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter</Body></Text></REUTERS>

Figure 4.1: Extract of an RN text (Manning, 2008: p. 281)
The RN collection was selected as one of the test collections, although it is more often used to evaluate text classification. Nevertheless, it was still suitable for the purpose of the experiments, because it covers a wide range of topics and is much smaller (8MB) than other IR test collections.

Although only one of them is a standard data set, both the ACA and SciAm collections have been used for the following reasons:

1) Unlike some standard IR test collections, such as Text Retrieval Conference (TREC) collections, the size of these collections chosen for this study both balance the need to have representative texts and the computing power needed by the author. Test collections can be enormous; for example, the size of TREC collections ranges from 2GB to 2.25TB (Baeza-Yates, 2011:p.162). It is impossible to process such quantities of text using only one personal computer.

2) By the time the RN test collection was obtained, it had already been saved in machine-readable form with either no special formatting or formatting that was easily removable. This has minimised the difficulty of using these texts for the experiments.

However, it is still necessary to pre-process these three document collections to enable them to be processed by the ESA IR platform and to provide textual statistics. Detailed information on the test collections is usually available, as are comparisons of the performances of different IR systems or retrieval methods. Since the three test collections need to be adjusted and pre-processed to meet the specific aims of the experiments, it was the researcher’s responsibility to provide the information for describing the new ad hoc test collections: for example, size, number of sentences, and the type and token count of the test collections. Information on how the data is used, as well as details of the pre-processing phase are given in Section 4.2.3.

Once the test collections are ready, potential queries that are sentences for testing need to be chosen according to the corresponding genre groups. The basic principle involved in finding such text segments is that they are from sources similar to the corresponding genre groups. This increases the possibility of having potential matches; in other words, the query sentences were taken from semantically similar texts.

The ACA test collection mainly comprises aircraft accident reports produced by the AAIB prior to 2001. It is reasonable to get similar reports from the same source to have a higher
chance of having similar text matches. Potential queries from publications on the AAIB website (aaib.gov.uk), including recent bulletins, annual safety reports and formal reports are therefore used as queries.

In the case of SciAm, the potential queries were mainly taken from articles on Chinese-English learner websites. These websites offer free bilingually aligned texts from various English language sources, and articles are organised into different topics. The SciAm test collection mainly contained popular scientific articles, although there were many difficult terms in the text. The syntactic structure is relatively easy compared to ACA texts. The information on such websites is an ideal reference for translators, and in general is expected to offer greater matching possibilities. For such websites, it is also more convenient to find articles about similar subject areas to those of SciAm test collections.

In the case of the RN test collection, the text segments for testing included sentences from websites offering financial and economic news, as they provide news similar to Reuters and could therefore result in potential matches. The list of websites for sources of potential queries can be found in Appendix A.

4.2.3 Experimental Procedure

The experiments were conducted in order to obtain results in answer to the research questions posed in Section 4.1.4. The experimental procedure was divided into two phases: preparing the experiments and answering the research questions.

The first phase only involved the tools for Task 1, while both Task 2 and Task 3 were often applied in the second phase. Each of the two phases is first presented in a table giving details of each step.
4.2.3.1 Phase 1: Preparation

Task 1 was carried out during the preparation phase: Task 1.1 involved the pre-processing of raw documents and the creation of a test collection; the aim of Task 1.2 was to obtain detailed information relating to the test collection. Task 1.3 concerned the configuration of the ESA software. The above mentioned tasks provide the technical architecture and the necessary information for experiments A and B. Table 4.1 below summarises the basic steps in Phase 1:

<table>
<thead>
<tr>
<th>Phrase 1</th>
<th>Tasks</th>
<th>Steps</th>
<th>Data or files produced</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>Task 1.1: Pre-processing raw documents and producing test collection (s)</td>
<td>Step 1 Placing All Raw Documents in a Folder</td>
<td>Raw document collection</td>
<td>DOS-command line: copy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 2 Merging All Documents to Create a Large Raw Document Collection</td>
<td>Raw document collection</td>
<td>P1-1rt</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 3 Removing Unnecessary Tags, Symbols, and Text segments</td>
<td>Raw document collection</td>
<td>P1-1ss</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 4 Segmenting Sentences of the Selected Document Collection</td>
<td>Test Collection</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task 1.2: Obtained Details of Test Collections</td>
<td>Step 5 Calculating the Size of a Test Collection</td>
<td>Size of a test collection</td>
<td>P1-2wc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 6 Calculating the number of sentences in a test collection</td>
<td>Number of sentences</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 7 Calculating the ASL of a Test Collection</td>
<td>ASL</td>
<td>Excel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 8 Calculating Type and Token count of a Test Collection</td>
<td>Type and Token</td>
<td>TextStats</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 9 Calculating the TTR</td>
<td>TTR</td>
<td>Excel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 10 Topic Modelling</td>
<td>Topic model</td>
<td>TopicModeling Tool-google</td>
</tr>
<tr>
<td></td>
<td>Task 1.3 Installation and Configuration of ESA IR Platform</td>
<td>Step 11 Downloading and Configuring the ESA Source Package</td>
<td>ESA source package</td>
<td></td>
</tr>
<tr>
<td>Step 12</td>
<td>Downloading the Wikipedia Dump File for Training the ESA IR Platform</td>
<td>Wikipedia dump 'newwiki-20130821-pages-meta-history.xml.bz2'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 13</td>
<td>Pre-processing Wikipedia Dumps</td>
<td>Page.sql, Revision.sql, and Text.sql</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>XML2SQL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 14</td>
<td>Importing the Data to MySQL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MySQL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 15</td>
<td>Training the ESA Knowledge Base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESA IR platform</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 16</td>
<td>Adding Programs</td>
<td>ComputeESASimilarity15Results.java and SimilarityComputeTwoUnits.java</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ESA IR platform</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: An overview of phase 1

This table focuses on the SciAm test collection in Tasks 1.1 and 1.2 for illustrative purposes only. The remaining two test collections were produced using the same procedure.

Task 1.1: Pre-processing Raw Documents and Producing New Test Collections

- **Step 1: Placing All Raw Documents in a Folder**

The *Scientific American* articles had been collected manually for other purposes, and the documents were saved separately as text files, as shown in Figure 4.2. They were then moved to a dedicated folder.
For test collections ACA and RN, the same procedure was repeated.

- **Step 2: Merging all Documents to Create a Large Raw Document Collection**

The MS-DOS ‘Copy’ command was used to concatenate the text files into a single file in the folder, as one large text file is easier to process than multiple files.

The syntax used was as follows: ‘copy *.txt SciAmrt-1’.

The new single file was named ‘SciAmrt-1’.

For genre groups ACA and RN, the same action was repeated, and produced the files ‘ACArt-1’ and ‘RNrt-1’ respectively.

- **Step 3: Removing Unnecessary Tags, Symbols and Text**

After being merged, many text segments formed independent sentences, but could not function as translation suggestions. The reason for this is that the presence of a large number of tags, symbols and text can affect the function of the ESA in tasks. Thus, the P1-1rt program was used to remove unnecessary tags, symbols and text.
The tags and symbols included the following:

Long white spaces, ellipses, hyphens, bullet points and brackets contained in the text segments.

The Figure 4.3 shows some typical instances of the contents before processing:

Figure 4.3: A screenshot of an original AAIB aircraft accident report: The text contained a large number of white spaces.
The following text segments were deleted:

1) Entire segments consisting of single strings of fewer than four letters (such as ‘2. 56C’)
2) Text segments containing fewer than three words
3) Web addresses
4) Addresses
5) Contact details (telephone etc.)
6) Names of organisations
7) References
8) Text segments containing incomprehensible codes (such as ‘Chevron Corp &lt;CHV&gt; jumped1-1/8 to 58-7/8’)
9) Text segments containing mainly numbers (such as ‘8.1 75.7 75.7 CORN
1,800 2,400 2,589 2,589 Special Producer Storage Loan Program -- WHEAT
165 150 163 163 FEEDGRAINS’)

These text segments contributed little value as translation suggestions and were also removed to improve the quality of translation suggestions.

Following this step, the new document collection was called SciAm_2.

For genre groups ACA and RN, the same action was repeated, and produced the files ACA _2 and RN_2, respectively.

- **Step 4: Segmenting Sentences from the Selected Document Collection**

The P1-1ss program was used to segment SciAm-2. As SciAm.txt included texts that were copied manually from the internet, those articles remained in formats that could be easily read by human readers. This means that they were organised into different paragraphs. However, the format was not useful for processing by the ESA IR platform, because the platform provides one sentence as one translation suggestion unit. After Step 2, the original format was corrupted in places.
For the convenience of retrieval by the ESA IR platform, the sentences were segmented by adding a line break between each.

By default, a sentence is counted as one unit of a translation suggestion. By counting the number of line breaks, it was also easy to ascertain how many translation suggestions were contained in the test collection. For these two reasons, the P1-1ss program was used to segment the sentences.

P1-1ss program was instructed to add line breaks where the sentence was followed by punctuation marks, including ‘.’ (Full stop), ‘?’ (question mark), ‘!’ (exclamation mark), and ‘:’, (semicolon), as well as sentences longer than 23 words with ‘,’ (comma).

It should be noted that ‘sentences’ does not just refer to full sentences but also to long clauses, as they can be useful to translators as well. Therefore, such text segments were also treated as independent ‘sentences’. Hence, they were also regarded as one unit of a translation suggestion.

After Step 4, ‘SciAm_2’ became the SciAm test collection and was ready for task 1.2.

For genre groups ACA and RN, the same action was repeated, and produced the ACA and RN test collections.

Task 1.2: Obtaining Details of the Test Collections

- **Step 5: Calculating the Size of a Test Collection**

Once the SciAm test collection was ready, the Java P1-2wc program was used to count the number of words in the collection.

- **Step 6: Calculating the number of sentences in a test collection**

P1-1rt was used to segment sentences of test collections, so line breakers at the end of each sentence were added in Step 3. Every sentence was separated each other by its line breaker. Knowing the number of line breakers makes it possible to assess the number of sentences (i.e., potential translation suggestions) in the test collection.
• **Step 7: Calculating the ASL of a Test Collection**

After calculating the size and number of sentences in SciAm, it was easy to calculate the ASL of SciAm in characters.

The actual calculation was performed in Excel.

• **Step 8: Calculating Type and Token count of a Test Collection**

TextStats was used to ascertain the type and token counts of the SciAm test collection.

• **Step 9: Calculating the TTR**

TextStats was used to calculate the TTR. As Figure: 4.5: shows, it can indicate the type (total words) and token (total unique words) of a test collection:

![TextStats](image)

Figure 4.5: A screenshot of TextStats

This calculation was also conducted in Excel using the TTR formula given in Section 4.1.4.2.
For test collections ACA and RN, Steps 5-9 were repeated in order to acquire the necessary information.

- **Step 10: Topic Modelling**

The ‘TopicModelingTool-google’ program was employed to generate the topic model directly. This process was also carried out in order to obtain topic models for the ACA and RN test collections.

Task 1.3 Installation and Configuration of the ESA Platform

Step 11-6 covered the procedure for creating an ESA IR platform. The ESA source package is not an executable program and therefore had to be installed and modified according to the purpose of the evaluation.

- **Step 11: Downloading and Configuring the ESA Source Package**

The ESA source package was downloaded from Sorg’s site (2010). The following libraries were required in order for ESA to run on a local personal computer:

  - Matrix Toolkits for Java (MTJ)
  - Snowball Stemmer
  - Terrier (Version 3)
  - Apache Commons CLI
  - Apache Commons Collections
  - Apache Commons Configuration
  - Apache Commons Lang
  - Apache Commons Logging
  - Apache log4j
  - GNU Trove
  - Spring Framework (Version 3)
- **Step 12: Downloading the Wikipedia Dump File for Training the ESA IR Platform**

A suitable Wikipedia dump was selected comprising a collection of Wikipedia articles from Wikimedia, and it was decided to use this dump as the source of the knowledge base to be used by the ESA to measure semantic similarity. Most of these are very large files that cannot be processed by personal computers. Not only are they too large to download, but the training process also requires huge computing power. After several attempts of testing, the maximum space available on the authors’ computer was 100 MB.

For this study, the Wikipedia dump ‘newwiki-20130821-pages-meta-history.xml.bz2’ was selected, because it contains a large amount of information in English and is also of a suitable size. Its compressed size was 63.6 MB.

- **Step 13: Pre-processing Wikipedia Dumps**

The Wikipedia dump was saved in XML format. It needed to be converted into SQL files in order to be used by the ESA platform. XML2sql, a Perl program, was recommended by the original developer and was used to process the Wikipedia dump. A screenshot of XML2sql is shown in Figure 4.6.

![XML2sql](image)

Figure 4.6: A screenshot of XML2sql
After being processed, three SQL files were produced by XML2sql, as follows: Page.sql, Revision.sql and Text.sql.

- **Step 14: Importing the Data to MySQL**

Using MySQL the processed Wikipedia dump files can be imported into the MySQL database.

- **Step 15: Training the ESA Knowledge Base**

Once the Wikipedia dump had been imported into the MySQL database, it was possible for the ESA platform to use the knowledge stored in the Wikipedia dump.

‘BuildWikipediaIndex’ was used to train the ESA platform, thus allowing the knowledge for the ESA IR platform to be fully populated.

- **Step 16: Adding Programs**

The last step was to add the required programs ‘ComputeESASimilarity15Results.java’ and ‘SimilarityComputeTwoUnits.java’. This was achieved by adding two classes to the ‘demo’ package. See Figure 4.7 below.

![Figure 4.7: A screenshot of the ESA IR platform: only ComputeESASimilarity15Results.java'and'SimilarityComputeTwoUnits.java' were used in the experiments](image-url)
On completion of these steps, the installation and configuration of the ESA IR platform for conducting experiments was complete.

### 4.2.3.2 Phase 2: Answering the Research Questions

Phase 2 involved experiment for the research questions. ESA matching experiments involved collecting the results for answering research questions A.1, A.2 and B. Readers are reminded that the illustration of all the experiments was based on test collection SciAm. However, the same action was repeated to obtain the data based on RN and ACA.

An overview of phase 2 is given in Table 4.2 below:

<table>
<thead>
<tr>
<th>Phase 2</th>
<th>Experiment</th>
<th>Steps</th>
<th>Data or files produced</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answering Research Questions</td>
<td>ESA matching experiment: Completing Tasks 2.1 and 2.3 for Each Test Collection</td>
<td>Step 1 Selecting a Test Collection</td>
<td>A test collection</td>
<td>ESA IR platform</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 2 Finding a Suitable Query from Sources Similar to the Test Collection</td>
<td>Suitable Queries</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 3 Computing the ESA Similarity Scores between the Selected Queries and Text Segmentations from the Test Collection</td>
<td>translation suggestions with ESA similarity scores</td>
<td>ComputeESASimilarity15Results.java</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 4 Computing Levenshtein distance scores between the Valid Query and Sentences from the Test Collections</td>
<td>Levenshtein distance scores</td>
<td>OmegaT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Step 5 Repeat Steps 2-4 until Collecting 40 Results for Each Text Collection</td>
<td>Results for answering Questions A.1, A.2, and B.</td>
<td>Excel</td>
</tr>
</tbody>
</table>

Table 4.2: An overview of phase 2
ESA matching experiment: Completing Tasks 2.1 and 2.3 for Each Test Collection

- **Step 1: Selecting a Test Collection**

As explained above, three test collections were used in the experiments and the research required that each test collection produces its own set of results. In order to achieve this, the ESA platform was instructed to use a particular test collection. In this case, test collection SciAm was selected first.

- **Step 2: Finding a Suitable Query from Sources Similar to the Test Collection**

The sources of potential queries mentioned in Section 4.2.2 were manually searched for potential and meaningful sentences that could be used as queries. A brief understanding of the content of the test collection can improve the process of searching for potential queries.

It should also be remembered that a particular preference was used in searching for potential sentences, which is described below:

1) Sentences had to be representative in the test collections and have to be meaningful to be translated. It was inevitable that all texts would contain fixed expressions used in other genres or subject domains, although attempts were made to avoid this before conducting the experiments. An example of this is an introductory phrase such as ‘this paper presents a new method for clustering texts’. Such sentences that may contain different subjects, but are uniform in terms of syntactic structure, were avoided as potential queries.

2) Potential queries need to be somewhat similar to sentences from the test collections. Due to the preparatory work done in Phase 1, the researcher was already familiar with test collections from the perspective of sentence length, topics and terms.

For test collections ACA and RN, the same actions were performed during this step.
• **Step 3: Computing the ESA Similarity Scores between the Selected Queries and Text Segmentations from the Test Collection**

As soon as a potential query was selected, the ‘ComputeESASimilarity15Results.java’ program was used to compute its ESA similarity scores with all the sentences from the SciAm test collection.

ComputeESASimilarity15Results.java had been configured to only show the top 15 potential results. In theory, the ESA IR platform could display a large number of results from a test collection. However, most existing TMS platforms do not display that many results for translators, because of the limitations of a typical TM database. It is also an unrealistic expectation for translators to find one useful translation suggestion from among a huge number of results every time, in the same way that Internet users would expect their queries to be answered by the first few hits from a Google search, for example. To create a balance between such a realistic situation of TMS use and the actual need to evaluate ESA performance, I decided on only using the top 15 results provided by ESA.

These 15 potential results were read manually to check for valid translation suggestions. The ‘valid result’ here is understood as a valid translation suggestion that shows either semantic or syntactical similarity to the query. A single query may have one, multiple matches, or no match which has very low ESA similarity scores.

Here, the evaluation was not intended to establish a general principle to identify and define valid translation suggestions, because that was not the aim of this research. The judgement of identifying valid translation suggestions should be based on human intuition rather than sets of rules, as the former is closer to the actual working conditions of translators. Translators are unlikely to follow a set of strict rules to translate texts because translating is a highly flexible and personal act (Pym, 2003). In this case, if no valid translation suggestion was found, the researcher had to repeat Step 2. If a valid translation suggestion was identified, it had to be recorded in an Excel file.

The default ESA similarity scores computed by the ESA IR platform were cosine values that were not easily identified. This is because default values were distributed in a very concentrated range. For example, the original cosine values for ACA ranged from 0.908 to 0.999. A conversion had to be carried out to make the original ESA values more identifiable.
This means that all original values were then converted in their antilogarithm form to base 10 and calculated as follows:

$$ESA \text{ similarity score} = \log_{10}^{-1}(\text{default cosine value})$$

Where, $0 \ll ESA \text{ similarity score} \ll 10$.

A score of ‘10’ means semantically identical, and a score of ‘0’ means not similar at all. In this study, an ESA similarity score of 10 is defined as being 100% equivalent according to the Levenshtein distance score. A cosine value of 0.908 has a converted antilogarithm ESA similarity score of 8.091. In theory, there are many different options for presenting the scores, for example the antilogarithm base can be 20. But the conversion is not designed to change the original scores in order to make the ESA similarity score functionally equal to the fuzzy matching score in terms of its rating of translation suggestions, but is to present the scores in a form that is more accessible for researchers.

It should be noted that Steps 2 and 3 can be laborious and time consuming, because a potential query does not necessarily have valid translation suggestions matched in the test collection.

- **Step 4: Computing Levenshtein Distance Scores between the Valid Query and Sentences from the Test Collections**

After acquiring a valid translation suggestion retrieved by ESA, it is interesting to see if the Levenshtein distance method can also identify valid translation suggestions. OmegaT was the tool used to compute the Levenshtein distance.

For test collections ACA and RN, the same action was conducted during this step.

- **Step 5: Repeat Steps 2-4 until 40 Results have been Collected for each Test Collection**

40 results were collected as data in response to questions A.1, A.2 and B. Obtaining these results was a time-consuming process and the total number of valid translation suggestions was 120 on completion of all three test collections.
The data resulting from the ESA matching experiment was saved in Excel, and were used to answer research question questions A.1, A.2 and B. Details will be given in the next chapter.

### 4.2.4 Other Evaluation Methods

In this study, ESA is not evaluated only as a semantic similarity measure, but its use in TMS tasks is also considered. This is to suggest that the evaluation of ESA is an application-based evaluation and should be incorporated in the TMS task. The methods to evaluate semantic similarity measures are reviewed in Section 3.2.3. The methodology described in Section 4.2.3 is similar to the evaluation methods for IR and MT systems because, from certain perspectives, the results retrieved by the ESA IR system are similar to the output of MT or of IR systems. However, the method applied in this study is different in terms of test collections and measures. The main reason is that ESA is a technique to measure semantic similarity, and the evaluation method in this present work considers the use of ESA in a TMS scenario, whereas so far neither ESA nor similar techniques appear to have been implemented in translation memory systems. This section summarises several evaluation methods for IR and MT systems, highlighting why they are not applicable to a TMS scenario.

IR evaluation is potentially applicable to translation memory systems. As has often been observed, the TMS task is another form of the IR task (Whyman & Somers, 1999; Macken, 2010; Colominas, 2008). TMS is generally described as a type of database that stores bilingually aligned texts, and searches the relevant translation units when new texts are inputted as queries (Trujillo, 1999: p.60). The evaluation of IR systems is usually based on a set of test collections that contain a large number of documents, including Text Retrieval Conference (TREC) collections (Baeza-Yates, 2011: p.160). As has been mentioned in Chapter 1, IR systems are mainly measured by precision and recall metrics. These two metrics are calculated differently depending on the types of applications or accuracy requirements (Baeza-Yates, 2011: p.158). Precision is a measure of the accuracy of an IR system in retrieving relevant documents. Recall is a measure of the number of relevant documents retrieved by the IR system.

For instance, let \( |A| \) be the number of the answers and \( |R| \) be the number of relevant documents; then, \( |A \cap R| \) is the number of relevant documents existing in an answer set. Precision and recall are defined as follows (Baeza-Yates, 2011: p. 135):
Precision = \( p = \frac{|R \cap A|}{|A|} \)

Recall = \( r = \frac{|R \cap A|}{|R|} \)

It is possible to make a comparison between systems by allowing different systems to work with the same test collection or by judging the performance of an IR system by comparing its scores with a baseline that is already known by the producer of the test collections (Baeza-Yates, 2011: p. 136). Many researchers place TMS evaluation in the same framework as IR evaluation because of the technical similarities between TMS tasks and IR tasks (Whyman et al., 1999).

Despite these similarities, translation memory systems cannot be evaluated the same way. An immediate problem in TMS evaluation is that the IR test collection is not suitable for TMS. The results provided by TMS are fundamentally different from the feedback of a text categorisation or IR application. For such applications, the feedback is relatively easy to judge, as correct or incorrect. However, the results in the case of translation memory systems are more complicated than IR system results because of the inevitable involvement of human judgement in translation suggestions.

The concept of precision is not applicable in the context of translation memory systems. Different fuzzy matching scores indicate that the usefulness and quality of translation suggestions are also different. The use of precision scores from Macken's research is ill defined. Macken (2010) does not show a way to quantify how higher match scores could be more accurate than lower match scores. For example, translation suggestions with a 90% fuzzy matching score are likely to be closer to the query sentences than translation suggestions with a 75% fuzzy matching score. However, they would all be considered equally relevant answers to count as \(|R|\) (i.e., the actual number of matched translation suggestions) in the precision metric.

Therefore, the precision score from the IR evaluation is not applicable to TMS tasks. It is also very difficult to obtain recall for the TMS evaluation. As for every query used in the evaluation, researchers would have to determine manually whether it was matched with every sentence of a corpus to know if any matched pair was lost by translation memory systems. In this way, the recall scores can be calculated. In other words, value \(|R|\) should be known before conducting
the evaluation. It is impossible to judge a query manually if it is matched with every sentence from a large test collection, because the size of the test collection is too great for one researcher. Therefore, IR evaluation methods are not perfectly transferable to the context of TMS evaluation.

Another possible approach would have been to evaluate TMS as if it were a type of MT system, as has been suggested by the Expert Advisory Group on Language Engineering Standards (EAGLES, 1996). Evaluation methods, such as reference-based automatic metrics and The Translation Edit Rate (TER), are very popular in MT evaluation because they tend to be reusable and quick to implement (Kohen, 2011; Schwartz et al., 2006). TER defines the closeness based on the number of edits required to correct an MT output to reference human translations (Schwartz et al., 2006). Similarly, the principle of all reference-based automatic metrics is that 'The closer a machine translation is to a professional human translation, the better it is' (Schwartz et al., 2006). Reference-based automatic metrics take human translations as references and measure similarity scores between machine translation output and human translations (Specia, 2010 & 2013; Kohen, 2011). Different reference-based metrics incorporate different methods to compute their scores, according to different perspectives of viewing 'closeness'. Bilingual Evaluation Understudy (BLEU) compares n-gram matches and sentence lengths of MT output with good human reference translations from a corpus. Some reference-based automatic metrics, such as BLEU, are able to achieve correlation with human judgements of quality (Papineni et al., 2002).

It seems that reference-based automatic metrics are applicable for evaluating the translation suggestions retrieved by ESA, as translation suggestions can be conceptually seen as a type of output of MT. However, these metrics are more suitable to measure the quality of translation output from an MT system that may have a sustained developer group (Kohen, 2013; Specia et al., 2009). Reference-based automatic metrics, such as BLEU, are relatively expensive to implement in terms of demanding suitable resources, which was a problem in this study. It is difficult to collect a set of reference translation suggestions to create an ad-hoc corpus for automatic metrics. More importantly, reference-based automatic metrics aim to evaluate the quality of MT output, rather than the possibility of using a matching method. In the context of translation memory systems, the quality of the translation suggestions retrieved by translation memory systems is a heuristic rather than an absolute concept. For example, the quality of the translation suggestions can be defined by the usefulness of the translation suggestions to translators. However, it is unlikely that two translation suggestions with 80% matching scores
are equally useful for every translator in every situation. Thus, understanding the quality of translation suggestions is beyond the scope of this study. The quality of translation suggestions does not necessarily ensure the possibility of using ESA in translation memory systems. As a semantic measure, the evaluation method in this study prioritises the demonstration of ESA performance; it does not view the use of ESA as a mature technique. As it is beyond the scope of this thesis to present a prototype, it is not possible to have end-users judge the results that are retrieved by this ESA platform. Furthermore, from an economic perspective, it is wise to test the possibility of using a technique in a system before developing a software prototype. Thus, different perspectives for viewing translation memory systems and the role of ESA are required.

One final option might be to evaluate how translation memory systems process paraphrases of texts, where paraphrasing is defined as ‘pairing of phrases that have similar meaning, but different lexical composition and syntactic realisation’ (Pekar & Mitkov, 2007: p.3). As mentioned in Chapter 1, it is highly desirable that translation memory systems should be able to retrieve translation suggestions expressed in different linguistic forms. Researchers such as Pekar and Mitkov (2007) and Marsye (2011) conduct evaluations based on the ability of a TMS’s retrieval methods to recognise paraphrases that investigate the impact of paraphrasing on the ranking of translation suggestions and the ability to capture paraphrased translation suggestions. This approach can also potentially reflect the performance of translation memory systems that are enhanced with semantic similarity measures. However, this approach is an indirect way of evaluating the actual performance of TMS, as the output of translation memory systems is not directly examined by end users. Rather, researchers usually have to paraphrase text segments manually based on a set of rules (Marsye, 2011; Timonera, 2014). Such limitations mean that a possible paraphrasing test would not fully reflect the conditions of realistic TM file use. If only a single word is replaced in a text segment, it is still possible for the segment to be retrieved by the traditional edit distance measure. Researchers such as Marsye – who has a similarly broad understanding of paraphrase to the one that I am using (2011:p.16) – may adopt many paraphrasing methods to transform text segments into different forms, but they may still not cover all the paraphrasing possibilities that can exist in naturally occurring situations other than those that involve the use of synonyms, different voices or different word orders, with an abbreviated form of expression arguably also being considered as a paraphrase. I would therefore argue that it is not feasible to encapsulate all the possible forms that paraphrasing may assume in real situations. Consequently, it may possibly overlook
some other potential uses of translation suggestions. Therefore an evaluation method based on retrieving naturally occurring text is an indispensable step for applying semantic similarity measures in TMS tasks. For this reason, it has been decided not to include a test based on the notion of paraphrasing in this research.

4.3 Conclusion

This chapter presents the methodology of this study. The methodology employs different experiments to evaluate the performance of ESA in TMS tasks. Although there have been several similar research studies presented previously, techniques for the evaluation of IR and MT systems are not suitable for this study. This study emphasises the role of ESA in retrieving translation suggestions, since this is the most important function of translation memory systems, and the involvement of ESA does not change this function.

The test collections and queries should be from similar sources, and should be categorised by genres rather than topics. This is because genres are more indicative than ‘topics’ or ‘subject domain’, as explained in Section 4.1.1. Texts belonging to three types of genre were used, to see how their textual factors can affect the performance of ESA. The textual factors include query length, QTR, ASL, TTR and test collection size. Each test collection provided 40 results to demonstrate the performance of ESA and these are compared with the Levenshtein distance method.

The experimental procedure consisted of the following two phases: Phase 1 consisting of 16 steps, and Phase 2 of 5 steps. Both phases need to be repeated for each test collection.

The next chapter will provide the results of the proposed research questions.
Chapter 5 Results

This chapter presents the results of the evaluation. There were two experiments performed in order to answer the three research questions. These research questions provided the results of the performance of ESA in TMS tasks. The results of the evaluation not only give us an understanding of how ESA can be applied in TMS tasks, but they are also useful for examining certain textual factors that may have an impact on the performance of translation memory systems.

5.1 Overview

The data and results generated from the experimental procedure will be presented in five sections. Section 5.2.1 is concerned with the details of three test collections. Sections 5.2.2 to 5.2.4 present the results from the ESA matching experiment. The results of research question A.1 confirm that ESA can be used in TMS tasks, and specify a possible threshold value. Threshold values of ESA similarity scores are then given based on different genre groups. In Section 5.2.3, the results of research question A.2 show that two types of translation suggestion can be retrieved by ESA: formally similar (FS) and conceptually related (CR) translation suggestions. In Section 5.2.4, the findings from research question B demonstrate how textual factors could affect the performance of ESA.

5.2 Overview of Test Collections and Results of Research Questions

Section 5.2.1 provides the textual statistics of the test collections used in the two experiments. From Section 5.2.2 to Section 5.2.4, each section provides the result of one research question.
5.2.1 Textual Statistics of Test Collections

In the first phase of the experiment, from steps 1 to 10, several tools were used to acquire the textual factors (such as, sizes, TTRs, ASLs) and the topic models of the newly created test collections. As stated in Section 4.2.2, the test collections used in the experiments were created specifically for the purpose of evaluating ESA in TMS tasks. These three test collections, i.e. SciAm, ACA and RN, were modified for the specific purpose of the experiments. Hence, details of these test collections relevant to the evaluation were not available in the first place. The relevancy of the details of the test collections is that they show the differences between the various genres from a statistical perspective. The details also include the textual statistics of each test collection.

The details of test collections are given below:

<table>
<thead>
<tr>
<th>Genres</th>
<th>ACA</th>
<th>RN</th>
<th>SciAm</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>363,395</td>
<td>127,705</td>
<td>202,673</td>
<td>693,773</td>
</tr>
<tr>
<td>Type</td>
<td>17955</td>
<td>13364</td>
<td>15346</td>
<td>46,665</td>
</tr>
<tr>
<td>TTR</td>
<td>4.941%</td>
<td>10.465%</td>
<td>7.572%</td>
<td>6.73%</td>
</tr>
<tr>
<td>ASL</td>
<td>21.963(^{10})</td>
<td>20.046</td>
<td>18.749</td>
<td>20.253</td>
</tr>
</tbody>
</table>

Table 5.1: Statistics from each of the three test collections

As Table 5.1 shows, the size (i.e. number of tokens) of the three test collections varies. The ACA test collection is the largest and the RN is the smallest. Due to its size, ACA would be a better choice than RN and ACA to be TM files, assuming these test collections were bilingual. As clarified in Section 4.1.1, bilingual versions are not necessary for the purpose of this study. Thus, ACA may have a higher possibility of achieving translation suggestions in the case of the experiments. However, ACA has the lowest TTR, meaning its choice of vocabulary is relatively small, particularly in comparison with RN.

With regard to ASL, ACA has the greatest ASL and SciAm the shortest. A likely reason for this may be seen in the fact that articles from SciAm had a wider range of readership, and thus the authors of the popular science articles may intentionally employ shorter sentences in order to make their texts accessible to a broader base of target readers. Texts from ACA are mainly

\(^{10}\) Values of ASL, as well as ESA similarity scores, SD and variation, were only corrected to three decimal places for the convenience of presentation.
written for professionals and but also for the general public, in particular for relatives of the crash victims. Thus, texts from ACA contain more serious technical information.

The topic models of the three test collections are given below:

<table>
<thead>
<tr>
<th>ACA</th>
<th>RN</th>
<th>SciAm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 00 wing</td>
<td>Topic 00 meeting</td>
<td>Topic 00 cancer</td>
</tr>
<tr>
<td>Topic 01 pressure</td>
<td>Topic 01 statement</td>
<td>Topic 01 galaxies</td>
</tr>
<tr>
<td>Topic 02 aircraft</td>
<td>Topic 02 market</td>
<td>Topic 02 past</td>
</tr>
<tr>
<td>Topic 03 rudder</td>
<td>Topic 03 added</td>
<td>Topic 03 theory</td>
</tr>
<tr>
<td>Topic 04 flight</td>
<td>Topic 04 economic</td>
<td>Topic 04 level</td>
</tr>
<tr>
<td>Topic 05 failure</td>
<td>Topic 05 rates</td>
<td>Topic 05 drug</td>
</tr>
<tr>
<td>Topic 06 engine</td>
<td>Topic 06 oil</td>
<td>Topic 06 data</td>
</tr>
<tr>
<td>Topic 07 flap</td>
<td>Topic 07 net</td>
<td>Topic 07 universe</td>
</tr>
<tr>
<td>Topic 08 damage</td>
<td>Topic 08 Japanese</td>
<td>Topic 08 researchers</td>
</tr>
<tr>
<td>Topic 09 system</td>
<td>Topic 09 prices</td>
<td>Topic 09 number</td>
</tr>
<tr>
<td>Topic 10 air</td>
<td>Topic 10 west</td>
<td>Topic 10 university</td>
</tr>
<tr>
<td>Topic 11 speed</td>
<td>Topic 11 exports</td>
<td>Topic 11 sun</td>
</tr>
<tr>
<td>Topic 12 control</td>
<td>Topic 12 official</td>
<td>Topic 12 human</td>
</tr>
<tr>
<td>Topic 13 captain</td>
<td>Topic 13 states</td>
<td>Topic 13 genes</td>
</tr>
<tr>
<td>Topic 14 aircraft</td>
<td>Topic 14 sugar</td>
<td>Topic 14 scientific</td>
</tr>
<tr>
<td>Topic 15 position</td>
<td>Topic 15 trade</td>
<td>Topic 15 cells</td>
</tr>
<tr>
<td>Topic 16 check</td>
<td>Topic 16 coffee</td>
<td>Topic 16 black</td>
</tr>
<tr>
<td>Topic 17 droop</td>
<td>Topic 17 gold</td>
<td>Topic 17 scientists</td>
</tr>
<tr>
<td>Topic 18 maintenance</td>
<td>Topic 18 bank</td>
<td>Topic 18 feathers</td>
</tr>
<tr>
<td>Topic 19 conditions</td>
<td>Topic 19 reserves</td>
<td>Topic 19 color</td>
</tr>
</tbody>
</table>

Table 5.2: Topic models of test collections

Although it is hard to have a minimum standard to specify a large TM, some other standards can be referred to some TMS distributors—for example, AutoSuggest function from SDL Trados—require users to have at least 10,000 translation suggestions to enable translation memory systems to predict the rest of the characters or letters of a word after users type the first few letters (SDL, 2014a). Hence, it is possible to say that a large TM may contain 10,000 translation suggestions from the perspective of enabling TMs to have an autocomplete function. If the ASL of a genre group is 12, then the size of the test collection should be approximately
120,000 words. Therefore, all three test collections are fairly large, as all the test collections met this standard.

The topic models are also useful for indicating what subjects are included in the test collections. For example, RN included topics such as ‘coffee, sugar, exports, and gold’. These topics can be used as ‘key words’ to provide an indication that articles about those topics might have a greater likelihood of yielding match sentences. Moreover, these three test collections had different TTRs and ASLs. The differences between the test collections may cause the difference in ESA similarity scores for retrieving valid translation suggestions. These three test collections are therefore suitable for the purposes of the experiments.

The following three sections will provide the results for each research question. The first two questions were answered by ESA matching experiments, and the third question was answered by experiment B.

5.2.2 Results for Research Question A.1

The aim of research question A.1 was to test whether the ESA algorithm could be used in TMS tasks and to compare its performance in retrieving translation suggestions using the Levenshtein distance method. The question is as follows:

**What is the possible range of ESA similarity scores that can give rise to potential translation suggestions?**

As mentioned in Section 4.1, each of these results consisted of a pair of query and translation suggestions with an ESA similarity score assigned by the ESA IR platform. The optimum threshold value was based on the ESA similarity scores of the results, and specifically on the average range of ESA similarity. The data produced in ESA matching experiments included 40 results for each test collection; hence, 120 results were produced in total. In other words, 120 valid matches of translation suggestions were found by ESA. As mentioned in Step 3 of the experiment procedure, all ESA similarity scores were converted from their original cosine values. Following the conversion of all the original values, ESA similarity scores of valid translation suggestions were comparable to those derived from the Levenshtein distance method.
This study only focuses on the ESA similarity scores of valid translation suggestions. As noted in Section 4.2.3.2, valid translation suggestions were manually selected from potential results and then were recorded in accordance with the definition of valid translation suggestions presented in Section 4.1.1. Invalid results are usually the sentences that appear to possess no similarity to potential queries. As explained in the third step of the ESA matching experiment, the validity of these results was checked manually, starting from the bottom results and ending with those at the top. If a valid translation suggestion was found, then it was recorded.

In most instances the ESA IR platform returned invalid results with low ESA similarity scores, as due to their size the test collections do not normally contain texts that can be closely matched with specific queries. In theory, there are three possible outcomes after running ESA IR platform each time:

1. No valid translation suggestion;
2. One valid translation suggestion;
3. Multiple valid translation suggestions;

Taken together, outcomes 2) and 3) represent approximately between 10% and 20% of the total number of queries sent to ESA IR. Given the the size of the test collections it was usually necessary to send between five and ten queries before an outcome 2) or 3) result was obtained. If a result is invalid, then it is not seen as a translation suggestion in the first place. Details of the invalid results were not retained, because to achieve a valid suggestion the ESA IR platform is not the only factor as one of the test collections also needs to contain similar text segments corresponding to potential queries. An assessment of the number of queries that have no valid translation suggestion is largely a matter of the information richness of test collections and casts no particular light on the potential of ESA in TMS tasks.

The following examples of query sentences and the full sets of results that they returned have been selected because their spreads of ESA similarity scores are fairly typical of how the ESA IR platform performed for each kind of outcome. Each example shows the rough range and dispersion of ESA similarity scores.

During the experiment, as stated above the most common occurrence was that one or other test collection did not contain any matched text segments at all, and all the results were assigned fairly low ESA similarity scores, as the ESA IR platform was instructed to provide the closest results possible. It is clear that none of the sentences are at all relevant to the potential query, and so were naturally not considered to be valid translation suggestions.
The example below is the result of a potential query tested by the SciAm test collection, and it represents a typical spread of ESA similarity for a set of invalid results. The query sentence was selected as it was considered possible that it may have retrieved suitably matching text segments from the SciAm test collection. The query sentence was as follows: ‘Yin and yang can be thought of as complementary (rather than opposing) forces that interact to form a dynamic system in which the whole is greater than the assembled parts’. However, this query can be considered unsuccessful in the sense that no valid results were returned. As we can see, the ESA similarity scores ranged from 2.09 to 1.803 in this case and were not widely spread out.

<table>
<thead>
<tr>
<th>No.</th>
<th>ESA Similarity Score</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.102</td>
<td>This aerosol direct climate forcing is uncertain by at least percent, in part because aerosol amounts are not well measured and in part because of their complexity.</td>
</tr>
<tr>
<td>2</td>
<td>2.011</td>
<td>Even the IPCC's minimum added forcing would cause dangerous anthropogenic interference with the climate system based on our criterion.</td>
</tr>
<tr>
<td>3</td>
<td>1.897</td>
<td>Frictional forces cannot do that; only tidal forces can.</td>
</tr>
<tr>
<td>4</td>
<td>1.887</td>
<td>If electrons and quarks are composite, this unification fails: the forces do not become equal</td>
</tr>
<tr>
<td>5</td>
<td>1.875</td>
<td>Frictional forces may still be important, however</td>
</tr>
<tr>
<td>6</td>
<td>1.868</td>
<td>The branes exert attractive forces on one another and occasionally collide.</td>
</tr>
<tr>
<td>7</td>
<td>1.866</td>
<td>For example, physicists believe that the weak nuclear force and the electromagnetic force were different aspects of a single electro weak force when the universe was hotter than is kelvins.</td>
</tr>
<tr>
<td>8</td>
<td>1.865</td>
<td>The opposite is true for the ekpyrotic scenario, in which the collision occurs when forces are at their weakest.</td>
</tr>
<tr>
<td>9</td>
<td>1.863</td>
<td>Whether other instinctual forces exist, such as a rough-and-tumble &quot;play&quot; system, is also being investigated.</td>
</tr>
<tr>
<td>10</td>
<td>1.862</td>
<td>Despite the large uncertainties, there is evidence that this estimated net forcing is approximately correct.</td>
</tr>
<tr>
<td>11</td>
<td>1.862</td>
<td>The forces are mediated by force particles: photons for electromagnetism, the W and Z bosons for the weak force, and gluons for the strong force.</td>
</tr>
<tr>
<td>12</td>
<td>1.858</td>
<td>Moreover, they are actively withheld from consciousness by a repressive force.</td>
</tr>
<tr>
<td>13</td>
<td>1.842</td>
<td>It is the propelling force and the order of each milestone that are under active debate.</td>
</tr>
<tr>
<td>14</td>
<td>1.832</td>
<td>When we extrapolate, we find that the strengths of these three forces become very similar but are never all exactly the same.</td>
</tr>
</tbody>
</table>
One revision suggests, for example, that the negative and cognitive symptoms may stem from reduced dopamine levels in certain parts of the brain, such as the frontal lobes, and increased dopamine in other parts of the brain, such as the limbic system.

Table 5.3.1: The typical spread of ESA similarity scores of outcome one

It should be noted that different queries may have different values, either higher or lower, depending on how close they are to the content of corresponding test collections. However, the general spread of ESA similarity scores for the first outcome was usually low and hardly ever greater than 0.5, based on the general estimation of outcome during the experiment. It is possible to suggest that the ESA similarity scores of translation suggestions are unlikely to be used if they are too low.

Secondly, most matched queries only received one valid translation suggestion back from ESA IR platform. Although a few queries do receive valid translation suggestion, the ESA IR platform can eventually supply valid translation suggestions if more potential queries were tested. A typical example of the second outcome is provided. The query was from the RN test collection: ‘Russia will maintain its long-standing military and economic support for Syrian President Bashar al-Assad.’

The valid translation suggestion was highlighted in a different colour from the top 15 results. This example indicates a typical spread of ESA similarity for the second outcome. As we can see below, overall, the typical spread of ESA scores varied slightly from query to query.

<table>
<thead>
<tr>
<th>No.</th>
<th>ESA Similarity Score</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.581</td>
<td>Russia was trying to build up its military presence in the region.</td>
</tr>
<tr>
<td>2</td>
<td>6.468</td>
<td>A military spokesman quoted by the official Iraqi news agency said other Iranian boats fled.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>5.192</td>
<td>But there was no official word on military movements apart from a comment by the government spokesman that the Greek navy was no longer in port.</td>
</tr>
<tr>
<td>4</td>
<td>3.833</td>
<td>Facilities being installed to protect the field include aircraft detection equipment anti-aircraft missiles housing for military personnel and helicopter landing pads the sources said</td>
</tr>
<tr>
<td>5</td>
<td>3.602</td>
<td>The unit makes valves and marine specialty equipment for the military.</td>
</tr>
<tr>
<td>6</td>
<td>3.391</td>
<td>Said Iranians used the platforms for military purposes and had fired on an American helicopter from the rigs earlier this month.</td>
</tr>
<tr>
<td>7</td>
<td>2.624</td>
<td>Mike Synar said today that while President Reagan is ready to use military force to protect Kuwait tankers in the Gulf the United States is ill-prepared at home to deal with a new energy crisis.</td>
</tr>
<tr>
<td>8</td>
<td>1.923</td>
<td>Spring sowing is off to a slow start with planting two to four weeks behind schedule in many areas of the Ukraine Byelorussia and Russia because the winter was unusually cold.</td>
</tr>
<tr>
<td>9</td>
<td>1.696</td>
<td>Doubts the fund would be able to fulfil its objectives citing the lack of widespread support.</td>
</tr>
<tr>
<td>10</td>
<td>1.675</td>
<td>The lower estimates are supported by the belief that crude runs increased and imports fell</td>
</tr>
<tr>
<td>11</td>
<td>1.485</td>
<td>Security Pacific will provide expertise in consumer and commercial lending as well as data processing support.</td>
</tr>
<tr>
<td>12</td>
<td>1.439</td>
<td>Diplomats said they expect three Soviet tankers initially to reinforce other flags already supporting Kuwait's tanker fleet.</td>
</tr>
<tr>
<td>13</td>
<td>1.383</td>
<td>The Canadians postponed the original July deadline at the request of the European Commission which Denmark approached for support.</td>
</tr>
<tr>
<td>14</td>
<td>1.332</td>
<td>The Dutch Central Bank intervened modestly to support the dollar with spot market transactions, dealers said.</td>
</tr>
<tr>
<td>15</td>
<td>1.311</td>
<td>Maciej said he sees crude oil supply outweighing demand and doesn't believe a recent OPEC production accord will continue to support prices.</td>
</tr>
</tbody>
</table>

Table 5.3.2: The typical spread of ESA similarity scores of outcome two
In principle, the ESA IR platform should rank the most relevant results at the top, and in this case, the very first result can be considered as a valid translation suggestion, and it has received very high ESA similarity score. Most results received ESA similarity scores that were significantly lower (ranging from 6.47 to 1.31), compared with the score of the valid translation suggestion (8.581). Although the second and third results also received ESA similarity scores that were significant, they were not relevant to the query. Therefore, their scores would be likely to be considered to fall outside the possible range of validity. Although the ESA similarity scores in the case of outcome two were usually higher than the results from outcome one, the similarity scores received by valid translation suggestions tended to be only slightly higher than those for the other, invalid results. Like outcome one, the precise ESA similarity score varied from one match to the next, but the three test collections tended to have their own respective ranges of scores, as will be presented later.

The third kind of outcome occurred occasionally. It reflects the well-known fact that any TMS can potentially retrieve more than one translation suggestion for a sentence being translated if the TM database contains multiple corresponding matches. Following the common practice of TMSs, each query-result pair will be treated as a separate piece of data. A typical example of this kind is presented as follows, based on the following query sentence from the ACA test collection: ‘At the rear of the passenger cabin they observed indications of fire above the ceiling panels’.

<table>
<thead>
<tr>
<th>No.</th>
<th>ESA Similarity Score</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.524</td>
<td>The evidence indicated that there had been no pre-crash or post-crash fire.</td>
</tr>
<tr>
<td>2</td>
<td>9.519</td>
<td>There was no pathological indication of an in-flight fire and no evidence that any of the victims had been injured by shrapnel from the explosion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>9.366</td>
<td>There were no indications of fire and the commander did not order an evacuation.</td>
</tr>
<tr>
<td>4</td>
<td>8.058</td>
<td>The results of this analysis detected traces of the explosive used in the electrically initiated squib, which indicated that the squib had fired.</td>
</tr>
<tr>
<td>5</td>
<td>7.474</td>
<td>The extent of the developing fire indicated that controlled flight could only have continued for a very short time.</td>
</tr>
<tr>
<td>6</td>
<td>7.312</td>
<td>However, with the ‘FIRE’ light still illuminated and indications of hydraulic failures from both tactile and warning systems, the co-pilot alerted the commander to a suitable nearby landing area.</td>
</tr>
<tr>
<td>7</td>
<td>7.161</td>
<td>There was plenty of fuel on the aircraft and there were no indications of a fuel leak.</td>
</tr>
<tr>
<td>8</td>
<td>7.160</td>
<td>The PF also indicated some preoccupation with the failure to capture the localiser.</td>
</tr>
<tr>
<td>9</td>
<td>7.159</td>
<td>No pre-accident damage or deterioration of these bearings was indicated.</td>
</tr>
<tr>
<td>10</td>
<td>7.159</td>
<td>No evidence of indication or of whether gyroscope was rotating.</td>
</tr>
<tr>
<td>11</td>
<td>7.158</td>
<td>There was no evidence to indicate that there was more than one explosive charge.</td>
</tr>
<tr>
<td>12</td>
<td>7.158</td>
<td>No evidence of indication, some signs that gyroscope was rotating.</td>
</tr>
<tr>
<td>13</td>
<td>7.157</td>
<td>This would have resulted in a cockpit indication of one graduation nose down trim.</td>
</tr>
<tr>
<td>14</td>
<td>7.156</td>
<td>The dynamic analysis commissioned by the AAIB indicated that this is a possible explanation of the subsequent bearing failure.</td>
</tr>
</tbody>
</table>
This was reinforced by the consistency of the indications given by the attitude indicator with the attitude initially determined from the overall examination of the aircraft wreckage.

Table 5.3.3: The typical spread of ESA similarity scores of outcome three

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>7.155</td>
</tr>
</tbody>
</table>

As we can see, the three results with the highest ESA similarity scores were considered valid translation suggestions in response to the query. In this case, the ESA similarity scores of the top three results for this kind of this outcome was not very different to that for the second kind. However, the fourth and fifth results also received ESA similarity scores that were significant, but they would be likely to be considered to fall outside the possible range of validity, and there is a small gap of ESA similarity scores between valid translation suggestions and invalid results. This situation occurred in both outcome two and three. In this example, the range of ESA similarity scores is from 7.15 to 9.52. This reflects the fact that the query has many potential results from this test collection, and suggests that some sentences can be similarly expressed in real world conditions. The range and the spread of ESA similarity scores are largely dependent on the response to individual queries and their corresponding test collections. As it will be demonstrated in Section 5.2.2.2, the lowest ESA score for valid suggestions varies from one test collection to another, while a particular valid translation suggestion may have a lower ESA similarity score in the context of its test collection than another, invalid suggestion in the context of its collection. With larger databases to draw on, the top suggestions in the case of outcomes 2) and 3) can be expected to be of a higher quality than is currently the case with the relatively small test collections that have been used in the experiments.

In general, in the context of TMS the threshold value cannot be seen as an absolute cut-off point between valid and invalid suggestions or as an absolute discriminator to identify translation suggestions, but rather as a baseline to indicate the general distribution of the ESA similarity scores of valid translation suggestions. As discussed in Section 3.1, the threshold value in this study is seen as a compromise between the recall and precision of retrieving valid translation suggestions. The detail of the analysis is presented in the following Section 5.2.2.

Very occasionally, the ESA IR platform could possibly assign an unusually high ESA similarity score to an invalid translation suggestion among other, valid translation suggestions. This occurred a few times during the experiment, and was seen as a technical fault of the current ESA IR platform. And for the same reason it was also possible that a small number of
valid translation suggestions could have been missed. However, these matters of recall and precision can be expected to diminish as the technology becomes more mature. As claimed in Chapter 1, this study does not aim to resolve the technical limitations of the actual ESA implementation. Therefore, the original data concerning the error rate was not retained. Consequently, the relevant valid translation suggestions were simply discarded whenever this occurred. The primary focus was the ESA similarity scores of the valid translation suggestions that were obtained.

Section 5.2.2.1 will show the performance of the Levenshtein distance method in measuring the valid translation suggestions found by ESA. Section 5.2.2.2 will present the statistics of the ESA similarity scores of the three test collections.

### 5.2.2.1 Performance of Levenshtein Distance Method

It was necessary to obtain an overview of the data produced by the ESA matching experiment in order to see how the Levenshtein distance method processes translation suggestions retrieved by the ESA method. In Step 4 of the ESA matching experiment, these valid translation suggestions were measured by the Levenshtein distance method.

The performance of the Levenshtein distance method is shown in Table 5.3 below:

<table>
<thead>
<tr>
<th></th>
<th>RN</th>
<th>SciAm</th>
<th>ACA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Highest Score</strong></td>
<td>50%</td>
<td>48%</td>
<td>71%</td>
</tr>
<tr>
<td><strong>Lowest Score</strong></td>
<td>12%</td>
<td>18%</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>35%</td>
<td>33.5%</td>
<td>37%</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>33.4%</td>
<td>33.6%</td>
<td>36.6%</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.106</td>
<td>0.090</td>
<td>0.121</td>
</tr>
</tbody>
</table>

Table 5.3.4: Statistics of Levenshtein distance scores of the valid translation suggestions

Table 5.3 represents the Levenshtein distance scores pertaining to the valid translation suggestions produced by the ESA method.

Although the Levenshtein distance method was occasionally able to recognise the results produced by ESA, most of the valid translation suggestions given by ESA were missed by the Levenshtein distance method. Only one valid translation suggestion achieved 71% of the
Levenshtein distance score, but most valid translation suggestions received lower scores. The mean results for RN, SciAm and ACA were 33.4%, 33.6% and 36.6%, respectively. The SDs of the Levenshtein distance scores of the three test collections were 0.106, 0.09 and 0.121. These low SDs suggest that the Levenshtein distance scores were not different from one another and that most scores were very low. All the results were much lower than the normal threshold value (70%). In cases where the Levenshtein distance method assigned lower scores for these results, this indicated that the Levenshtein distance method was not able to identify the valid translation suggestions given by the ESA method. As mentioned previously, all of the results collected in ESA matching experiment were valid translation suggestions.

In theory, it is possible that ESA can miss translation suggestions that can be retrieved by Levenshtein distance method if no suitable training database is available. To overcome this situation in an actual implementation, every translation suggestion can be measured by the Levenshtein distance method first and then by ESA as the Levenshtein distance method is much faster and does not need an external knowledge base. Considering the nature of the Levenshtein distance method, it is most likely to be used with repetitive texts. The similarity of those texts does of course not need to be measured again by ESA. This study does not aim to replace the Levenshtein distance method with ESA; more discussion of the mutually beneficial relationship between these two approaches can be found in Section 6.1.1.

5.2.2.2 Statistics of ESA Similarity Scores of the Three Test Collections

Comparing the performance of the Levenshtein distance method with the ESA method proved that ESA was able to retrieve some translation suggestions that were missed by Levenshtein distance method. After that, it was necessary to examine the ESA similarity scores. An overview of the data from ESA matching experiment, including means, medians and SDs from all three test collections, is given in the Table 5.4 below:
As mentioned in Section 4.1.2, both the mean and the median were considered, because of the large number of results. SDs were also used to indicate the spread of the ESA similarity scores.

In general, the ESA similarity scores varied noticeably between different test collections. However, the data also suggests that the ESA similarity scores showed more similarities with respect to ACA and RN. The medians for the ESA similarity scores were 9.368 (ACA), 9.143 (RN) and 7.839 (SciAm), and the means for the scores were 9.17 (ACA), 8.982 (RN) and 8.048 (SciAm). The score ranges were 8.107–9.99 (ACA), 7.183–9.966 (RN) and 6.518–9.925 (SciAm), respectively. The differences between the medians and means of the ESA similarity scores were much smaller for ACA and RN compared to SciAm.

For each test collection, the difference between the means and medians was relatively small. Those differences were 0.198 (ACA), 0.161 (RN), and -0.209 (SciAm) respectively. Though the medians for the ESA similarity scores were different, their means were relatively similar, especially in the cases of ACA and RN. This revealed that the valid ESA similarity scores for the test collections were not particularly affected by extremely high or low scores.

The top ranges of ESA similarity scores for ACA, RN and SciAm were 9.999, 9.966 and 9.925 respectively. Translation suggestions with high ESA similarity scores occurred in all test collections, and the upper score limits of the three test collections were similar. This suggests that all test collections had sentences very similar to the queries.
By contrast, the lowest ranges of ESA were noticeably different. The lowest ESA similarity score for ACA (8.107) was higher than that for RN (7.183) and SciAm (6.518). Their SDs were 0.615 (ACA), 0.794 (RN), and 1.024 (SciAm) respectively. Lower ESA similarity scores increased the SD of the results. Low SDs suggest that the ESA similarity scores were relatively concentrated in certain ranges. In contrast, higher SDs suggest that the spread of the ESA similarity scores was relatively wide. With respect to the ESA similarity scores of individual test collections, the scores tended to show more stability in terms of ACA and RN, of which SD was only 0.615 and 0.794 respectively, compared to 1.024 for SciAm. Thus, the medians of the ESA similarity scores for the three test collections were more affected by the spread of the scores, especially for SciAm.

Based on the upper and lower ranges and SD of the ESA similarity scores, the mean was shown to be more suitable as a benchmark score. Hence, it is best to use means in the case of a range of ESA similarity scores for valid translation suggestions to ensure the reliability of future analysis for the questions, A.2 and B. Thus, the mean of overall ESA similarity scores was the overall threshold value of 8.733 for ESA in TMS tasks. The overall threshold value will be used for the rest of the research.

The distribution of these scores over the test collections is displayed in Chart 5.1. The blue column represents SciAm, the red ACA and the green RN. The number at the top of each column represents the number of results that fall in the corresponding range. For example, the green column labelled number ‘8’ indicates that there are eight RN results that fall in 9.8 of the range of the ESA similarity scores.
The relationship between the ESA similarity scores and the number of results is regarded as an issue of ESA's preference ranges; that is, whether ESA was capable of processing a query of a certain length. This issue will be illustrated further in Section 5.2.4.

5.2.2.3 Summary of Results for Research Question A.1

Research question A.1 aims at evaluating whether ESA can be applied in TMS tasks by comparing its performance with the Levenshtein distance method, and examining the ESA similarity scores of valid translation suggestions. As a result, all the genre groups mentioned here offered a sufficient number of translation suggestions. The ESA similarity scores of valid translation suggestions ranged from 6.518 to 9.99. The data from ESA matching experiments demonstrated that ESA was able to retrieve some translation suggestions that were likely to be missed by the Levenshtein distance method. However, most of the valid translation suggestions found by ESA received low Levenshtein distance scores (34% on average).
The statistics of the ESA similarity scores were noticeably different for each test collection. According to the statistics, the medians were more easily changed by the extreme scores of each individual test collection. The threshold value of the ESA similarity scores will be kept for later discussion, as it synthesises a wider range of values.

5.2.3 Results for Research Question A.2

After ascertaining the general results related to the ESA similarity scores, the next research question was concerned with the translation suggestions that are typically retrieved by this ESA IR platform. Since the results were all valid translation suggestions, this can be done by observing the differences between the translation suggestions. Research question A.2 is reiterated below:

What generalisations can be made about the kinds of translation suggestion that are typically retrieved?

To answer this question, all 120 results were examined individually. ‘Depending on their formal and semantic characteristics, they were divided into two types, while the link between the ESA similarity scores and translation suggestions of the results was examined. The following section reports on the two types of translation suggestion: a) the ‘Formally Similar translation suggestion (FS)’ and b) the ‘Conceptually Related translation suggestion (CR)’. The features and potential uses of the two types of translation suggestions are also examined. Statistical information on the ESA similarity scores will also be given with a view to comparing the two types of translation suggestion.

5.2.3.1 Types of Translation Suggestion

The first type of translation suggestion is the FS translation suggestion. This type is distinguished by the presence of textual fragments that are identical or similar to fragments contained in the corresponding queries. Suggestions of this kind can potentially be used in a similar manner to exact or fuzzy matches produced by traditional translation memory systems, in that they could realistically be simply post-edited by end users. Because of the semantic
motivation of ESA, we will of course expect to find in a result sentence items such as synonyms and hyponyms, which of course have a meaning-based connection to items in the query. In the light of this an example of a FS translation suggestion is the following.

Example 1:

<table>
<thead>
<tr>
<th>Query</th>
<th>The cause of the braking loss could not be positively established.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>The cause of the failure could not be assessed.</td>
</tr>
</tbody>
</table>

As it would be a relatively simple matter for a TMS to identify a suggestion as an FS translation suggestion, such hits could be automatically presented to the end user alongside hits from a traditional TM and/or from an MT engine.

Clearly, some parts of the query and result sentences are repeated, such as ‘The cause of the..’, meanwhile we can also find that ‘braking loss’ is the hyponymy of the ‘failure’, and can therefore be directly used as a translation suggestion.

The second type is the CR translation suggestion; suggestions falling within this category are composed of text that could simply occur in the same context as the queries but that does not contain directly repeated or similar textual fragments. Given the fact that most translation suggestions are at the sentential level, it is natural that translators are usually offered help of both a lexical and a syntactical nature. According to the results obtained from the experiments, CR translation suggestions may also be of help to translators.

From this point of view, the benefits of CR translation suggestions are indirect and less obvious compared with traditional translation suggestions and FS translation suggestions. Despite the features of the CR translation suggestions, this kind of translation suggestion can potentially be used by translators who are aware of translation issues.

The following illustrations will be based on English to Chinese translation. This does not suggest that this study is only of relevance to the use of ESA in this translation direction, but is simply determined by my language proficiency, as a native speaker of Chinese with advanced proficiency in English. In an actual implementation, all the results would be in Chinese, because of the bilingual aligned nature of the reference material that would be used. The translation problems that occurred in queries were selected for the purpose of analysing the possible uses of CR translation suggestions; all these problems occur frequently in translation.
practice, while many of them can probably be solved by experienced translators without any use of reference material.

When translating texts, especially specialised texts, choice of vocabulary is one of the commonest problems. As noted on page 26, non-technical words can have specific meanings in specialised texts, but translators may need to translate them using specific terms in the target language. Translators need to be aware of the precise implications of using different terms.

Example 2:

<table>
<thead>
<tr>
<th>Query</th>
<th>The value of credits given cannot exceed either one-fifth of the combined capital and reserves of the bank itself or two-fifths of the value of the stake owned in the bank.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>A Fed spokesman said that all of the borrowings were made yesterday by fewer than half the banks.</td>
</tr>
</tbody>
</table>

The term ‘value of credits’ in the query can be difficult to translate, because the word ‘credit’ can have several different translations, an initial thought being ‘信用’, which represents the literal meaning of the English word. In Chinese, the word ‘credit’ is also used in several existing financial terms such as ‘credit card’ (信用卡) and ‘credit rating’ (信用评级). This may lead some translators to make the decision to use ‘信用’. However, this would not be the best possible translation. This is because ‘信用的价值’, the literal translation of the phrase ‘value of credits’, would not be the most appropriate rendering if the target audience consisted of financial professionals. This is more than a wording issue. First, the word ‘信用’ does not carry the full meaning understood by financial professionals in a Chinese context. Secondly, ‘价值’ is the only suitable translation for ‘value’ that collocates with ‘信用’, but this word is more often used to describe qualities rather than quantifiable characteristics. Hence, translators could be aware of a translation problem, but could find themselves unable to locate a satisfactory translation based on searching for ‘信用’ (‘credit’). The results are also likely to be terms that contain the word ‘credit’, such as ‘credit spread option’, ‘credit-linked note’, ‘credit default swap’, and ‘credit derivatives’.

With regard to the result sentence, which would be in Chinese in an actual implementation, translators could very possibly find inspiration from the word used to translate ‘borrowing’ (e.g. 借款, 借贷额). This would remind translators of a possible meaning of the word ‘credit’, and at the very least may give them a better sense of what to look for when consulting relevant
resources. With the assistance of the result sentence, translators may produce ‘信贷额’ (‘value of credits’), which would be a more appropriate translation for ‘the value of credits’ in the context described above. The word ‘额’ (value) collocates naturally with the expression ‘信贷’ (‘credits’) and it is also a perfect solution given the quantifiable nature of the information contained in the query sentence.

Translators can also find it is difficult to translate some non-technical vocabulary that occurs in specialised texts.

Example 3:

<table>
<thead>
<tr>
<th>Query</th>
<th>The serious incident occurred to an Airbus A319-111 aircraft operating a scheduled passenger flight between Spain and UK.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>The aircraft was privately owned and operated.</td>
</tr>
</tbody>
</table>

The word ‘operate’ can hardly be considered a technical term in this case, but it can have two translations, either ‘操作’ (use machines) or ‘运营’ (run a business). Either option could make sense in Chinese, so the meaning of the word is ambiguous. With regard to the result sentence, it could be from a similar context to that of the query sentence, and it also contains the word ‘operate’. Although the two sentences do not need to have the same meaning, the result sentence could disambiguate the meaning of ‘operate’ as it represents an example of how it was translated previously. In this way, the use of ESA suggestions may mean that translators who are in need of lexical inspiration may not need to consult thesaurus-like resources as they will potentially be able to receive indications of suitable synonyms automatically from within the TMS interface.

Besides linguistic hints, a CR translation suggestion may offer the translator some explanation of the information contained in the query, should that be needed, although the quality of such explanation is likely to vary from one suggestion to another. At the same time, it can also provide background information relevant for the text as a whole.

In practice, the distinction between linguistic hints and conceptual clarification is not cast-iron. If translators understand the general meaning of the texts and the background knowledge, it is easier for them to identify solutions for specific translation problems from within the ESA translation suggestions. Furthermore, the results from the experiments show that some results are better described as being transitional between FS and CR.
Example 4:

<table>
<thead>
<tr>
<th>Query</th>
<th>The US government has painted a bearish outlook for the prices of corn, wheat and soyabean.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>Although prices are now considerably lower, consumers and producers have been unable to agree on re-introduction.</td>
</tr>
</tbody>
</table>

The word ‘bearish’ in the query can have two meanings. The literal meaning of the English word is ‘looking or behaving like a bear’, but it does not make sense in this context, as the translation could be ‘熊一样的前景’ (‘an outlook as that of a bear’). However, the result sentence does explain some part of the query sentence in the clause ‘Although prices are now considerably lower’. Therefore, the suggestion could if necessary help the translator to a correct understanding of the meaning of the query sentence and the selection of an appropriate financial term such as ‘看空’ or ‘看跌’ (both of which closely correspond to the English word). From this perspective, the clause highlighted is similar to the relevant parts in the query sentence.

This example, although potentially able to help translators to come up with the right term, has still been considered as a CR translation suggestion, because it is offering a conceptual clarification and can occur in a similar context. But it also suggests that the distinction between FS and CR translation suggestions may not always be totally clear-cut in practical situations, as the uses to which translation suggestions are put are largely based on translators’ preferences and needs. For this reason, highly detailed lists of linguistic features that may identify each of the two types of translation suggestion are possibly unnecessary. However, for the sake of simplicity translation suggestions were only classified as belonging to one type or the other in this thesis, depending on what was considered to be the main kind of benefit they were offering to translators. The decision to use just two broad categories was taken in view of the relatively small number of translation suggestions that were included in the study.

In a sense, the assistance provided by CR translation suggestions on the lexical level functions in a similar manner to the traditional concordance search features available within many translation memory systems, as translators may discover the meaning of particular words from large databases, but in the case of CR translation suggestions the hits are potentially relevant on a semantic level and occur not only at the word level but also at that of entire segments. Indeed, a possible future implementation might present the translator with a collection of CR hits that could be quickly scanned for possible useful information. From this point of view, CR
translation suggestions simply offer an additional search function that can be seen potentially as a semantically-enhanced concordance search.

CR translation suggestions can also be used to help on a sentence level. When translating texts, translators may need to change the sentence structure to conform to target norms. Translators may be able to obtain a grammatical clarification from a pair of texts that do not even have to contain the same words.

Example 5:

<table>
<thead>
<tr>
<th>Query</th>
<th>Scientists have been expecting an increase in solar activity because the sun is moving into a more volatile period of an 11-year cycle in which <strong>its magnetic field reverses its orientation</strong>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>Deep inside the sun, these two rates are similar, and the magnetic field is able to organize itself on large scale</td>
</tr>
</tbody>
</table>

The query sentence does not contain any difficult lexical problems. However, it can be difficult to translate into Chinese. A translator can translate the highlighted clause as ‘太阳磁场反转了方向’ (BT: The sun’s magnetic field reverses its orientation) which follows the most common Subject-Verb-Object word order. But, the problem is that the sentence then becomes a separate sentence and is hard to fit into the larger sentence, as the translator has to use linking words to reconstruct the query sentence. It is relatively uncommon to use linking words in Chinese sentences. Consequently, the translation of the whole sentence may risk reading like a piece of translationese.

Translators can benefit from the clause ‘the magnetic field is able to organize itself on a large scale’ from the result sentence. Ideally, its proper Chinese translation suggestion could be ‘磁场可大规模自我调节’ (BT: the magnetic field is able to organize itself on a large scale)’. From the translated text, the translator may find a solution to the clause highlighted in the query, as these two clauses would most naturally be presented in the same Chinese grammatical order: Subject-Object-Verb. Both clauses contain a subject (i.e., ‘magnetic field’), and in this specific case the subject should be followed by object or object-verb in order to be grammatically correct in Chinese. The phrase ‘大规模自我调节’ (BT: ‘be able to organise itself on a large scale’), a noun phrase ending with the verb ‘调节’ (BT: organise), can be easier to fit into the larger sentence. Thus, the translator could produce a more natural and neat translation of the highlighted clause as ‘磁场方向反转’ on the basis of the hint suggested by the result.
The above examples reveal that CR translation suggestions could also be used to resolve issues of a linguistic nature that occur in translation. One text fragment can be used to suggest a possible translation for another text fragment in a similar context, rather than the two text fragments containing a high proportion of overlapping information. A possible explanation could be that texts from the same context are likely to display similar ways of organising sentence structure and word choice. However, limited by the number of examples available, this study does not intend to develop a comprehensive understanding of how translation suggestions from similar contexts are used.

What is more important is that CR translation suggestions make us aware of a possible new way of exploiting translation suggestions. CR translation suggestions should be seen as the broadest method of finding potentially useful translation suggestions. Most translation suggestions provided by a TMS need to be post-edited by human translators. Translators may accept or reject the translation suggestions, or just produce a new translation from scratch. From the practical aspect, there are no bad or good translation suggestions. Thus, CR translation suggestions can be seen as an additional resource that translators may of course choose not to use at all. The development of a precise filter algorithm would be a likely next stage in the application of ESA techniques in TMS tasks, although at this stage of the study ESA cannot distinguish between FS and CR translation suggestions, nor can there be a guarantee that all CR translation suggestions will be of use to translators.

5.2.3.2 Statistical Information on the Explicit Semantic Analysis Similarity Scores for the Two Types of Translation Suggestion

After identifying the two types of translation suggestion, it was necessary to examine the link between ESA similarity scores and types of translation suggestion. Tables 5.5.1 and 5.5.2 below show that the chance of retrieving FS as opposed to CR translation suggestions was higher in SciAm than the cases of ACA (27 vs. 13) and RN (28 vs. 12). There was an equal
chance of retrieving the two translation suggestion types in the case of SciAm (20 vs. 20). The details of the two translation suggestion types are shown below:

<table>
<thead>
<tr>
<th></th>
<th>ACA</th>
<th>RN</th>
<th>SciAm</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Results</strong></td>
<td>27</td>
<td>28</td>
<td>20</td>
<td>75</td>
</tr>
<tr>
<td><strong>Lowest ESA Similarity Score</strong></td>
<td>8.107</td>
<td>7.183</td>
<td>6.521</td>
<td>7.270</td>
</tr>
<tr>
<td><strong>Highest ESA Similarity Score</strong></td>
<td>9.990</td>
<td>9.918</td>
<td>9.925</td>
<td>9.944</td>
</tr>
<tr>
<td><strong>Average Query Length</strong></td>
<td>17&lt;sup&gt;11&lt;/sup&gt;</td>
<td>16</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.615</td>
<td>0.780</td>
<td>1.047</td>
<td>-</td>
</tr>
<tr>
<td><strong>Means</strong></td>
<td>9.252</td>
<td>8.223</td>
<td>8.169</td>
<td>8.548</td>
</tr>
</tbody>
</table>

Table 5.5.1: Details of FS translation suggestions

<table>
<thead>
<tr>
<th></th>
<th>ACA</th>
<th>RN</th>
<th>SciAm</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of Results</strong></td>
<td>13</td>
<td>12</td>
<td>20</td>
<td>45</td>
</tr>
<tr>
<td><strong>Lowest ESA Similarity Score</strong></td>
<td>8.199</td>
<td>7.369</td>
<td>6.518</td>
<td>7.362</td>
</tr>
<tr>
<td><strong>Highest ESA Similarity Score</strong></td>
<td>9.960</td>
<td>9.967</td>
<td>9.467</td>
<td>9.798</td>
</tr>
<tr>
<td><strong>Average Query Length</strong></td>
<td>17</td>
<td>16</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.528</td>
<td>0.772</td>
<td>1.024</td>
<td>-</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>9.000</td>
<td>9.354</td>
<td>7.927</td>
<td>8.760</td>
</tr>
</tbody>
</table>

Table 5.5.2: Details of CR translation suggestions

<sup>11</sup> Values of query length and average translation suggestion length were displayed as integers.
As has been illustrated in the last section, the means rather than medians were used as threshold values for retrieving valid translation suggestions. When comparing the CR translation suggestions, it was found that the means of the ESA similarity scores for FS translation suggestions were closer to the overall threshold values. As Table 5.4 shows, the overall threshold values of ACA, RN and SciAm were 9.17, 8.982, and 8.048 respectively. The difference between the means of FS translation suggestions for each group and their corresponding overall threshold values were -0.082 (ACA), 0.759 (RN) and -0.121 (SciAm), while the difference between the means of CR translations for each test collection were 0.17 (ACA), −0.372 (RN) and 0.121 (SciAm). These figures showed that threshold values of ESA similarity scores for the CR translation suggestions were slightly lower than for the FS translation suggestions, except in RN. But the absolute values of the difference between the overall threshold values for RN and SciAm in FS translation suggestions (|0.759| vs. |-0.372| for RN, |-0.121| vs. |0.121| for SciAm), were larger than or equal to those for CR translation suggestions. This suggests that the threshold values of CR translation suggestions were closer to the overall threshold values of the two types taken together.

As can be seen from Tables 5.5.1 and 5.5.2, CR translation suggestions occurred much less frequently than FS suggestions. The ESA similarity scores for the two types fall within very similar parameters. In other words, in terms of ESA scores, CR cannot be seen as a ‘worse’ type than FS, but they are both consequences of the semantic motivation of the ESA retrieval method.

### 5.2.3.3 Summary of Results for Research Question A.2

Research Question A.2 is proposed in order to find generalisations regarding the kinds of translation suggestion that are retrieved by the method. More interestingly, there were two types of translation suggestion that were typically retrieved by the ESA IR platform. The FS translation suggestions could be used in the same way as the traditional fuzzy matching, while the CR translation suggestions could mainly be used for conceptual clarification of the subject, providing help that was either context-specific or perhaps relevant for an understanding of the text as a whole, or else as a source for translation suggestions on the lexical level, which it is
believed may be of particular use in the case of translation between languages that are not closely related to each other.

Although the ESA similarity scores of FS translation suggestions were very close to the overall threshold values, it was generally difficult to identify a translation suggestion as either FS or CR translation suggestion on the basis of the ESA similarity scores. Translators are supposed to make their own judgement regarding possible use of translation suggestions, just as they do with translation suggestions retrieved by the Levenshtein distance method.

The results of A.1 and A.2 have provided detailed statistics about the ESA similarity scores of the valid translation suggestions from the three test collections. According to the statistical details of ESA matching experiment, translation suggestions with higher ESA similarity scores were not invariably more useful than those with lower ESA similarity scores (that were, however, still above the threshold value), even though there was a strong tendency for them to be so. Considering the different textual statistics of the test collections displayed in Section 5.2.1, some textual factors of the genre may play a more significant role than the similarity of topics in the performance of ESA. In the next section, the impact of genres on the performance of ESA will be presented.

5.2.4 Results of Research Question B

After having gained a comprehensive understanding of the ESA similarity scores, it was then necessary to find why the performance between the different test collections differed. As mentioned in Section 4.1.4.3, the 'performance' was twofold, including the threshold values and lengths of valid translation suggestions. In examining this difference, it was reasonable to assume that certain textual factors in the test collections affected the threshold values of ESA. Two aspects of genre were considered: lexical complexity and length of sentences (including queries and sentences) in the test collections. Lexical complexity was quantified as TTR and the length of sentences included the ASL and the length of queries. This lead to research question B:
How do the following textual factors affect the performance of ESA: query length, type/token ratio (TTR), size, and average sentence length (ASL) of the genre group?

The results of research question B are also based on ESA matching experiments. Section 5.2.4.1 analyses the impact of TTR, ASL, and the size of the test collection on the ESA performance. Section 5.2.4.2 analyses the change of QTR. The first section mainly concerns the textual factors directly related to the test collections and the impact on the threshold values of ESA. The second section focuses on the ratio between queries and translation suggestion length as a measure of translators’ workload.

5.2.4.1 Impact of Type/Token Ratio, Average Sentence Length, and Size of the Test Collections on the Performance of Explicit Semantic Analysis

In this section, the impact of several textual factors related to the test collections on ESA similarity scores was examined. ESA similarity scores were further divided into three groups: overall translation suggestion, the FS translation suggestion (FS) and the CR translation suggestion (CR).

The relationship between TTR, ASL of the test collections and the ESA similarity scores on the three test collections is summarised in Table 5.6:

<table>
<thead>
<tr>
<th></th>
<th>ACA</th>
<th>RN</th>
<th>SciAm</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTR</td>
<td>0.049</td>
<td>0.105</td>
<td>0.076</td>
</tr>
<tr>
<td>ASL</td>
<td>21.963</td>
<td>20.046</td>
<td>18.749</td>
</tr>
<tr>
<td>Threshold values (Overall)</td>
<td>9.17</td>
<td>8.982</td>
<td>8.048</td>
</tr>
<tr>
<td>Threshold value (FS)</td>
<td>9.252</td>
<td>8.223</td>
<td>8.169</td>
</tr>
<tr>
<td>Threshold value (CR)</td>
<td>9</td>
<td>9.354</td>
<td>7.927</td>
</tr>
<tr>
<td>SD (Overall)</td>
<td>0.615</td>
<td>0.794</td>
<td>1.024</td>
</tr>
<tr>
<td>SD (FS)</td>
<td>0.615</td>
<td>0.78</td>
<td>1.047</td>
</tr>
<tr>
<td>SD (CR)</td>
<td>0.528</td>
<td>0.772</td>
<td>1.024</td>
</tr>
<tr>
<td>Query Length (Overall)</td>
<td>16</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>Query Length (FS)</td>
<td>17</td>
<td>16</td>
<td>15</td>
</tr>
</tbody>
</table>
Table 5.6: Relationship between TTR, ASL, size of test collections and the ESA similarity scores

<table>
<thead>
<tr>
<th>Query Length (CR)</th>
<th>17</th>
<th>16</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median (Overall)</td>
<td>9.368</td>
<td>9.143</td>
<td>7.839</td>
</tr>
<tr>
<td>Median (FS)</td>
<td>9.394</td>
<td>9.058</td>
<td>7.993</td>
</tr>
<tr>
<td>Median (CR)</td>
<td>8.665</td>
<td>9.174</td>
<td>7.866</td>
</tr>
</tbody>
</table>

First, the link between the TTR and the ESA similarity scores was examined. The ACA test collection had the lowest TTR (4.941%), SciAm had the second highest (7.572%) and RN had the highest (10.465%).

There was a possibility that TTR affected the threshold value of ESA similarity scores, especially in the case of the semantically similar types of translation suggestion. ACA, which had the lowest TTR, achieved the highest threshold value of the ESA similarity scores compared to the other test collections, although for the CR translation suggestion, the threshold value of RN was 9.354 (vs. 9.0 for ACA) and the median was 9.174 (vs. 8.665 for ACA). As stated in Section 4.1.4.3, a low TTR suggests that the vocabulary of a test collection was certain and predictable compared to a test collection with a higher TTR. Importantly, the threshold value could be relatively higher in a test collection with a low TTR. But, the threshold values for SciAm were lower than for RN, although SciAm had a lower TTR than RN. Therefore, TTR was not a dominant factor affecting the threshold values for other genres.

However, the effect of the SD on the ESA similarity scores could not be ignored. Lower median values may have led to higher SD values. For example, the medians of the three test collections were 9.368 (ACA), 9.143 (RN) and 7.839 (SciAm), while their SDs were 0.615 (ACA), 0.794 (RN) and 1.024 (SciAm). The SD of SciAm was highest of all. Lower SDs also decreased the threshold value of SciAm (i.e., 8.048). However, any correlation between the TTR and SD cannot be identified in the current data.

The size of a test collection was the second potential textual factor that could affect ESA performance, but it may not relate to any other factors or to the performance of ESA. It was known from the data in Section 5.2.1 that the sizes of the test collections were 363,395 (ACA), 127,705 (RN) and 202,673 words (SciAm). SciAm was the second largest test collection, but
its query length was the longest and its ASL was the shortest. RN was the smallest test collection, but its threshold value was similar to that of ACA, which was the largest test collection. This demonstrates that it was not likely that the size of the test collection was a factor in the performance of ESA.

The ASL of a test collection is another factor that can affect ESA performance. The ASL of ACA was the longest (21.963), while RN fell in the middle (20.046) and SciAm had the lowest (18.749).

The overall threshold values of these test collections were 9.17 (ACA), 8.982 (RN) and 8.048 (SciAm). This suggests that a higher ASL might have resulted in higher threshold values and lower SD values. It could also be seen from the data that the ASLs correlated with SDs. In general, a higher ASL caused a lower SD. For example, the ASL of ACA was 21.963 and its SD was 0.6154. The correlation between the ASL and the ESA similarity scores and SD was consistent for all test collections.

However, a low SD may be the indication that the ESA similarity scores are high enough to be valid for ACA. Although higher threshold value means a higher requirement for query matching and reduces the number of translation suggestions obtained, it offers the advantage that translation suggestions are more predictable for such test collection(s). This was different from RN and SciAm as in the case of these two test collections, the ESA similarity scores could be relatively low and still be used, which can be seen from the medians of the ESA similarity scores for ACA, RN and SciAm (9.368, 9.143 and 7.839 respectively).

The ASL of the test collections did not affect query lengths. The differences in ASL and query lengths between ACA and RN were not major. Their ASLs were 21.963 and 20.046, and their average query lengths were both 16. However, SciAm, the group with the lowest ASL, had the longest queries, particularly for CR translation suggestions (22). The query length of CR translation suggestions was significantly longer in the SciAm group than in the other two groups. It suggests that ESA provided more CR translation suggestions with longer queries. The impact of query lengths on the performance of ESA will be discussed in detail in the next section.
5.2.4.2 Query Lengths and the Performance of Explicit Semantic Analysis

As stated in Section 4.1.2, QTR was employed to indicate the lengths of translation suggestions regarding the workload of translators.

All the queries were divided equally, into different ranges of query length, and by their QTRs. The relationship between query length, QTR and the ESA similarity scores of each test collection is shown in Tables 5.7.1, 5.7.2 and 5.2.7.3. Each table presents the distribution of query lengths with their ESA similarity scores and QTRs. The query lengths from the three test collections ranged from 7 to 32 words. This range was subdivided into values of 7-8, 9-10, etc. For example, the value of column range 9-10 was 3 in Table 5.7.1. This implies that there were three queries that were either nine or ten words in length in ACA test collection. The value of the ATSL (average translation suggestion length) was 14, which suggests that the average word number of the translation suggestions was 14 in ACA test collection, corresponding to range 7-8. The QTR was calculated by dividing the mid value of the range by the ATSL, giving a result of 0.536 (i.e. 7.5/14 ≈ 0.536). The purpose of subdividing the query lengths rather than using distinct data points is to make the results more accessible and readable.

<table>
<thead>
<tr>
<th>ACA</th>
<th>Mean</th>
<th>ATSL</th>
<th>QTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Range (7~32)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7~8</td>
<td>9.66947</td>
<td>14</td>
<td>0.536</td>
</tr>
<tr>
<td>9~10</td>
<td>9.776</td>
<td>18.667</td>
<td>0.509</td>
</tr>
<tr>
<td>11~12</td>
<td>9.057</td>
<td>14.4</td>
<td>0.799</td>
</tr>
<tr>
<td>13~14</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15~16</td>
<td>9.051</td>
<td>15.846</td>
<td>0.978</td>
</tr>
<tr>
<td>17~18</td>
<td>9.116</td>
<td>10.2</td>
<td>1.716</td>
</tr>
<tr>
<td>19~20</td>
<td>9.074</td>
<td>17</td>
<td>1.147</td>
</tr>
<tr>
<td>21~22</td>
<td>9.045</td>
<td>24</td>
<td>0.896</td>
</tr>
<tr>
<td>23~24</td>
<td>8.871</td>
<td>13</td>
<td>1.808</td>
</tr>
<tr>
<td>25~26</td>
<td>9.385</td>
<td>11.5</td>
<td>2.217</td>
</tr>
<tr>
<td>27~28</td>
<td>9.46685</td>
<td>10</td>
<td>2.750</td>
</tr>
<tr>
<td>29~30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>31~32</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.7.1: Relationship between QTR and the ESA similarity scores; query length from the ACA test collection
<table>
<thead>
<tr>
<th>Query Range (7~32)</th>
<th>Number</th>
<th>Mean</th>
<th>SL</th>
<th>QTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>7~8</td>
<td>1</td>
<td>9.966763</td>
<td>19</td>
<td>0.395</td>
</tr>
<tr>
<td>9~10</td>
<td>2</td>
<td>8.591</td>
<td>20.5</td>
<td>0.463</td>
</tr>
<tr>
<td>11~12</td>
<td>6</td>
<td>9.096</td>
<td>14.333</td>
<td>0.802</td>
</tr>
<tr>
<td>13~14</td>
<td>9</td>
<td>9.098</td>
<td>15.1111</td>
<td>0.893</td>
</tr>
<tr>
<td>15~16</td>
<td>9</td>
<td>9.092</td>
<td>14.111</td>
<td>1.098</td>
</tr>
<tr>
<td>17~18</td>
<td>1</td>
<td>9.1711472</td>
<td>20</td>
<td>0.875</td>
</tr>
<tr>
<td>19~20</td>
<td>6</td>
<td>8.966</td>
<td>12.333</td>
<td>1.581</td>
</tr>
<tr>
<td>21~22</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23~24</td>
<td>3</td>
<td>8.411</td>
<td>19.333</td>
<td>1.216</td>
</tr>
<tr>
<td>25~26</td>
<td>1</td>
<td>9.0160595</td>
<td>13</td>
<td>1.962</td>
</tr>
<tr>
<td>27~28</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>29~30</td>
<td>1</td>
<td>7.36928</td>
<td>13</td>
<td>2.269</td>
</tr>
<tr>
<td>31~32</td>
<td>1</td>
<td>9.62485</td>
<td>18</td>
<td>1.750</td>
</tr>
</tbody>
</table>

Table 5.7.2: Relationship between QTR and the ESA similarity scores; query length from the RN test collection

<table>
<thead>
<tr>
<th>SciAm Query Range (7~32)</th>
<th>Number</th>
<th>Mean</th>
<th>ATSL</th>
<th>QTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>7~8</td>
<td>1</td>
<td>9.966763</td>
<td>19</td>
<td>0.395</td>
</tr>
<tr>
<td>9~10</td>
<td>2</td>
<td>8.591</td>
<td>20.5</td>
<td>0.463</td>
</tr>
<tr>
<td>11~12</td>
<td>6</td>
<td>9.096</td>
<td>14.333</td>
<td>0.802</td>
</tr>
<tr>
<td>13~14</td>
<td>9</td>
<td>9.098</td>
<td>15.1111</td>
<td>0.893</td>
</tr>
<tr>
<td>15~16</td>
<td>9</td>
<td>9.092</td>
<td>14.111</td>
<td>1.098</td>
</tr>
<tr>
<td>17~18</td>
<td>1</td>
<td>9.1711472</td>
<td>20</td>
<td>0.875</td>
</tr>
<tr>
<td>19~20</td>
<td>6</td>
<td>8.966</td>
<td>12.333</td>
<td>1.581</td>
</tr>
<tr>
<td>21~22</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23~24</td>
<td>3</td>
<td>8.411</td>
<td>19.333</td>
<td>1.216</td>
</tr>
<tr>
<td>25~26</td>
<td>1</td>
<td>9.0160595</td>
<td>13</td>
<td>1.962</td>
</tr>
<tr>
<td>27~28</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>29~30</td>
<td>1</td>
<td>7.36928</td>
<td>13</td>
<td>2.269</td>
</tr>
<tr>
<td>31~32</td>
<td>1</td>
<td>9.62485</td>
<td>18</td>
<td>1.750</td>
</tr>
</tbody>
</table>

Table 5.7.3: Relationship between QTR and the ESA similarity scores; query length from the SciAm test collection

These three tables indicate that some ranges had more results compared to others and that the means of the ESA similarity scores were also different. Some ranges retrieved just one result,
which was not representative of any statistical feature for these cases: for example, query ranges 7-8, 17-18, and 29-30 in RN test collection.

A different perspective of the data is that the relationship between QTRs and ranges of query lengths is also given. The distribution of QTRs of each test collection is displayed in Chart 5.2.

As Chart 5.2 shows, the QTRs also increased in general, as the query range increased. Shorter ranges, such as 7~8 and 9~10, generally had lower QTRs. In this study, 2 was treated as the threshold QTR value, as once a QTR is > 2, the word number in the translation suggestion may be only half that in the query. So, translators probably feel uncomfortable to read more words when QTRs > 2. There were four ranges in the three test collections that had QTRs higher than 2: ranges 25~26 and 29~30 for ACA and range 29~30 for RN and SciAm.

The distribution of the results in each query range and the trend of QTRs demonstrated that the ESA IR platform was particularly suitable for retrieving certain query ranges. Some ranges of query showed a significantly higher number of results than others, and QTRs increased significantly for ranges of query longer than 24 words. Translators would be able to receive less information with longer queries.
5.2.4.3 Optimum Ranges of Query

It is possible that ESA is more capable of processing certain ranges of query, i.e., that optimum ranges exist. Thus, the following table was compiled, here including only the ranges having the two largest numbers of results, and these ranges are seen as representative ranges.

<table>
<thead>
<tr>
<th>ACA</th>
<th>Number</th>
<th>Mean</th>
<th>ATSL</th>
<th>QTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range (7~32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11~12</td>
<td>5</td>
<td>9.057</td>
<td>14.4</td>
<td>0.799</td>
</tr>
<tr>
<td>15~16</td>
<td>13</td>
<td>9.051</td>
<td>15.846</td>
<td>0.978</td>
</tr>
<tr>
<td>17~18</td>
<td>5</td>
<td>9.116</td>
<td>10.2</td>
<td>1.716</td>
</tr>
<tr>
<td>RN</td>
<td>Number</td>
<td>Mean</td>
<td>ASL</td>
<td>QTR</td>
</tr>
<tr>
<td>Range (7~32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11~12</td>
<td>6</td>
<td>9.096</td>
<td>14.333</td>
<td>0.802</td>
</tr>
<tr>
<td>13~14</td>
<td>9</td>
<td>9.098</td>
<td>15.1111</td>
<td>0.893</td>
</tr>
<tr>
<td>15~16</td>
<td>9</td>
<td>9.092</td>
<td>14.111</td>
<td>1.098</td>
</tr>
<tr>
<td>19~20</td>
<td>6</td>
<td>8.966</td>
<td>12.333</td>
<td>1.581</td>
</tr>
<tr>
<td>SciAm</td>
<td>Number</td>
<td>Mean</td>
<td>ASL</td>
<td>QTR</td>
</tr>
<tr>
<td>Range (7~32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15~16</td>
<td>8</td>
<td>8.285</td>
<td>15.75</td>
<td>0.98413</td>
</tr>
<tr>
<td>19~20</td>
<td>7</td>
<td>8.791</td>
<td>13.57</td>
<td>1.43699</td>
</tr>
</tbody>
</table>

Table 5.8: Relationship between QTR and the ESA similarity scores; query length from representative ranges with the largest numbers of results

Table 5.8 clearly shows that range 15~16 had the best chance of producing translation suggestions. The largest range in the case of ACA was 15~16, with 13 results, and ranges 11~12 and 17~18 were the second largest, each containing five results. In the case of RN, ranges 13~14, and 15~16 achieved nine results and each ranges 11~12, 19~20 were the second largest ranges, each containing six results. In the case of SciAm, range 15~16 was the largest range, with eight results. The second largest ranges were 19~20 with seven results. Therefore, range 15~16 is the optimum range of ESA, because it is in the top range for all three test collections.
5.2.4.4 Summary of the Results for Research Question B

Research question B aims at examining how textual factors affect the performance of ESA. The data analysis revealed that the ASLs have affected the performance of ESA. Higher ASLs increased the threshold values of the ESA similarity scores and were negatively correlated with the SDs. TTR was linked to the ESA similarity scores of all test collections. However, the TTR was not a highly influential factor affecting the ESA similarity scores. TTR also had a very weak link to SD, which indicates the spread of the ESA similarity scores. The size of the test collections were not related to any other textual factors or any aspect of the performance of ESA.

The evidence is clear that query lengths are positively correlated with QTRs. Consequently, the translation suggestions of some exceptionally long queries were relatively short. This showed that ESA had a strong preference for queries ranging between 15 and 16.

5.2.5 Limitations of the Evaluation

Two limitations are worth noting. First, clearer conclusions may have been produced if larger quantities of results had been collected. However, it was unrealistic to implement larger-scale experiments given the limitations of this PhD research. The quantity of results in this study is primarily designed to ensure a valid conclusion regarding the possibility of using ESA for TMS tasks. Possible directions for further research in the context of a larger-scale project are suggested in Section 7.2.

Secondly, the results produced by the evaluation do not establish a causal relationship between the textual factors of the text genre groups and the performance of ESA. In fact, the evaluation was not designed to establish such a link in the first place. Ultimately, developers will control how ESA assigns semantic similarity scores across different texts. The evaluation can only show how particular textual factors affect the performance of ESA in TMS tasks. The further analysis of the performance of ESA will be discussed in Sections 6.1.3 and 6.1.4. Some suggestions for optimising the performance of ESA will also be given.
5.3 Conclusion

This chapter presents the results for the research questions based on the data from the ESA matching experiment. The experiment involved the ACA, RN and SciAm test collections. As mentioned in Section 4.2, several tools, as well as the ESA IR platform, were used to provide the necessary information and data for the experiment.

ESA matching experiments provided the data for research questions A.1, A.2 and B. ESA proved to be able to retrieve translation suggestions that were missed by the Levenshtein distance method. The data also shows that different genres required different threshold values and that the ESA similarity scores of valid translation suggestions were distributed in certain ranges. All of these results were further divided into the following two types: FS and CR. ESA performance was more closely related to ASL than to any of the other textual factors. Queries that ranged 15~16 words had greater chances of retrieval by ESA. The TTR also had an impact on the ESA similarity scores in two test collections. However, the size of the test collection was not a factor affecting ESA performance.

According to the SDs of the three test collections, the median and mean of the ESA similarity scores and the query lengths, RN and ACA were far more similar to one another than they were to SciAm. However, from the human translator's perspective, the topics in ACA (technical reports) and SciAm (popular science articles) have more in common with one another than they do with the financial news.
Chapter 6 Discussion

This chapter discusses several issues relating to the results of the evaluation. The relevance of the significant findings of the research to translation memory systems is highlighted. Based on the performance of ESA in the experiments, I also outline some bottlenecks in the use of ESA or similar techniques in translation memory systems. Finally, the perspective of translation memory systems in the future is also discussed. It suggests that the development of translation memory systems will rely on the support of large databases such as Wikipedia to enable their semantic measure function to strengthen their core function. With the involvement of large online databases, more and more translation memory systems are likely to use Software as a Service (SaaS) model, such as Wordfast Anywhere, but may also allow more flexibility for developers to customise translation memory systems. Plenty of potential functions may be available for both translation software and that used by other industries.

6.1 Significant Findings and Their Relevance to Translation Memory Systems

This section covers four significant findings of the evaluation. Each of these represents a distinct aspect of ESA performance. The next section will illustrate these important findings and their relevance to translation memory systems and future research.

6.1.1 Comparing Explicit Semantic Analysis and the Levenshtein Distance Method

According to ESA matching experiments, virtually all of the valid translation suggestions retrieved by ESA received very low fuzzy matching scores when measured by the Levenshtein distance method. As a result, these translation suggestions were not found by the Levenshtein distance method. As stated in Chapter 1, the core function of translation memory systems is to provide translation suggestions, so it seems that ESA has a useful role to fulfil here where deployed alongside Levenshtein-based matching. Compared with the Levenshtein distance method, the most important advantage of ESA is that it enables translation memory systems to
match queries with translation suggestions on the basis of semantic similarity. Consequently, queries do not need to have the exact same words in order to be matched by translation memory systems.

ESA is also able to match longer queries, because it considers semantic similarity rather than string similarity. As mentioned in Section 3.1.1, the Levenshtein distance method assigns penalty scores for bridging any differences between two text segments. It is unlikely for traditional translation memory systems to match two sentences of significantly different lengths, because high penalty scores would be assigned to deletions. ESA increases the likelihood of discovering useful translation suggestions, which is a big advantage over the Levenshtein distance method. Changing sentences may vary the ESA similarity scores compared with the Levenshtein distance method. First, ESA computes similarity scores based on the Wikipedia features of two sentences, i.e. if both sentences contain concepts that correspond to the same semantic space. Secondly, as some research teams have pointed out, ESA is a variation of the GVSM model (Gottron et al., 2011). Technically, the vectors of two sentences should be normalised. Normalisation means that two vectors have been mathematically converted to two unit vectors with the same length and direction (Manning, 2008:p. 121-3).

String similarity measures and semantic similarity measures are two methods of defining translation suggestions. Different definitions of translation suggestions imply that the usage of the two kinds of translation suggestion is also different. The Levenshtein distance method recognises translation suggestions from the perspective of minimising editing operations; therefore, the Levenshtein distance method measures the operations of editing sentences into segments of a translated text. String similarity measured by the Levenshtein distance method is more like a measure to indicate how many keystrokes the translator needs to make in order to bring the translated segment into an acceptable form, rather than a direct indication of the overall potential usefulness of translation suggestions. A Levenshtein translation suggestion needs to share at least one character string with the query sentence that it is intended to be assisting with. Hence, translators will typically reuse part of the translation suggestion and then translate the remaining part. Some translation memory systems, such as Déjà Vu, also have a ‘pre-translate’ function, in which the translation memory system inserts all previously stored similar translation suggestions into the appropriate positions (Kilgray, 2014; Atril, 2011). Hence, translators manually translate the rest of the text while contributing new translated
material. As a result, the Levenshtein distance method is mainly suitable for highly repetitive texts.

By contrast, ESA adopts a different perspective on the concept of the translation suggestion: it measures the semantic similarity of two texts or text segments. It is not in any way ‘better’ than Levenshtein distance method; it is just that the nature of many of the translation suggestions that it is likely to retrieve is fundamentally different from those that would be found by the other method. A translation suggestion’s ESA similarity score shows how similar it is to the segment being translated on both a formal and a conceptual level. So, ESA similarity scores are an indication of the likelihood of a text segment being useful as a reference or by giving the translator ideas for possible translations of material contained within the segment.

However, this thesis is certainly not making the claim that the Levenshtein distance method should be thought of as redundant, particularly in contexts where the technical solution of using knowledge directly from Wikipedia for translation memory systems is impractical. There are many practical issues involved in the use of ESA that are not covered in this thesis, such as, the precise form that an actual implementation of ESA in a TMS will take. As mentioned in Chapter 4, it is a time-consuming process to find appropriate Wikipedia dumps relevant to a particular translation job with which to train ESA. The Levenshtein distance method measures string similarity, and ESA measures semantic similarity. From this perspective, ESA does not conflict with the Levenshtein distance method. ESA, as a semantic processing technique, is more complex and specialised than the Levenshtein distance method. But the implementation of ESA is also a complicated process that requires the support of a knowledge base. As presented in Chapter 4, ESA training requires intensive computing power and access to an external knowledge base. Consequently, ESA is not intended to replace the Levenshtein distance method, because the simplicity of the Levenshtein distance method ensures that it does not require an external database to perform TMS tasks. Therefore, there is no conflict between the two approaches, which are in fact complementary.

6.1.2 Two Types of Translation Suggestion

The second significant finding of the experiment is that the translation suggestions found by ESA can be categorised into two types: FS and CR translation suggestions. The first type of
translation suggestion, semantically similar, can be used in an analogous manner to the
translation suggestions found by the Levenshtein distance method. Translators may find similar
terms, words or syntactical structures from FS translation suggestions. The second type of
translation suggestion is the CR. As explained in Section 6.2.1, this kind of translation
suggestion is measured differently, because ESA is a fundamentally different technique.
Inevitably, these CR translation suggestions do not primarily provide translation equivalents;
instead, they mainly offer conceptual clarification of the texts being translated. In translating
texts from specialised fields, translators can frequently face content that is difficult to
comprehend. For example:

Sentence 1: 在入骨髓房前，病人必须用消毒药水清洗全身，然后穿上经消毒
的衣服，便进入骨髓 房内的无菌室 (Bao, 1997)。

BT: Prior to entering to the BMT (Bone Marrow Transplant) room, the patient
must bathe using sterile water, then put on sterilized robe before entering to the
germ-free room inside the BMT room (Cook, 2008).

Without relevant medical knowledge, translators may find it difficult to comprehend the text
and make a factual mistake in the sentence. In fact, ‘便进入骨髓 房内的无菌室’ (the germ-
free room inside the BMT room) is incorrect information: there is no other room inside the
BMT room.

In such circumstances, translators may normally consult reference works, such as
encyclopaedias, online resources, or consult professionals. But it may be a time-consuming
process to find the relevant information, as most reference works were not designed for
translators’ needs. Thus, it is important to find a cost-effective way to locate and present
background information. A CR translation suggestion is potentially a good form of reference,
because it is directly generated from within the translation memory system depending on the
text currently being translated. Consequently, translators do not need to try to comprehend the
texts and then search for relevant information based on their initial guesses. Translators may
need to consult many resources in order to obtain the specific information required for
translating. If a TMS is able to provide a conceptual clarification, it may save the time taken,
both by searching for the relevant information and by specifying the scope of the relevant
information. Cases of how conceptual clarification may help translators are demonstrated as
follows.
Sentence 2: The idea is that countries would coordinate counter-cyclical regulatory measures, thereby preventing capital from skirting them by moving across borders. (Yueh, 2011)

Sentence 2 may be difficult for translators to comprehend in the first place. But it may be helpful for translators to have a sentence which is CR to this completed sentence, such as ‘This inflow of funds into the stock exchange, occurring also in other countries, may continue’. With this translation suggestion, a translator may quickly recognise that the sentence is describing the movement of money and financial regulation.

Another example is provided by sentence 3:

‘Dark energy will continue to push galaxies ever faster away until they fade completely from view. (emphasis added).

Sentence 3 is also relatively difficult to translate, in particular the highlighted part. It would be helpful for a translator to receive a translation suggestion, even one as simple as ‘But if the dark energy density decreases and matter becomes dominant again, our cosmic horizon will grow, revealing more of the universe’. Such a CR translation suggestion can potentially offer translators useful insights into the meaning of the sentence. In this case, it may remind translators that the key information of sentence 3 is about an effect of dark energy itself. As explained above, it should be noted that the pseudo-TMS tasks conducted in this study do not offer a translated version of the translation suggestion; but at the same time, this translation suggestion could also provide the translation equivalent of 'dark energy' if its target language version were available.

The above cases demonstrate the usefulness and importance of CR translation suggestions. Obviously, a translator is unlikely to receive this type of translation suggestion from string measure methods. A CR translation suggestion should have been a product of the semantic processing techniques used in translation memory systems. It offers an additional form of translation suggestion along with traditional fuzzy translation suggestions and rough MT output. However, it should be noted that the use of CR translation suggestions still requires awareness of linguistic issues and translation techniques and the ability to distinguish between useful and non-useful suggestions. Otherwise, translators may not have the competence to employ such suggestions. The advancement of the technology does not aim to decrease the level of intellectual engagement in translating; it merely eliminates some unnecessary effort in searching and provides greater scope for creativity. Therefore, from this perspective, the involvement of ESA still enhances the original purpose of a TMS as a type of CAT tool.
6.1.3 The Impact of Textual Factors Relating to Test Collections on ESA Performance

The evaluation examined the impact of textual factors on ESA performance. Although ACA and RN test collections relate to different topics, ESA performed similarly in these two test collections. Surprisingly, SciAm and ACA test collections should have appeared more similar in performance because most texts from RN were on financial topics. The fact that ESA had similar performance in ACA and RN test collections is possibly caused by textual factors relating to genres.

The results of the experiments suggest that the performance of ESA in the ACA and RN test collections was very similar. First, both means and medians of their ESA similarity scores of valid translation suggestions were very close. Secondly, the FS/CR translation suggestion ratios were similar. Thirdly, the performance of the optimum ranges of ACA and RN were more closely correlated to their overall ranges than was the case with SciAm, and the numbers of results that fell in the representative range were also similar.

Among several textual factors, the ASL and query length of the test collections were identified as the key factors affecting ESA performance. ASL is the average number of words within sentences from a test collection. Query length is the number of words in a sentence being used to test ESA performance. These two figures are more closely related to ESA performance than other textual factors, such as TTR. Many data demonstrated the impact of these two textual factors on ESA performance.

First, according to the overall statistics, the performance of RN and ACA is more closely related to their threshold values, the ratio of FS/CR translation suggestions, SDs, and number of frequent ranges of queries. Their threshold values are 9.17 (ACA) and 8.982 (RN), respectively, versus 8.048 (SciAm). The ratios of semantic-similar/related translation suggestions are: 27:13 (ACA), 28:12 (RN), and 1:1 (SciAm). When focused on optimum ranges of query\textsuperscript{12}, the numbers for each test collection are 23 (ACA), 30 (RN), and 15 (SciAm). Despite this, ACA and SciAm are more suitably categorised as one group in terms of the subject domain from the human perspective.

\textsuperscript{12} As mentioned in Section 5.2.4.3, the top two ranges that have the most translation suggestions.
The most important difference between the subgroup of RN, ACA and SciAm was the ASLs. It is possible to explain that their different performances were caused by different ASLs, which were 21.963 (ACA), 20.046 (RN), and 18.749 (ACA). It is clear that ACA and RN performed similarly, because their ASLs were close. Moreover, ASLs of test collections were also related to the SDs of ESA similarity scores. A lower ASL, as in the case of SciAm, increased the spread of the results. Secondly, the query length of CR translation suggestions 22 was much greater than those of other groups, those 17 for ACA and 16 for RN.

Although texts with different numbers of words are technically the same for ESA to process, the difference is still obvious in practice. Queries ranging from 7 to 32 words in length were matched in the test collections. As the data also showed, most queries were from the ranges 11~12, 15-16, and 19~20. In other words, queries from these ranges have a better chance of being matched by ESA. It is possible that the query range of these is also close to the average sentence length of test collections. Another interesting finding relates to QTR, the ratio of queries to translation suggestions. As explained in Section 4.1.2, the query is the same length as the translation suggestion when QTR is 1. For test collections, QTRs normally come close to 1 until range 14. The QTR for short queries (range 7- 8, 9~10) for each test collection were 0.536 and 0.509 (ACA), 0.395 and 0.463 (RN), and 0.697 and 0.633 (SciAm) respectively. This suggests that ESA can only provide longer translation suggestions for short queries.

In fact, in many cases ASLs were also related to the length of translation suggestions. QTRs were close to 1 for ranges that were close to the ASLs of the test collections. The ASL of SciAm was 18.749; thus the QTRs of ranges ‘15~16’ and ‘17~18’ were 0.984 and 1.23, respectively. The ASL of ACA was 21.963; then the QTRs of ranges ‘19~20’ and ‘21~22’ were 1.147 and 0.896 respectively. As an exception, the ASL of RN was 20.046 but the QTR of range ‘19~20’ was 1.581.

Some very short and very long sentences were more likely to be affected by the preference of ESA. Even when TM files were large enough, an ESA-enhanced TMS did not provide translation suggestions for short sentences which were shorter than half of the ASL of the TM files. It is possible for the TMS to provide translation suggestions for very long sentences, but they face high QTRs. If QTRs are too high, it means that queries are much longer than translation suggestions. Consequently, translation may provide very little translation equivalent for translators, especially for FS translation suggestions.
If ESA is implemented in translation memory systems, a possible solution is to segment the sentences of TM files for short queries, while grouping some sentences together as single translation suggestions. Sentences can consist of clauses and phrases. For short sentences, such sub-sentential structures are also meaningful as translation suggestions. However, segmentation is also challenging from both conceptual and technical perspectives, as determining the levels of granularity can be very difficult for some East Asian languages, especially Chinese. Some languages, such as Chinese, do not have as many grammatical tags (such as inflection, and white spaces between words) that can be recognised. Sometimes, it is difficult to identify a group of Chinese words accurately as either a sentence or a phrase. Section 6.4 will explain this in more detail.

Grouping different short sentences into a single long unit can be more of a challenge. The following technical problems need to be solved:

1) Under what conditions should this kind of translation suggestion be adopted? Is it possible to judge the condition according to QTRs or other indicators? A possible solution would be for translation memory systems to group enough sentences until QTR approaches 1.

2) How do translation memory systems find these sentences? They are found from sentences neighbouring the sentence with the highest ESA similarity scores, or by extracting sentences from the entire test collections.

3) What method should be used to aggregate the similarity scores of these sentences?

Every aspect requires an intensive period of experimentation, which is beyond the scope of this research. It should be noted that the above problems can be related to some existing applications. For example, the second problem can be very similar to the technical architecture of automatic abstracting applications. Both basically relate to producing aggregated information based on certain features, such as specific terms of a topic. In contrast, the perception of QTR is more relevant to translators’ ease of use in employing translation suggestions. Thus, it is necessary to involve end users at a certain point in the research.

Because of the importance of the length of queries and sentences, developers need to consider processing different lengths of texts in different ways in order to optimise the use of ESA in translation memory systems. As discussed above, ESA most frequently retrieves texts that are equal in length to the ASL. Currently, weakness of ESA is processing both long and short texts. Therefore, one potential solution is to find ways to avoid processing texts that are too long or
too short. Some applications may be helpful for converting those different-length texts to make them approximate more closely to the ASL of selected TM files. More comprehensive document analysis tools are required to break document collections down into different levels of granularity as for TM files, and researchers also need to find suitable large texts, for instance, the Internet can be used as a corpus for translation purposes. Further research may be conducted in this area.

6.1.4 Reflection on Measures Used in the Evaluation

This section presents a reflection on measures used in the evaluation. The evaluation of ESA differed significantly from other research attempts. The evaluation of ESA should always be founded on the core function of translation memory systems, which is to provide translation suggestions to translators. As noted in Section 4.2.3.2, a sufficient number of results should be gathered for a convincing evaluation, and then valid translation suggestions should be recorded and measured in the proper way. As a result, some special measures are used to demonstrate the performance of the matching methods. However, some of these are not only useful for the evaluation but also for end-users.

The experiments collected many results; thus, the statistical measures, mean and median, were used to summarise average performances. Two aspects were measured, due to their relevance to translators: (1) scores of translation suggestions and (2) lengths of queries. These are directly related to the use of semantic processing techniques in translation memory systems. Moreover, query length is important to translators' workloads. Obviously, reading and translating longer texts increase translators’ workloads. In this study, QTR was employed to indicate the length of queries.

The ESA similarity scores are used as indicators of an individual valid translation suggestion and to observe a method's performance. All the ESA similarity scores should be converted first in order to be used as an indicator for end-users, because the original cosine values given by the ESA IR platform were not easily identified, as explained in Section 4.2.3. The ESA matching experiment shows that translation suggestions with higher ESA similarity scores were not necessarily more useful than those with lower scores. However, this does not suggest that ESA similarity scores were not useful. As mentioned in Section 5.2.2, translation
suggestions with higher ESA similarity scores can improve the chance of retrieval by the ESA IR platform. This is because most translation memory systems rank translation suggestions with higher scores first. By default, the ESA IR platform lists the sentences in descending order. This can be particularly important for genre groups that are similar to ACA; for those genre groups, the threshold value of ESA similarity scores can be higher. Thus, the ESA similarity score is still an important indicator for end users.

The performance of ESA was measured by how a matching method processed queries from the test collection and then aggregated single results into a general performance. In the evaluation, mean, median and SD were proposed as measures of the average of ESA similarity scores. The means of scores are recommended for threshold values because most valid scores are distributed in ranges according to the performance of ESA. In addition, the standard deviation was used to measure the spread of scores and lengths of queries. Although it is not an indicator for end-users, researchers should measure the stability of a TMS that has been enhanced with the use of a new matching method.

The QTR can indicate whether the matching method prefers translation suggestions of particular length ranges and also is used as a measure to quantify the translator’s workload. This is meaningful for both researchers and users. For end-users low QTRs (< 0.5) may increase translators’ workloads, because they need to read more translation suggestions. For researchers, either very low or very high QTRs are not a good sign, as they indicate that the matching method may not be capable of fully considering sentences with different numbers of words.

The measures used in the evaluation are useful for both developers and end-users. The means of valid translation suggestion scores can be used as threshold values for translators to identify whether translation suggestions are suitable before they read all the translation suggestions. However, the means of valid translation suggestion scores are also useful for developers as benchmarks to compare the performances of different methods.

In the future, a new kind of matching score for retrieving translation suggestions can consider aspects of similarity and translators’ workload based on both ESA similarity score and QTR. Here, it is called ‘similarity-length score’. Ideally, this kind of score may filter too long or too short translation suggestions that are, however, also semantically similar. This matching score may be especially useful in the case of a particularly large and repetitive TM database that may
bring too many translation suggestions. In fact, an ideal match score should not be regarded solely on the basis of the similarity relation between texts and translation units, but is an index calculated from many factors that are related to the productivity of translation. However, the study of those factors goes beyond the scope of this study. It can be foreseen that designing the matching score will require both knowledge of the technical ability of semantic processing techniques and the needs of translation from a human perspective.

6.1.5 Summary

This section discussed five significant findings: (1) the advantage of ESA compared with the Levenshtein distance method; (2) the usefulness and importance of the CR translation suggestion; (3) the importance of ASL and queries and their impact on designing translation memory systems; and (4) measures used in evaluation. More importantly, the third finding raises the technical problem of grouping short sentences together as longer units to overcome the QTR problem. The fourth section proposes a new idea for defining matching scores for translation memory systems.
6.2 Bottlenecks when Using Explicit Semantic Analysis in Translation Memory Systems

The evaluation confirms the possibility of using ESA in TMS tasks. However, there are many problems when implementing ESA in practice. Some problems demand the involvement of other current techniques, but other problems are bottlenecks of NLP in general: the knowledge presentation and knowledge base. This section describes how bottlenecks affect the use of ESA in translation memory systems, especially for Chinese. In fact, some problems are not only technical, but also involve many other areas, such as commercial involvement and the study of grammar.

6.2.1 Problems of Representing Knowledge in Chinese Texts for Explicit Semantic Analysis Technique

The nature of translation is still about processing human languages, so it is best to have knowledge of translating as well as subject domain knowledge when using ESA for translation purposes. Unlike other applications, such as recommender or IR systems, most translation suggestions are untagged sentences, rather than a set of texts or units that have more explicit tags (or annotations). For example, a DVD sold on Amazon can have tags such as ‘directors’, ‘actors’ or ‘price’. The fundamental aim of ESA is to measure the semantic similarity between queries and sentences from TM files. Information on language knowledge, such as syntactic structure, word-order, contained in translation suggestions should be identified and extracted first, but is less precise when compared to tags used in IR systems. So, as a NLP application, TMS tasks are more complicated.

ESA employs a different form of tagging of translation suggestions, as it considers the meanings of texts by referring to their positions in Wikipedia. Wikipedia mainly offers encyclopaedic information and is meant for human readers (Dornescu & Orasan, 2014). Currently, ESA only measures the semantic similarity of texts according to the representation of the knowledge domain from sources such as Wikipedia, based on the correspondence of terms. It is possible to extend the TMX tag-set to manage translation suggestions by combining
language features, such as syntactical structure, into semantic measures. For example, if a query is an interrogative sentence, translation memory systems may assign higher similarity scores to other interrogative sentences from a TM database. Thus, annotations—such as genres, sentence types, level of formality, and rhetorical features—may be helpful for ESA or other techniques to retrieve more accurately, because they include more features.

However, there are several problems with having this kind of information in the implementation of ESA for Chinese. Compared to English, some linguistic features of the Chinese language may be very difficult for ESA or other similar techniques to represent. When obtaining the linguistic information, syntactic analysis (parsing) —namely, taking an input and producing some sort of linguistic structure for it (Jurafsky & Martin, 2009: p. 79)—is necessary. Currently, ‘the linguistic structures’ of sentences are normally based on models such as context-free grammar (CFG) or dependency grammars (Jurafsky & Martin, 2009: p.423). A sentence is parsed according to a set of rules into different components, and their relation can be shown in a parse tree. For example, the parse tree of sentence ‘the woman saw him’ is as follows:

![Parse Tree Example](Collins, 2013)

Where

S stands for sentence.

NP stands for noun phrase.

D stands for determiner.

N stands for noun.

VP stands for verb phrase.

V stands for verb.
P stands for pronoun.

Syntactic analysis provides some more comprehensive information on texts, and some functions are available for translation memory systems. Translation memory systems use chunk-based segmentation. Unlike most translation memory systems that measure similarities between sentences, a chunk-based TMS is able to process translation suggestions on a sub-sentential level, i.e. on the noun phrase level (Colominas, 2008). In addition, some scholars also attempt to measure semantic similarity with the involvement of syntactic analysis in Chinese NLP. Li et al. (2003) employ a method that combines the dependency structure of sentences with the counting of common words as the basis for computing semantic similarity. It achieved 81% accuracy based on 43 sentences (Li et al., 2003).

POS tagging and segmentation is vital for syntactic analysis techniques (Jurafsky & Martin, 2009: p. 461). Parsers may have to identify the number of constituencies (groups of words as single phrase units) (Jurafsky & Martin, 2009: p. 429) in a sentence, and then analyse their relation according to grammatical rules. However, the problem is that the grammatical rules on which this is based are not always applicable to Chinese, and some grammatical rules are yet to be formalised.

In NLP tasks for English texts, POS tagging can be accomplished very effectively. Most lexical inflexions are relatively regular in English. It is easy to identify tenses and voices according to verb forms, such as ‘have done’, ‘had done’ and ‘be done’ in English. The language features of the Chinese language make POS tagging very difficult. Although some ending functional words (着，了，过，的) can have similar indications in Chinese, they are not very accurate and there are no clear grammatical rules to distinguish them. According to the corpus of ‘现代汉语语法信息词典’ (Grammatical Knowledge-Base of Contemporary Chinese, GKBCC), 1,601 verbs can be attached with either ‘了’ or ’过’, and 1,151 verbs with the resultative verbs and directional verbal tags ‘着’，’了’，’过’ and ’在’ (Zhan, 2004 and 2010). Except on some occasions, such as ‘您/你’ (honorifics) and ‘它/她/他’ (grammatical gender), most Chinese words do not have inflexions corresponding to English, such as case, number, gender, voice, and tense as grammatical tags.
Compared with English, POS tagging is less important in Chinese NLP tasks, as the relationship between syntactical structure and word classes is more flexible in Chinese. In English, the sentence structure is relatively strict, as most sentences ‘contain at least one main clause with an independent subject and verb and express a complete thought’ (Simmons, 2014). The grammatical rules of constructing compound sentences, complex sentences and clauses are usually clear, as indicated by most Chinese texts books of English grammar. The relationship between clauses is also more explicit, as cohesive devices are used more often (ibid.). More importantly, sentence units (subjects, objects, verbs) also demonstrate a relatively strict relationship between words of different POS (ibid.). For example, it is not common for a verb to be the subject of a sentence, unless it changes its POS.

Chinese relies more on the change of word order and on function words to express meaning, although both English and Chinese are ‘subject-verb-object’ structure languages (Lin, 2006: p.4). Sometimes, Chinese also adopts ‘subject-object-verb’ structure in ‘把字句’ (ba constructions) and ‘被字句’ (bei constructions) (Lin, 2006:p.85). The typical ‘subject +verb’ structure does not always exist in real world texts, as it is not necessary according to Chinese grammar. Many usages of Chinese grammar are very different from English. These differences may create ambiguity if Chinese sentences are parsed and POS tagged according to English grammar.

Verbs can be used as the subjects of sentences. For example:

竞争提高考试成绩。

(BT: To compete improves examination scores (literal); Competition improves examination scores.)

‘竞争’ competition is a verb in the above sentence.

Two or multiple verbs can be used in one clause.

战士们趴在地下匍匐前进。

(BT: Soldiers lay on the ground crawled (literal); Soldiers lay on the ground and then crawling.)
They take plane fly to Beijing (literal)/ They fly to Beijing.

‘坐’ is a verb meaning ‘take’ or ‘use’; ‘飞’ means ‘fly to’.

Adjectives and nouns can function as predicates in some sentences:

武汉天气炎热。 (BT: Wuhan weather scorching hot (literal)/ It is very hot in Wuhan)

苹果多少钱一斤? (BT: Apples how much money one pound (literal)/ How much is a pound of apples?)

The copula ‘是’ (be) is not used in these two sentences.

According to these examples, Chinese is a type of topic-prominent language, which tends to place the topic, rather than the subject, at the front of the sentence (Lin, 2006: p.2). A sentence is formed primarily as a group of phrases to describe a topic, rather than according to a typical grammar structure (SVO). As a result, the parsing of a Chinese sentence is more contextually dependent and flexible. To disambiguate the ambiguities in Chinese language, it is sometimes necessary to apply common sense, logical reasoning, and contextual referencing to other parts of the text (Pan, 1997: p 158). Moreover, no word boundaries are used between Chinese characters, and punctuation is the only indication of breaks in clauses and sentences. The following sentences can be either a set of independent short sentences or a single long sentence:

Sentence 1: 我在一头租来的骡上骑了一整天，直到曾担任我向导的邮差指着远处远边上的小屋，默默地拒绝了我给他的钱，骑着骡走了。(BT: I had spent a long day on a hired mule before the mail carrier who had been my guide pointed to a cabin on the far side of a stream, mutely refused the money I offered, and rode on.)

Sentence 2: 我在一头租来的骡上骑了一整天，直到曾担任我向导的邮差指着远处远边上的小屋，默默地拒绝了我给他的钱，骑着骡走了。(BT: I had spent a long day on a hired mule. I stopped until the mail carrier who had been my guide pointed to a cabin on the far side of a stream, The mail carrier mutely refused the money I offered. The mail carrier rode on.)
GKBCC uses up to 40 grammatical attributes to aim at more explicit Chinese POS. With this standard, 14,479 words can be classified into 7,897 categories (Yu et al., 1996). Accordingly, 6,105 categories contain only one word, while 903 categories have two words (ibid.). This suggests that the usage of 6,105 words is unique. For example, ‘当然’ is an adverb, meaning ‘of course’ or ‘obvious’. Its usages include being placed after a pronoun as an adverb, or being placed in front of a sentence or between two clauses to be a cohesive device:

今天雨很大，他当然不会来。(BT: As it is heavily raining today, he will of course not come.)
当然，努力就会成功。(BT: Of course, working hard will be successful.)

Moreover, it can be an adjective to describe another ‘state’:
获胜，心里高兴是当然的。(BT: It is normal to be happy about victory.)

Studying the usage of ‘当然’ does not mean all adverbs can be used in this way. While a large number of corpora are available, the practicality of having explicit categories of words is quite low. From a corpus linguistics perspective, the distribution of ‘当然’ is unique in terms of its co-occurrence with other words. Zhan (1997 & 2010) and Dong (2002) recognise that researchers are not able to train useful models to recognise phrasal structure based on a collection of Chinese words, as constructions of phrasal structures are so diverse in terms of the possible combinations of word types.

It may suggest that POS tagging is not applicable to Chinese NLP tasks. Given the ambiguity of Chinese POS tagging, it is difficult to represent the different meanings of two sentences with ambiguous sentence structures and words. For example, ‘他上海外贸易学院的’ can be parsed as follows:

1) 他 上海外贸易学院 的 (BT: He+ Shanghai Institute of Foreign Trade+ ‘De’ particle)

2) 他 上 海 外 贸易学院 + 的 (BT: He+ Go + Foreign/overseas+ Institute of Foreign Trade+ ‘De’ particle)

Most Chinese native speakers would probably prefer the first Interpretation. However, it is difficult to instruct ESA or other similar techniques to parse the sentence correctly. Both 上海 and 海外 are legitimate words. ‘上海外贸易学院’ can also be a Wikipedia entry. Thus, sentence segmentation, especially on a sub-sentential level, is a vital factor for using ESA and other
similar techniques. Unlike English, Chinese relies more on functional words and the change in word-order. Therefore, the method of presenting this kind of feature is the next stage for Chinese NLP.

Normally, the Chinese language places the most important topic of the sentence at the beginning of a sentence, and the meaning of a sentence is different corresponding to different arrangements of units; for example, ‘It’s Father’s book’ can have two Chinese expressions:

1) 这本书是爸爸的。 (BT: This book is Father's.)

2) 这是爸爸的书。 (BT: It's Father's book.)

The difference between sentence 1 and sentence 2 is that the first sentence specifies that ‘father’ owns this book. The latter emphasises that the ‘book’ is one of father’s books. Unfortunately, there is not a very effective method to represent the meanings according to different word order, as there are no systematic studies about the word order for the purpose of semantic similarity measure. These two expressions are likely to receive the same or very close similarity scores if they are processed by current semantic similarity measure.

Most Chinese scholars agree about the typology of phrases and sentence structures, as the number of sentence and phrase types is formalised and strict. As in English, sentences in Chinese can also be categorised according to different types, such as declarative sentences, interrogative sentences, imperative sentences and exclamatory sentences (Lin, 2006:p.8). Chinese sentences (and also clauses) consist of phrases that are made up of different words. The typology of phrases is also explicit, such as endocentric phrases, subject-predicate phrases, verb-object phrases, coordinative phrases and prepositional phrases (Lin, 2006: p.10). But there is much controversy over the construction of Chinese phrases and words, especially what kind of words can constitute what kind of phrases and the relations between categories of words. Several explanations are offered for this.

Shen (2009) argues that certain groups of verbs can be seen as a subcategory of nouns in Chinese. Some scholars attempt to build a set of rules to describe the conversion between each category of Chinese words that can be used as a verb, noun or adjective. As a result, categories of words overlap each other. However, this school of thought is not useful for the application of NLP tasks, because these rules cannot form a functional system to explain the construction of Chinese words. For example, 数学 (mathematics) is obviously a proper Chinese word.
However, it is difficult to explain why 数学学习 (Learning mathematics) is not a word while 数学成绩 (mathematics scores) is a word. Even though it has a conceptual explanation, we are still a very long way from enabling a computer to identify Chinese phrases correctly.

Another attempt is to focus explicitly on usage of words that are regarded as different grammatical attributes, e.g., ‘collocated with two-object structure’. Such research assumes that Chinese words require a more explicit category framework (Song, 2010). However, unlike English, most Chinese words share little in terms of usage, even words within the same category.

Some scholars recognise that the study of combinations of semantic meanings may be a breakthrough for the next stage (Dong et al., 2010; Zhan, 2010; Shen, 2009). A way of describing Chinese grammar based on semantics is probably more useful. They suggest that the construction of Chinese syntax and phrases is arranged according to several semantic orders, so Chinese native speakers comprehend texts based on collective recognition. Human understanding of language is fairly complicated and differs fundamentally from that of computers, so not even a state-of-the-art technique is able to apply the complicated linguistic knowledge of grammar to NLP tasks.

To conclude, the study of Chinese grammar does not match the level of NLP applications, as a different Chinese grammar system is required to describe both the use of functional words and the word order. More comprehensive research is required to explain these two very important phenomena in Chinese.

### 6.2.2 Problems with Knowledge Bases in Chinese

As a type of machine learning technique, the ability of ESA is highly relevant to the training data. The current implementation of ESA is primarily based on Wikipedia, which is the largest online encyclopaedia, based on a combination of average daily visitors and page views according to Alexa (2014).

Wikipedia is a multilingual encyclopaedia; thus, it is potentially feasible for ESA to perform semantic measurement in languages other than English. Currently, the English Wikipedia is the largest language version: by March 2014, it had approximately 4,478,441 articles (Wikipedia,
Meanwhile, there are at least one million articles in eight other language versions, all of them European: Dutch, German, Swedish, French, Italian, Russian, Spanish, and Polish (ibid.).

The size of Chinese Wikipedia is relatively small, considering that Chinese is the most widely spoken language in the world. Chinese is the 15th largest language version; by March 2014, it had 753,629 articles (Wikipedia, 2014d). In terms of the number of articles, Chinese Wikipedia is much smaller than English Wikipedia. There are 132,269 active users who are registered and have made at least one new or edited English Wikipedia entry in the last 30 days (Wikipedia, 2014d). By contrast, Chinese Wikipedia has 6,943 active users (Wikipedia, 2014d). These figures indicate that Wikipedia is much more popular in the English-speaking world than the Chinese-speaking world, even though there are far more native Mandarin Chinese speakers than native English speakers.

Of all 285 languages available in the encyclopaedia, Chinese Wikipedia is one of the largest. All of the top 50 language versions exceed 10,000 articles. Wikipedia uses editing depth to indicate the collaborative quality, i.e., the frequency with which Wikipedia articles are updated (Wikimedia, 2014).

\[
\text{Depth} = \frac{\text{Edits}}{\text{Total}} \cdot \left( \frac{\text{NonArticles}}{\text{Articles}} \right)^2
\]

NonArticles include ‘user pages, redirects, images, “project” pages, categories, templates, and all talk pages’, and Articles are Wikipedia articles. The total is the sum of Articles and NonArticles. Edits are the total number of times that articles have been edited (ibid.).

The editing depth of Chinese Wikipedia is the 10\textsuperscript{th} highest (112.4), compared to 849.5 for English (Wikipedia, 2014d). Hebrew Wikipedia has the second greatest depth, at 254 (Wikipedia, 2014d). The editing group of Chinese Wikipedia works more closely than many other large Wikipedia language versions, such as Cebuano, Waray-Waray, Vietnamese and Japanese. Cebuano and Waray-Waray Wikipedias have 892,558 and 959,446 articles respectively, but their editing depths are both 3, meaning that articles are not edited and updated many times after being created (Wikipedia, 2014e). Vietnamese Wikipedia has 886,597 articles and an editing depth of 17 (Wikipedia, 2014d). Among all of these, the Japanese community performs relatively better, with 897,522 articles and 66 editing depth. These figures demonstrate that Chinese Wikipedia articles were edited more frequently,
showing very high levels of collaboration and communication. In addition, not only Mandarin Chinese, the most common standard Chinese, is available online; there are several other variations of Chinese on Wikipedia, including Wu Chinese, Cantonese, Min Nan, Min Dong, Hakka, and even Classical Chinese (Wikipedia, 2014).

Other online encyclopaedias in Chinese also adopt similar technical infrastructure and community rules. Baidu Baike (百度百科) and Hudong Baike (互动百科) are the two largest encyclopaedias in mainland China. They both aim to cover broad areas of knowledge. By March 2014, Baidu Baike had 7,624,849 entries, and Hudong had 8,128,164 entries (Baidu, 2014; Hudong Baike, 2014). Although they both allow users to collaborate in the editing of the contents, they do not allow developers to download the contents for exploitation in other applications. MBA lib (MBA智库百科) is a specialised encyclopaedia that has 343,348 articles covering business and financial topics (MBA lib, 2014).

The biggest problem with resources is whether they can maintain neutral points of view. Unlike Wikipedia, all of them are maintained by commercial companies. Baidu Baike and Hudong actively use the website as a platform for advertising. For example, Hudong charges companies for adding product information as encyclopaedia entries (Qiye Baike, 2014). The business model of MBA lib is to use the website as a platform for recruiting and job hunters. With sponsorship from large corporations, the content of such websites grows much faster than that of Wikipedia. However, it may be very difficult to maintain the quality. For example, Wikipedia, Hudong and Baidu Baike have entries about ‘haemorrhoid’. The entries of Hudong and Baidu Baike contain commercial advertisements for medical treatments. If many such entries were included in a training dataset, the noise of the knowledge domain would be very high, as many irrelevant terms would also be included, e.g. the brand names of certain drugs and hospitals.

In addition to the quality of other Chinese encyclopaedias, the Chinese language itself may also create problems in the use of translation memory systems. The challenges primarily relate to different standards of translation and terminology, and the two different Chinese writing systems.

Although both China and Taiwan adopt Mandarin Chinese as their official language, the differences between the use of Mandarin Chinese in these two countries are noteworthy. The lexical differences are noticeable. For example, software is ‘软件’ in mainland China, while it
is ‘軟體’ in Taiwan. Some words also have different meanings; for example, ‘男生’ ‘女生’ refer to young school boys and school girls, but in Taiwan can also refer to young men and women. There are many syntactical and grammatical differences as well. In both varieties of Chinese, ‘有’ is a verb meaning ‘to own something’, e.g. 我有一本書 (BT: I own a book). But in Taiwan ‘有’ is also often linked with another verbal phrase to describe an action that has occurred, starting from the past until the current time. ‘有’ functions similarly to the English present perfect tense; for example, 我有學習 (BT: I have been studying) (Zhang, 2004). This kind of use is not prevalent in mainland China. Moreover, Taiwanese Chinese sentences are usually shorter than sentences written in mainland Chinese texts, because link words and parallel structures occur more frequently in mainland Chinese texts.

Other than China and Taiwan, Mandarin Chinese Wikipedia is also edited from some south-east Asian countries. Both Singapore and Malaysia have their own standardized forms of Mandarin Chinese. After 1976, Singapore adopted the simplified Chinese characters of mainland China, and local Chinese schools started to use simplified Chinese in Malaysia after 1981 (Xu, 2009; Yeap, 2012; Hu & Zhou, 2009). Hence, Singaporean and Malaysian simplified Chinese versions can also be found in Wikipedia. Therefore, most Chinese Wikipedia articles normally have five versions— simplified Mandarin Chinese characters, Hong Kong traditional Chinese characters, Macau traditional Chinese characters, Taiwan traditional Chinese characters, and Singaporean and Malaysian simplified Chinese (Liao, 2008):

Figure 6.2: An example of a Wikipedia page about ‘微信’ (WeChat). As displayed, the page is available in five Chinese versions.

Although all of these languages are considered Mandarin Chinese, there are some issues when converting traditional and simplified Chinese characters. One of biggest issues is that

---

13 WeChat is a mobile text and voice messaging app.
ambiguities can be caused by simplifying Chinese characters, because many characters with different meanings are merged into single simplified Chinese characters. For example, ‘發’ and ‘髮’ are different characters in traditional Chinese writing systems. But they were all converted to ‘发’ in the PRC simplified standard. ‘發’ usually means ‘development’, and ‘髮’ usually means ‘hair’. This can be a problem when converting the simplified Chinese characters into traditional characters. For example, converting ‘发菜’ (a kind of moss) is not straightforward, as either ‘發’ or ‘髮’ can be acceptable. More importantly, the confusion also affects the identification of words, as most Chinese words are combinations of two characters. For example, both ‘搜’ and ‘蒐’ exist in traditional Chinese forms. Normally, ‘蒐集’ should convert into ‘搜集’ (to collect something) in the simplified form, as ‘蒐’ is not used in the simplified version when used for the meaning of ‘collection’. However, many Chinese characters contain multiple meanings. It is questionable whether ‘搜’ can replace ‘蒐’ in the word ‘搜索’ (to investigate), as no clarification was actually made. In practice, ‘搜索’ cannot be converted into ‘蒐集’ by using the Chinese inputting system.

Similar cases are also caused by China’s language policy. In 1986, the second round of simplified Chinese characters was officially declared to be cancelled, with many simplified characters remaining unresolved (Su, 2003). Some characters were newly created but were not given a clear status of official characters, for example ‘付’ (Su, 2003). ‘付’ can be the simplified version of ‘傅’, which is normally used as a surname and in some ancient governmental titles (太傅) or 師傅, an honorific for highly skilled workers or tutors. But both ‘付’ and ‘傅’ are still used today. It is a problem if these two characters can be used interchangeably. In this case, developers of a Chinese knowledge base may have a problem to choose either ‘太傅’ or ‘太付’ to be included as an entry for the database, and cost a huge amount of man power to identify and organise the use of ‘付’ from Chinese texts written in the 1970s. As Chinese is a living language, new words are constantly being produced. Resolving such ambiguity can be considered a new task for Chinese NLP, especially when some NLP applications are based on statistical approaches. Since Chinese corpora contain many ambiguous words, the performance of such applications can be compromised. The lexical and grammatical differences make it relatively difficult to use different versions of Chinese on the same platform, as they may affect the calculation of the semantic measure.
As discussed, there are many problems in using knowledge bases for Chinese NLP, and not all of them are technical issues. In fact, these problems relate to many different aspects, including the commercialisation of online resources, the negative consequence of the Chinese character simplification scheme, and the diversity of Chinese character systems, and of the language itself.

6.3 Perspective of Translation Memory Systems in the Future

As explained in Section 2.5, translation memory systems are a conceptualised type of KMS that follows a certain KM process. Semantic similarity measures may greatly enhance knowledge codification and knowledge application steps to improve the efficiency of translating. The evaluation of ESA suggests that such semantic similarity measures are applicable in TMS tasks, although several aspects of these techniques need to be adjusted and developed. The functionality of translation memory systems do not necessarily change with the involvement of ESA, because such techniques do not aim to change the basic interaction between translator and software. However, new technology could lead to change on the developer and commercial sides.

6.3.1 Translation Memory Systems as a Type of Knowledge Management System

As explained in Chapter 3, most semantic similarity measures require support from knowledge bases manually developed by either experts (i.e., WordNet and OpenMind) or online users (Wikipedia). Regardless of the technical perspective, semantic similarity measures and knowledge bases are two sides of the same coin’. In other words, there is a close connection between applications and knowledge bases. However, developers tend to design applications based on types of knowledge base, rather than developing a knowledge base for a specific application. The former is more efficient, as it can employ an existing resource. Some applications can be based on core applications that rely on certain forms of database. For example, applications such as recommender systems, IR, answering systems, translation memory systems and plagiarism checkers are primarily based on the use of semantic similarity
measures. The process would be much more efficient if all of these applications were based on a large database for general use. Employing a large, easily accessible database is the key to developing the most useful and efficient applications. Because of the reliance on large online databases, it can be foreseen that translation memory systems may adopt the SaaS model that provides the application for end-users via the web rather than implement in desk software (Aitken, et al., 2012). This section provides a vision of the infrastructure of translation memory systems as a type of KMS.

Currently, translation memory systems are distributed by several companies. The mainstream software, such as SDL Trados Studio, memoQ, Wordfast and Déjà Vu, is already compatible with standardised TM formats, such as TMX. The core function of different translation memory systems, as a type of IR system to match queries and translation suggestions, are homogeneous. From an economic perspective, most commercial translation memory systems are not very different from one another, but can be collectively described as a homogeneous good, which 'has uniform properties, any unit being interchangeable with any other' (Black, Hashimzade & Myles, 2009). The differences between translation memory systems are found mostly in the extra features that each of them offers. For example, one TMS may be strong in verification functions, while another may be strong in terminology management (SDL, 2014b; Kilgray, 2014). There is a recent trend for TM companies to attempt to employ large datasets to create competitive advantages. For example, SDL Trados Studio offers large dictionaries from the European Union to enable AutoSuggest. WordFast provides the ‘Very Large Translation Memory’ (VLTM) for Wordfast Anywhere users. Some third parties have started to provide language resources to integrate with translation memory systems. For example, WordFinder is an online dictionary that supports 15 languages (WordFinder Software International, 2013a). WordFinder claims that up to 120 dictionaries are available to subscribers within a single interface (WordFinder Software International, 2013b). In addition, some big data owners, such as Google, have made their MT applications available for some Translation memory systems, such as SDL Trados Studio.

Translation memory systems are always used with certain resources that provide references for translators. Without the translation suggestions of TM files, translation memory systems are mere text editors customised for translators. The development of translation memory systems is not to provide convenient features, but is dependent on whether translation memory systems can be platforms to collect and analyse new sources of information, especially the vast amount of information online. Translation memory systems may not only have better chances of
finding translation suggestions within large databases, but also most newly available applications, such as statistical machine translation and semantic similarity applications, which require a large amount of data or structured information.

It is possible that the traditional form of translation memory systems has changed. Traditionally, different companies have released their own versions of software, and translators have been required to download the software locally, thereby acquiring licensing fees. The problem is that the local client text editors may not significantly improve productivity, because current translation memory systems do little to manage translation resources (i.e., finding, collecting, analysing and presenting translation resources). Translators are still required to complete many trivial and repetitive tasks, such as finding the right words each time they translate new texts (Bowker & Barlow, 2008: p.12-3).

Instead of adding new convenient features for text editors, translation memory systems should become types of KMS that manage the knowledge required for translation. That knowledge is twofold: (1) knowledge for human translators and 2) knowledge for computers to enable new features, especially semantic measures.

Translation memory systems should be treated as types of KMS because of the complexity of translation. As stated earlier, the current drawback to translation memory systems providing FS translation suggestions is that the systems only contain information related to translators; they should also contain the 'knowledge' necessary to manage other types of knowledge. The purpose of managing is to improve the performance of translation memory systems by analysing knowledge that is better saved in certain structures. Single TMS companies are not likely to create a successful and functional database, particularly for the knowledge base for measuring semantic similarity, because most companies focus on the convenience features of their products rather than on matching methods and translation resources. Commercial companies seem reluctant to develop and maintain an open database for sharing.

A more realistic plan is to employ some existing open knowledge bases, such as Wikipedia. Ideally, a single, giant database for translation knowledge will emerge by combining several smaller databases. Most translation memory systems could be connected to this knowledge base. In this case, most commercial translation memory systems should not be treated as independently released software, but as different interfaces to a core knowledge base. This relationship between TMS developers and data owners could be similar to Linux distribution and Linux Kernel. As different Linux distributions are released, users may select freely, based
on their preferences and needs; but some key functions (e.g., retrieving translation suggestions and semantic measures) depend on the large data selector. Obviously, translation memory systems would work much better with larger knowledge bases. In offline mode, translation memory systems may also function as traditional text editors and still perform some simple matching methods, such as those based on the Levenshtein distance measures.

The trend to transform translation memory systems into types of KMS would render TMS platforms capable of generating knowledge and information for translators, rather than serving as expensive text editors sold by commercial companies. This transformation would allow more independent translation memory systems to appear in the market, and possibly allow more independent developers to release extensions (or add-ons) for translation memory systems, similar to those of Google Chrome and Mozilla Firefox. Arguably, extensions are easier and cheaper to develop than software, as they only require small programs or script languages, such as CSS. The benefit is that translators would be able to modify or improve translation memory systems using these small applications, rather than paying for alternative software. Most features provided by commercial software (e.g., aligning TM files or verifying functions) amount to text processing that can be done online. Translation memory systems are more than text editors; they are knowledge management systems that process different types of knowledge. The transformation of TMS into an online platform could greatly reduce costs and correspond more closely to the KM process of using TMS. The transformation could also provide more developers for the enrichment of the user experience.

The ideal architecture of translation memory systems in the future can be mapped as follows:
As Figure 6.3 shows, the architecture of translation memory systems consists of three layers: the database, platform and end-user interface layers.

The database layer is the foundation for the other two layers and contains most of the supporting information for translation memory systems. The important applications from the platform layer (e.g., semantic similarity measures and retrieval of TM files) of translation memory systems require information from the database layer. Different databases are included in this layer, such as knowledge bases, TM files and term bases. The construction of the database layer requires a huge amount of information. It is unrealistic to collect all the information from scratch. Alternatively, resources from existing databases should be accessed and merged.

The platform layer is designed as an intermediary between the database and interface layers. This platform receives and processes the information from the database layer and transfers it to the end-user applications. These platforms are likely to be used by current TMS companies rather than retaining traditional methods of software distribution, as mentioned earlier. Different platforms may offer slightly different approaches. This enables the core applications
of translation memory systems; i.e., different platforms may employ different semantic similarity measures and different options for retrieving translation suggestions. Ideally, the platform layer would serve as a software environment for different parties to implement particular applications. As long as the platform remains open, anyone with the ability to create useful functions for translation memory systems should be able to interact with this layer. Hence, there would be a place for small, independent developers to release their extensions, which would provide more functions for users to choose from. For example, an independent developer may provide their own MT component along with different platforms. Wordfast Anywhere, for example, operates similarly to a 'platform layer' in that it develops a cloud platform that allows users to access the platform online. However, Wordfast Anywhere is not supported by a functional database layer, so the size of the VLTM (i.e., the TM files that have been collected by the developer and/or uploaded by users themselves) is still far too small. Moreover, Wordfast Anywhere as a platform is still closed to the contributions of other developers. But, developers cannot customise the functions of the platform according to their needs, as in the case of open source tools such as OmegaT.

The end-user interface layer is the main interface with which translators interact. Ideally, the end-user interface layer consists only of essential functions in the beginning. However, user experiences can be improved by allowing end users to modify translation memory systems for themselves. As mentioned, state-of-the art products such as Wordfast Anywhere have operated similarly to the platform layer. The modification methods do not have to be highly technical, especially for the non-functional features of translation memory systems. For example, translation memory systems allow users to adopt different graphical appearances through themes. Every detail can be edited in the end-user interface layers. Ideally, translators would be able to control the interface as easily as they enable/disable a web-browser extension.

To conclude, having a different technical structure may not necessarily change the key function of translation memory systems but neither does it necessarily change the workflow. Dependence on an online resource would build a closer relationship between users and the Internet. It is possible to foresee that translation memory systems may become a kind of online platform, rather than being installed on local drives, along the lines of existing SaaS systems but with the involvement of large knowledge bases.
Translation memory systems as a type of KMS could also give rise to an online community that shares a certain similarity with Wikipedia. As mentioned in Chapter 2, TMS workflow can be conceptualised as a KM process that has knowledge capture, knowledge codification, knowledge application and knowledge creation steps. Knowledge creation, the fourth step of the KM process of translation memory systems, gives rise to a mutually beneficial relationship between users and translation memory systems. Traditionally, the only outcome of knowledge creation is a better growth of the scarce TM files, because translators could then work more effectively. However, the content of knowledge creation may be enriched with the changing landscape of translation memory systems. An online TMS could also play a role as a collaborative writing platform and a personal KMS at the same time. Once such a TMS has attracted a considerable number of users, it might store a huge amount of knowledge, along similar lines to Wikipedia. With the advancement of new technology, translation memory systems are likely to become more widely used in the future and offer more benefits for people inside and outside the translation industry.

Translation, especially translating texts from specialised areas, requires the support of reference information. The need for information and sharing information already forms some online translators’ communities, such as TranslatorsCafe.com. As probably the most important CAT tool, translation memory systems can be integrated with an online community, especially since the development of translation memory systems of the kind that I envisage requires large amounts of data from online. With the enhancement of a large database, it is possible to let translation memory systems attract different user groups that also need to acquire information in foreign languages.

The purpose of translation is to allow information to be disseminated from a speaker (or speakers) of one language to those of another language. Consequently, translation memory systems might also be useful to more than translation-related professionals. The knowledge required both for translation, such as background knowledge, and on translation, such as the bilingual aligned translation suggestions, is complicated. Currently, translation memory systems are a type of specialised tool that is only used by professionals. Some researchers have already explored the use of translation memory systems for other purposes, such as language learning and quick information dissemination (Shi, 2004). Because the original users of translation memory systems must be translators, it is possible to generate huge TM files and
use them to measure semantic similarity. Translation memory systems as an online platform may offer enormous potential for other kinds of application and usage, based on technical similarity.

As translation memory systems ideally contain huge TM files in a bilingual form that enable users to retrieve similar information, users may explore foreign information relating to certain domains within the knowledge base, even from different TMS interfaces. Translation memory systems provide the information in sentence-length segments, a format that will satisfy users who wish to acquire a large volume of information quickly, as is the case with micro-blogs. Some kinds of information are short but are also informative, such as titles, prefaces and abstracts. Translation memory systems may provide a form of news feed, reporting the same topic from different language versions in a selected target language. For example, the disappearance of Malaysia Airlines Flight 370 was a topic that attracted considerable interest in many countries in 2014. It would be convenient and interesting if users could read the relevant headlines in Chinese, Malay and English. It would be useful to overcome bias from the media in each country if a user could compare texts at the same time. It could be a form of news that is able to indicate the differences in reporting in different language perspectives.

Many potential new applications could be developed if a large, semantically structured multilingual knowledge base were available. Another use of translation memory systems would be to recommend relevant short pieces of information based on queries given by users. It is not necessarily for translation purposes, but can also be used for language learning purposes, as semantically similar texts are also very likely to have analogous grammatical styles. On the basis of particular queries, translation memory systems may also recommend useful background information. Not only translators, but also other readers may benefit from such a platform.

If the online TMS attracts enough users, the translation behaviours and the use of the words may result in another type of resource, other than TM files and a knowledge base. Keywords for searching information, finding relevant articles and modifying translation suggestions are very important for translation scholars to understand the user’s translation process, and may also be useful in themselves for other developers. More importantly, with the benefits of techniques such as data mining, translation scholars are able to study the use of language beyond the textual level. It is reasonable to assume that some patterns and features of language may only be observed at corpus level (Mu, 2012). Obviously, it is hard to ensure the
representativeness of the trend of neologism based on a few blogs. With large-scale language resources, statistics on translation behaviours can be made available (Sotov & Ji, 2012). These resources might also demonstrate different perspectives on a topic from different language groups or a different timescale.

It is possible to establish a relationship between language features and consumption habits. For instance, the information on Xiaomi, a smartphone designer, and developing companies is presented very differently in English and Chinese. It is interesting to analyse the impact of Xiaomi in the database to estimate its influence in English-speaking communities. More marketing-related terms, such as ‘狂放’ (wildly release) or ‘五折起’ (50% discount), are likely to be used in Chinese texts, as Xiaomi is mainly targeted at young, low-budget consumers in China.

Figure 6.4.1: A screenshot from the Chinese version of Xiaomi mobile: Xiaomi tends to publish more promotions; the packages include gift vouchers, free mobile phones, free themes and E-books.

However, Xiaomi’s business model is more appealing to Western investors, because they are surprised by the growth rate of Xiaomi in the Chinese market (Ballve, 2014; BBC, 2014). Consequently, Xiaomi often is mentioned in more serious texts and the target readership is mainly business-oriented, as Xiaomi tends to emphasise the technical advantage of its models and use more stylish design on its English website.
This information can be a useful source of texts for research topics, such as the rhetorical features or stylistics of online information for translation scholars. With a large database, researchers may have a more comprehensive and broad perspective regarding their research topics. For example, it can be used to compare the language used in other electronic devices, such as game consoles. Texts relating to PlayStation 3 (a type of TV game console) may be very different between Chinese and English user groups, as PlayStation 3 was not officially sold in mainland China until 2014 (Carsten, 2014). By using a text clustering technique, researchers may identify the different topics that are focused on. They may find that, for example, the database has information in Chinese on mail order of PS3s from Hong Kong, but such information may not appear in English-language texts. Hence, researchers may be able to observe the impact of different regions and societies on language use itself. It can also be the other way round: the language-related information can also be used for commercial purposes, as different uses of language can also be one dimension according to which consumers can be classified. All of this research would be conducted more efficiently and more convincingly with the support of a large-scale database.
It should be noted that implementing this plan poses several challenges, such as privacy, copyright, quality control and the cost of hosting enormous databases. However, the nature of both translation and translation memory systems determines the form translation memory systems should take. Some commercial translation memory systems already have one or more of the features outlined above. WordFast employs cloud computing to make one of its versions a web-based platform, for example.

To conclude, translation memory systems can become a new kind of media that serves different purposes. Translation memory systems as online platforms could provide a social community for translation professionals. The content of translation memory systems could also be used for information dissemination and language learning or as databases for studying translator behaviour. Such translator behaviour provides a complex primary source for translation process studies and also for business.

6.4 Conclusion

In this chapter, a range of issues are discussed based on the results of two experiments.

First, the significant findings of the evaluation of ESA in translation memory systems were reviewed. ESA has the advantage of being able to retrieve more translation suggestions compared to the Levenshtein distance method, and therefore can be applied in TMS tasks. However, the Levenshtein distance method is not necessarily redundant, because there may still be a long way to go before ESA can actually be implemented in translation memory systems. CR translation suggestions and their relevance to translation are also discussed. The analysis suggests that ESA performance is more related to the number of words in a query: as a matching method used in translation memory systems, ESA is more likely to retrieve sentences ranging from 12~13 and 16~17 words long. Not only is the size of the training data important, but also the knowledge that it covers, because ESA needs to match terms and words from relevant topics in order to measure semantic similarity more properly.

Secondly, an evaluation framework is proposed to measure the use of semantic similarity instruments in translation memory systems. Three elements are specified: the purpose of evaluation, standards for tools and materials, and measures. The difference between the
evaluation of translation resources and that of matching methods are emphasised. The purpose is mainly reflected in procedures to quantify the performance of a specific semantic similarity measure as a matching method. Finally, materials and tools should meet certain standards in order to produce satisfying and convincing results.

Thirdly, two problems with using ESA in translation memory systems are discussed with special attention given to English > Chinese translation. ESA may be enhanced if it could also employ the linguistic information contained in texts. However, the knowledge representation of Chinese is limited because of our current inadequate understanding of Chinese grammatical structure. Further research is recommended to explore a method that represents the meaning of Chinese texts without reference to POS. The second problem is inconsistencies between the various Chinese writing systems, while the commercialisation of Chinese online encyclopaedias may also affect their performance as training data for ESA. Some of these problems are not only technical issues, but are also related to the development of the Internet and the existence of different language policies in different regions.

Lastly, the future perspective of translation memory systems is also outlined. A consequence of involving semantic similarity measures is a new possible interdependence between translation memory systems and online knowledge bases. TMS software developers may play a less important role in the future, but may start to act as designers of different interfaces that are all supported by major data owners. This perspective may provide more opportunities for individual developers to create convenient features as ‘add-ons’ to enrich the user experience. As a result, translation memory systems can become similar to different platforms. A large unified platform can be seen as a new platform that serves not only translators but also different users from other related industries looking for information in different languages.
Chapter 7 Conclusions

This thesis presents a study of the application of ESA in translation memory systems. As mentioned in Section 1.3.1, there are five aims. In this chapter, I summarise the conclusions of these aims and address future work.

7.1 Assessment of Aims of the Thesis

This section assesses the five aims in the light of the findings of the thesis. Each aim constitutes a specific part of this thesis that studies the application of ESA in TMS tasks.

7.1.1 Aim 1: Construction of an operational framework for discussion of the use of semantic processing techniques in Translation Memory Systems

Aim 1 primarily demonstrates that translation memory systems are a type of KMS, as the operational framework provides a new perspective for conceptualising the workflow of translation memory systems within the context of KM. The conceptual framework of KM was modified to serve as an operational framework for the discussion of semantic processing techniques in translation memory systems.

The workflow of translation memory systems comprises a four-step KM process: knowledge capture, knowledge codification, knowledge application and knowledge creation. This conceptualises the interaction between translation memory systems and TM files as a complex sequential process of converting and transforming (i.e. using) knowledge from different categories and in different forms. In this study, it was found that the knowledge capture step is relatively successful, but knowledge codification (i.e., managing TM files) is not effective. This is mainly because the matching methods used in current translation memory systems are inadequate.
The research done for Aim 1 indicates the feasibility of involving semantic similarity measures to enhance the performance of translation memory systems, and hence providing a conceptual framework that contains a proper definition of concepts for evaluating ESA in TMS tasks.

7.1.2 Aim 2: Investigation of semantic similarity measures

Aim 2 both reviews major semantic similarity measures and identifies what kind of semantic similarity measure can best enhance the knowledge management process for translation memory systems as a type of KMS. For the purpose of this study, semantic similarity measures are methods of quantifying the relationship between queries and translation suggestions according to their semantic meaning. After reviewing two kinds of semantic similarity measure, thesaurus methods and corpus-based methods, ESA is chosen. Compared with thesaurus methods, corpus-based methods overcome the problem of lacking sufficient domain knowledge. ESA is a technique very similar to corpus-based methods, but ESA respects a sensible knowledge structure because it employs knowledge from Wikipedia.

It is expected that the use of ESA would permit a TMS to retrieve a greater number of translation suggestions. Thus, ESA is used as a representative semantic similarity measure to enhance the knowledge capture step in translation memory systems as a type of KMS. Aim 2 predicts the potentials of using ESA in translation memory systems and therefore an actual implementation to test possibility of ESA in TMS tasks is needed.

7.1.3 Aim 3: Design of a software platform for evaluating the possibility of using Explicit Semantic Analysis

For Aim 3, a software platform is designed to test the performance of ESA in TMS tasks. It included the creation of an ESA implementation and preparation of relevant information and materials. For the needs of the evaluation, the software platform was modified accordingly, and proper test collections were used to establish representative and conceivable results. The most important feature of the implementation is that a software platform was designed to simulate the core function of translation memory systems, instead of building completed prototype
software. In this study, the core function of translation memory systems is recognised as the IR task of retrieving translation suggestions based on queries given. The software platform is primarily used to conduct a simulation of the core function, called pseudo-TMS tasks. Moreover, the evaluation was only implemented in English texts because the TMS task is a type of monolingual IR task for short texts. Translation memory systems do not translate queries to match translation suggestions, but pair queries with target texts contained in translation suggestions. In this circumstance, the use of neither a bilingual ESA implementation nor a bilingually aligned corpus was necessary.

Sorg's implementation of ESA (2010) was modified as a software platform for the evaluation by adding two programs. One program ‘ComputeESASimilarity15Results.java’ was designed to implement the IR tasks of translation memory systems. Another program ‘SimilarityComputeTwoUnits.java’ was designed to measure the semantic similarities between a pair of texts. These were used to collect data addressing the research questions proposed in Section 1.3.2. As a result, the software platform became an IR platform, called as the ESA IR platform.

The ESA IR platform also required relevant information and materials. An appropriate Wikipedia dump (newwiki-20130821-pages-meta-history.xml.bz2) was chosen, as its size was suitable for ESA process training. The test collections were categorised according to genres rather than topics, because texts from different genres were more indicative of discriminative features than topics. Test collections are from different genres, including technical reports, popular scientific articles and news articles. Several tools were used to obtain the performance results of ESA and details of the test collections, because the details of the test collections (textual factors) were also related to the performance of the study. The following textual factors were considered: size, length of queries, type/token ratio (TTR) and average sentence length (ASL).

As a semantic similarity measure, ESA has not been implemented in translation memory systems, so its effects have not been studied. Aim 3 provides a technical solution for testing the performance of ESA for use in translation memory systems.
7.1.4 Aim 4: Defining of measures for evaluating Explicit Semantic Analysis

For Aim 4, several measures were defined to determine the possibility of using ESA for the purpose of the evaluation. As well as newly defined measures, the concept of a 'valid translation suggestion' was developed in this evaluation. The valid translation suggestion is defined as a sentence with one or more components similar or identical to the corresponding query.

The possibility is defined as ESA's ability to provide more translation suggestions than would be possible by using the Levenshtein distance method alone, and the measures are primarily to describe and demonstrate the similarity scores of translation suggestions given by the ESA IR platform. In this circumstance, translation memory systems should aim to provide valid translation suggestions that are potentially useful to translators. The method’s efficiency would depend on whether it was able to assign sufficient similarity scores (ESA similarity scores) for valid translation suggestions that would possibly have been missed by the Levenshtein distance method. That is because higher similarity scores were more likely to be retrieved by translators. The ESA similarity scores were measured by mean, median and SD. The first two summarised the average values of ESA similarity scores for valid translation suggestions, and the SD was used to show the spread of ESA similarity scores.

This evaluation differed significantly from the evaluation measures used in the IR evaluation, despite the technical similarities between translation memory systems and IR. The two IR evaluation metrics (recall and precision) are not applicable in this evaluation because they were difficult to obtain in TMS evaluations. In fact, the involvement of the ESA technique does not change the fact that TMS is still a type of CAT tool that minimises translator workload by providing more translation suggestions.

After confirming the proper measures to show the performance of ESA, two research questions are proposed to provide a comprehensive understanding of the performance of ESA. Variations in performance attributable to genre differences and textual factors affecting ESA were observed. A result was defined as a pair comprising a query and a sentence matched from the test collection. Accordingly, the results totalled 120 examples; that is, 40 for each test collection. Question A tested whether ESA was applicable for translation memory systems by observing the translation suggestions and their ESA similarity scores. Question B tested how textual factors affected the performance of ESA. After gaining a comprehensive understanding
of ESA performance, it was reasonable to assume that some textual factors in the test collections affected the threshold values of ESA, as different test collections had different textual factors.

The measures for evaluating ESA are crucially important to quantify the performance of semantic similarity measures in the use of translation memory systems. The accomplishment of Aim 4 not only finds a way of demonstrating the performance of ESA to compare with other matching methods, but also structures the research questions and experimental design.

7.1.5 Aim 5: Examination of relevant issues regarding involvement of Explicit Semantic Analysis in translation memory systems

For Aim 5, four relevant issues are discussed concerning the employment of the results in the development of TMS.

First, important evaluation results were highlighted and their relevance to translation memory systems was analysed. Based on the results, it was deemed feasible to employ ESA as a matching method in TMS tasks. In the meantime, the evaluation method concluded with an evaluation framework for using semantic processing techniques in TMS.

Second, some problems related to using ESA in translation memory systems were discussed. Notably, none was related to technical issues. Other problems, such as China's language policy and the impact of commercial interests, affected the quality of the knowledge base.

Lastly, the potential future of translation memory systems was outlined. Using online resources, the potential for change in the technical architecture of translation memory systems and functions was addressed. In the future, translation memory systems could serve as an online community for both translation professionals and users in the language industry.

The discussion contributes to the understanding of using semantic processing techniques in translation memory systems, and outlining the impact of online resources, particularly knowledge bases, on translation memory systems.
7.2 Future Work

This thesis has evaluated the possibility of using ESA in TMS tasks and conceptualised translation memory systems as a type of KMS. It demonstrates that ESA can be applied in translation memory systems. A possible future for translation memory systems that exploits online resources, in the way envisaged in the thesis, is also outlined. However, some aspects of the research could usefully be extended in future work.

First, future research should test the performance of ESA in other genre groups. This thesis only tested the performance of ESA in three test collections from three different genres, namely, technical reports, financial texts and popular scientific articles. In the evaluation experiments, the ESA similarity score threshold values for valid translation suggestions were different between the genre groups. Future research should concentrate on other high demand areas, such as legal and medical texts. ACA and RN demonstrated more similar features than SciAm; but, it would be worthwhile to determine if those similar features were also reproduced in other test collections. With results available from more test collections, it may be possible to produce a comprehensive outline of how ESA performs.

Secondly, because this thesis proved the possibility of using ESA, an end-user evaluation (i.e., user-centred style evaluation) is necessary. It is feasible to implement ESA as a matching method in an actual TMS. A possible option is to implement ESA in OmegaT because it releases its source code for developers to modify. The perception of new functions generated by ESA should receive priority from researchers. As the evaluation suggested, the ESA IR platform is able to provide conceptual clarification of translated texts. In this study this type of translation suggestion is called a 'CR translation suggestion'. The usefulness of CR translation suggestions is better assessed directly by translators, as it is unlikely that the nature of CR translation suggestions will be able to be judged using mechanical methods, such as the number of keystrokes required for post-editing translation suggestions (Aziz, Specia, and Mitkov, 2013). A comprehensive study using CR/similar translation suggestions could be based on translator reviews. The end-user evaluation, then, should focus on the impact of translation suggestions on translating. Ideally, actual TM files should be used rather than test collections for end-user evaluations. However, this would mean higher requirements for evaluation materials, as bilingually aligned texts and texts to be translated would have to be
prepared. It is possible for translators to explore different uses of translation suggestions based on their hands-on experience.

Thirdly, it is feasible to form an automatic metric for measuring the quality of translation suggestions by employing the same statistics and feedback given by translators from their experiences using an ESA-enhanced TMS. The aim of the automatic metric would be to have a metric similar to BLEU for MT evaluation. This thesis considers many measures of ESA performance and textual factors relating to genres, but the evaluation measures used were primarily for testing ESA as a matching method. The metric used to measure the quality of translation suggestions would require many different judgments. For example, end-user evaluations should be designed to determine an acceptable QTR for translators. Hence, this data can be used as an aspect of the metric that defines the quality of a translation suggestion. If the metric is properly defined, it will reduce the workload of translation suggestion assessments.

Fourthly, the performance of ESA should be compared with other semantic similarity measures using similar evaluation methods. As demonstrated in Section 6.2, this thesis proposes a general framework for evaluating semantic processing techniques. It suggests a method to compare different semantic measures in TMS tasks. Other than focusing on the performance of ESA in different test collections, research should concentrate on the performance of different semantic measures. It is possible that some semantic similarity measures can outperform others. In the light of this, a hybrid semantic similarity measure could be developed. For example, the ESA technique is suitable for retrieving texts of certain lengths, while another semantic similarity measure might be more capable of processing texts of different lengths, or indeed ones with particular rhetorical structures. A hybrid matching method could use multiple semantic similarity measures simultaneously to improve the performance of the translation memory system.

Fifthly, the practical implementation of ESA in translation memory systems can also consider measuring the semantic similarity of two languages directly. It is advised that the possibility of using bilingual ESA implementation should be conducted. As a result, it requires the ESA implementation to be trained on Wikipedia for two languages. Although it is unnecessary to conduct a bilingual semantic similarity measure in this evaluation and practical TMS tasks in the state of arts level, the benefit of bilingual semantic similarity measures is very clear. The greatest benefit would be to enable translation memory systems to use information from
documents which are not necessarily available in both source and target languages. As explained in Section 2.4, translation memory systems only need to match texts according to source languages, and then show the target language as translation suggestions. Obviously, all TM files need to be translated in the first place, and it raises costs at the same time. However, if translation memory systems could compute the semantic similarity of target languages directly, it would greatly enlarge the choice of translation resources for translators, hence improving the chance of matching texts being translated. But bilingual ESA implementation is also a more complex technique compared with the monolingual computation. As can be foreseen, the challenge will be the difference between Wikipedia dumps in different languages. For example, by August 2014 the size of English Wikipedia is 4,582,136 articles, and Chinese 782,239 (Wikipedia, 2014d). Inevitably, some entries from these two Wikipedia dumps are not always matched. Moreover, the contents of Wikipedia entries vary. Different language user groups may use different words to describe the same concepts. For example, concepts such as ‘Tai-chi-ch’uan’ are described very differently in Wikipedia. English Wikipedia emphasises the different styles, historical origin and philosophical traditions of Tai-chi-ch’uan, while the article in the Chinese Wikipedia is even shorter. A possible explanation is that some concepts, such as ‘tai-ji’ (太极), are just common sense to most Chinese people. Not only culturally related concepts, but some technical concepts, such as entropy (information theory), may also be described differently in terms of length of the entry, level of comprehensiveness, and related concepts mentioned in the end of entries. It is possible to assume that different ESA scores can be given, based on those different knowledge bases, especially if two Wikipedia are developed independently, rather than being translated or paraphrased from one to another. Future work should first concentrate on how ESA can overcome the differences between Wikipedia dumps and impact on ESA similarity scores. An investigation of Wikipedia from the perspective of a knowledge base for the use of translation memory systems is also necessary.
7.3 Final Word

This thesis contributes to the understanding of semantic processing techniques in TMS tasks by evaluating ESA, a semantic similarity measure that employs open online knowledge bases, such as Wikipedia. The framework of knowledge management is used to conceptualise the TMS workflow. This work demonstrates that translation memory systems would very probably be improved by incorporating a further matching method. The evaluation method focuses on the ability of ESA to add a further useful matching methodology to translation memory systems by providing more translation suggestions. The results of the evaluation demonstrate that ESA is capable of retrieving potentially useful translation suggestions that are missed by the traditional matching method and provides an understanding of how ESA is affected by various textual factors related to different genres. Some relevant issues concerning the use of ESA in translation memory systems are also discussed, particularly the effects of online resources on translation memory systems. It is hoped that this thesis provides an initial understanding of using ESA in TMS tasks and will support the development of translation memory systems in the future.
Appendices

Appendix A: Sources of Potential Queries

A.1 Website for ACA

http://www.aaib.gov.uk/home/index.cfm

A.2 Websites for RN

http://www.bbc.co.uk/news/business/
http://www.commodities-now.com/
http://blogs.wsj.com/money
http://www.reuters.com/article/
http://talkingbiznews.com/
http://www.bls.gov/opub/mlr/2012/03
http://www.wto.org/english/
http://www.moneycontrol.com/
http://documents.foodandwaterwatch.org/doc/
http://www.marketwatch.com/s
http://www.ft.com/
http://www.economist.com/
A.3 Websites for SciAm

http://www.fanyitie.com/
http://www.tingvoa.com/
http://www.dxy.cn/bbs/
http://www.tingvoa.com/
Appendix B: Data of experiments A

Appendix B.1 Data of ESA matching experiment (ACA)

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>WN</th>
<th>Translation Suggestion</th>
<th>Original cosine value</th>
<th>ESA Similarity Score</th>
<th>WN2</th>
<th>Type</th>
<th>LS Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>It is recommended that the Federal Aviation Administration initiate action for making inert the Honeywell International RESCU406AFN fixed Emergency Locator Transmitter system in Boeing 787 aircraft.</td>
<td>26</td>
<td>The Boeing 747 aircraft was originally certificated by the Federal Aviation Administration</td>
<td>0.974</td>
<td>9.429</td>
<td>12</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>2</td>
<td>At the rear of the passenger cabin they observed indications of fire above the ceiling panels.</td>
<td>16</td>
<td>The evidence indicated that there had been no pre-crash or post-crash fire</td>
<td>0.979</td>
<td>9.524</td>
<td>12</td>
<td>1</td>
<td>0.42</td>
</tr>
<tr>
<td>3</td>
<td>At the rear of the passenger cabin they observed indications of fire above the ceiling panels</td>
<td>16</td>
<td>There was no pathological indication of an in-flight fire and no evidence that any of the victims had been injured by shrapnel from the explosion</td>
<td>0.979</td>
<td>9.519</td>
<td>25</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>4</td>
<td>At the rear of the passenger cabin they observed indications of fire above the ceiling panels.</td>
<td>16</td>
<td>There were no indications of fire and the commander did not order an evacuation</td>
<td>0.972</td>
<td>9.366</td>
<td>14</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>The lack of damage on the recovered areas of the bearing outer race indicated that the initiation was not entirely consistent with the understood characteristics of spalling</td>
<td>27</td>
<td>No evidence of indication or of whether gyroscope was rotating</td>
<td>0.976</td>
<td>9.467</td>
<td>10</td>
<td>2</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>The serious incident occurred to an Airbus A319-111 aircraft operating a scheduled passenger flight between Spain and UK.</td>
<td>18</td>
<td>It continued to operate until it was dumped.</td>
<td>0.931</td>
<td>8.529</td>
<td>8</td>
<td>2</td>
<td>0.47</td>
</tr>
</tbody>
</table>

In Appendix B, LS score stands for Levenshtein distance score. WN stands for word number. Type 1 refers to formally similar translation suggestions; Type 2 refers to conceptually related translation suggestions.
<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>The serious incident occurred to an Airbus A319-111 aircraft operating a scheduled passenger flight between Spain and UK.</td>
<td></td>
<td>18</td>
<td></td>
<td></td>
<td>0.44</td>
</tr>
<tr>
<td>8</td>
<td>The inability of the flight crew to reconfigure the electrical system.</td>
<td>The crew did not report any defect in the electrical system.</td>
<td>11</td>
<td></td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>9</td>
<td>The accident probably occurred as a result of the pilot attempting an unplanned rolling manoeuvre.</td>
<td>The commander's demand for right roll lasted for less than three seconds</td>
<td>15</td>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>10</td>
<td>This contained recommendations to operators as to what maintenance action should be carried out following a cabin fumes event.</td>
<td>Initially, the operator in this case was carrying out annual audits of the maintenance organisation.</td>
<td>19</td>
<td></td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>11</td>
<td>And also what scheduled maintenance inspections of the air conditioning systems and engines should be conducted.</td>
<td>The outlet air from both air conditioning packs is directed to a cabin and flight deck distribution system.</td>
<td>16</td>
<td></td>
<td></td>
<td>0.42</td>
</tr>
<tr>
<td>12</td>
<td>The investigation identified the following causal factors in this incident.</td>
<td>The investigation into G-OJEM's failure identified no further factors that were likely to have affected the HPT vibratory characteristics.</td>
<td>10</td>
<td></td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>13</td>
<td>There was no flight deck indication of brake system malfunction, and this delayed the crew's recognition of the loss of braking.</td>
<td>and at that moment had been accurately assessed, indicated that there was no major mechanical malfunction had been an intention to dump the recovery system.</td>
<td>21</td>
<td></td>
<td></td>
<td>0.52</td>
</tr>
<tr>
<td>14</td>
<td>The cause of the braking loss could not be positively established.</td>
<td>This provided no signs of a failure, malfunction or anomaly of the aircraft or its equipment likely to have been relevant to the cause of the accident.</td>
<td>11</td>
<td></td>
<td></td>
<td>0.47</td>
</tr>
<tr>
<td>15</td>
<td>The cause of the braking loss could not be positively established.</td>
<td>The cause of the failure could not be assessed.</td>
<td>11</td>
<td></td>
<td></td>
<td>0.71</td>
</tr>
<tr>
<td>16</td>
<td>The findings raised concerns about the aircraft manufacturer's procedures intended to ensure design quality.</td>
<td>Consequently, the quality of spar design and manufacture is satisfactory.</td>
<td>17</td>
<td></td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>The findings raised concerns about the aircraft manufacturer’s procedures intended to ensure design quality and continued airworthiness</td>
<td>17</td>
<td>The repair was inadequate either in design or implementation.</td>
<td>0.990</td>
<td>9.783</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>At a groundspeed of around 70 kt the brakes ceased operating, for about 17 second</td>
<td>15</td>
<td>The stick shaker operated intermittently for about 30 seconds.</td>
<td>1.000</td>
<td>9.990</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>The aircraft commenced a descent from 2,000 ft to the published minimum descent altitude of 740 ft whilst still 11 nm from the runway threshold</td>
<td>25</td>
<td>The highest rate of descent was 8,500 ft/min at 30,000 feet.</td>
<td>0.972</td>
<td>9.367</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>The aircraft commenced a descent from 2,000 ft to the published minimum descent altitude of 740 ft whilst still 11 nm from the runway threshold</td>
<td>25</td>
<td>Outbound the aircraft commenced a descent from 3,000 ft</td>
<td>0.972</td>
<td>9.377</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>There was no flight deck indication of brake system malfunction, and this delayed the crew’s recognition of the loss of braking</td>
<td>21</td>
<td>The Stall Recovery Fail lamps light steadily to indicate a failure of one half of the duplicated stall recovery system.</td>
<td>0.983</td>
<td>9.616</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>The aircraft had insufficient nose-up pitch attitude at the point of entry to ensure the safe execution of the manoeuvre.</td>
<td>20</td>
<td>the helicopter would have had the lowest apparent weight and therefore the lowest stability in the absence of any other influence.</td>
<td>0.953</td>
<td>8.985</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>A somatogravic illusion is a non-visual illusion that produces a false sensation of helicopter attitude.</td>
<td>15</td>
<td>False sensations about the pitch attitude of the aircraft are caused by a misinterpretation of the gravity vertical, known as ‘somatogravic illusion’.</td>
<td>0.942</td>
<td>8.759</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>The accident probably occurred as a result of the pilot attempting an unplanned rolling manoeuvre.</td>
<td>15</td>
<td>A pronounced dip in the runway prevented the pilots from seeing the end of the paved surface until late in the ground roll.</td>
<td>0.985</td>
<td>9.650</td>
<td></td>
</tr>
</tbody>
</table>
Manual wheel braking commence shortly after main wheel touchdown. Consequently, although touchdown occurred slightly beyond the TDZ, had the brakes not malfunctioned it should have been possible to stop the aircraft within the main runway using manual braking.

The aircraft had insufficient nose-up pitch attitude at the point of entry to ensure the safe execution of the manoeuvre. Consequently, although touchdown occurred slightly beyond the TDZ, had the brakes not malfunctioned it should have been possible to stop the aircraft within the main runway using manual braking.

The aircraft commenced a descent from 2,000 ft to the published minimum descent altitude of 740 ft whilst still 11 nm from the runway threshold. It is very unlikely that the descent below the outbound minimum altitude was deliberate.

The aircraft landed uneventfully at Bristol, with the radios and several other systems still inoperative. VHF radio communications with Gatwick ATC were satisfactory.

Manual wheel braking commence shortly after main wheel touchdown. They commenced manual fuel transfer at 0347 hrs.

The aircraft landed uneventfully at Bristol, with the radios and several other systems still inoperative. No further radio transmissions were heard from the aircraft.

The cause of the braking loss could not be positively established. one of the causes of the loss of speed.

The two aircraft involved in the accident were engaged on air experience flights when they collided in midair. Each boot incorporates two interleaved sets of integral tubes to which either pressure or suction can be applied.

The aircraft systems remained in a significantly degraded condition for the remainder of the flight. It was thus highly unlikely that both diaphragms had ruptured before the accident.

and also what scheduled maintenance inspections of the air conditioning systems and engines should be conducted. The cabin pressurisation, air conditioning and various other systems on the aircraft require pressurized air, which is supplied by the engines.
<table>
<thead>
<tr>
<th>No.</th>
<th>Text</th>
<th>Text (Expanded)</th>
<th>Probability</th>
<th>Similarity</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>There was no flight deck indication of brake system malfunction, and this delayed the crew’s recognition of the loss of braking.</td>
<td>The lack of a flight deck indication of the brake failure very probably made it more difficult for the flight crew to assess the cause of the problem.</td>
<td>0.953</td>
<td>8.977</td>
<td>28</td>
</tr>
<tr>
<td>36</td>
<td>An elevator down spring assists in lowering the nose of the aircraft to prevent a stall caused by the aft CG position.</td>
<td>Calculation showed that the feather blade angle with the washer in the position found would be approximately 70° instead of the specified 80°</td>
<td>0.920</td>
<td>8.310</td>
<td>23</td>
</tr>
<tr>
<td>37</td>
<td>The aircraft stalled shortly before entering some treetops.</td>
<td>The commander subsequently performed a manual landing, with touchdown occurring at about 0410 hrs</td>
<td>0.985</td>
<td>9.669</td>
<td>14</td>
</tr>
<tr>
<td>38</td>
<td>Each aircraft was likely to have been obscured from the view of the pilot of the other aircraft by his aircraft’s canopy structure.</td>
<td>though from the pilot's point of view speed may be the dominant consideration.</td>
<td>0.948</td>
<td>8.871</td>
<td>13</td>
</tr>
<tr>
<td>39</td>
<td>The accident probably occurred as a result of the pilot attempting an unplanned rolling manoeuvre.</td>
<td>The initial ground roll phase of the takeoff was unremarkable and the aircraft did not behave abnormally until it became airborne.</td>
<td>0.992</td>
<td>9.815</td>
<td>21</td>
</tr>
<tr>
<td>40</td>
<td>There was some disruption of the fuselage before it struck the ground.</td>
<td>In this incident, fuel had been deliberately jettisoned and the fuel leak ceased before the aircraft landed</td>
<td>0.920</td>
<td>8.310</td>
<td>17</td>
</tr>
</tbody>
</table>
## Appendix B.2 Data of ESA matching experiment (RN)

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>WN</th>
<th>Translation Suggestion</th>
<th>Original cosine value</th>
<th>ESA Similarity Score</th>
<th>WN2</th>
<th>Type</th>
<th>LS Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Other major companies have also tapped the market.</td>
<td>8</td>
<td>He also said that petroleum futures are a major market for oil companies as well as for commodity traders</td>
<td>0.999</td>
<td>9.967</td>
<td>19</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>Exports are growing, but too slowly to rescue the economy</td>
<td>10</td>
<td>But imports do not seem to be picking up much because the Japanese economy remains sluggish, he said</td>
<td>0.962</td>
<td>9.164</td>
<td>18</td>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>Higher U.S. exports narrowed the country's current account deficit in the second quarter.</td>
<td>13</td>
<td>Current account deficit could be offset by an inflow of foreign funds into the U.S.</td>
<td>0.937</td>
<td>8.652</td>
<td>15</td>
<td>1</td>
<td>0.41</td>
</tr>
<tr>
<td>4</td>
<td>A devaluation in the early 1990s helped Britain export its way out of recession.</td>
<td>14</td>
<td>Its plan is expected to help increase economic growth led by domestic demand, officials said.</td>
<td>0.931</td>
<td>8.534</td>
<td>14</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>Russia will maintain its long-standing military and economic support for Syrian President Bashar al-Assad.</td>
<td>14</td>
<td>Russia was trying to build up its military presence in the region.</td>
<td>0.934</td>
<td>8.581</td>
<td>12</td>
<td>1</td>
<td>0.23</td>
</tr>
<tr>
<td>6</td>
<td>The dollar index was last little changed after the previous session's 1.1 percent drop.</td>
<td>14</td>
<td>a senior economist predicted the US dollar would decline another 30 pct by year-end, but said he foresees no significant change in US interest rates</td>
<td>0.987</td>
<td>9.704</td>
<td>25</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>The dollar index was last little changed after the previous session's 1.1 percent drop.</td>
<td>14</td>
<td>The Fed has to determine if this represents a fundamental change for the dollar</td>
<td>0.994</td>
<td>9.855</td>
<td>13</td>
<td>1</td>
<td>0.38</td>
</tr>
<tr>
<td>8</td>
<td>The surge in bank lending has come as bond issuance in the region has declined.</td>
<td>15</td>
<td>Treasury bonds has been bought by the Japanese.</td>
<td>0.856</td>
<td>7.183</td>
<td>7</td>
<td>1</td>
<td>0.38</td>
</tr>
<tr>
<td>9</td>
<td>Many of the loans are five years in duration, an increase from three years previously</td>
<td>15</td>
<td>Lending is now about one billion a year</td>
<td>0.893</td>
<td>7.824</td>
<td>8</td>
<td>1</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>Despite the downward revisions to previous months, the trend still shows an improving labor market.</td>
<td>15</td>
<td>The updated calculations, which may take several months to complete, are expected to show substantial increases in labor market.</td>
<td>0.987</td>
<td>9.700</td>
<td>19</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>11</td>
<td>US manufacturing employment recovered to nearly 12 million in mid-2012 but since then has stagnated.</td>
<td>15</td>
<td>Interest rate rebound to halt further dollar depreciation.</td>
<td>0.988</td>
<td>9.724</td>
<td>8</td>
<td>2</td>
<td>0.46</td>
</tr>
<tr>
<td>12</td>
<td>The US government has painted a bearish outlook for the prices of corn, wheat and soyabean</td>
<td>16</td>
<td>But he cautioned that the short-term outlook is very sensitive to crude oil prices.</td>
<td>0.974</td>
<td>9.424</td>
<td>14</td>
<td>1</td>
<td>0.36</td>
</tr>
<tr>
<td>13</td>
<td>The US government has painted a bearish outlook for the prices of corn, wheat and soyabean</td>
<td>16</td>
<td>Although prices are now considerably lower, consumers and producers have been unable to agree on re-introduction.</td>
<td>0.987</td>
<td>9.714</td>
<td>16</td>
<td>2</td>
<td>0.43</td>
</tr>
<tr>
<td>14</td>
<td>It would hurt export competitiveness in other countries and could trigger large capital inflows to China and push up inflation</td>
<td>20</td>
<td>This inflow of funds into the stock exchange, occurring also in other countries, may continue</td>
<td>0.874</td>
<td>7.477</td>
<td>15</td>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>15</td>
<td>It would hurt export competitiveness in other countries and could trigger large capital inflows to China and push up inflation</td>
<td>20</td>
<td>China may have to import</td>
<td>0.963</td>
<td>9.184</td>
<td>5</td>
<td>2</td>
<td>0.38</td>
</tr>
<tr>
<td>16</td>
<td>It would hurt export competitiveness in other countries and could trigger large capital inflows to China and push up inflation</td>
<td>20</td>
<td>It also found that trade among developing countries was the biggest factor in that expansion.</td>
<td>0.904</td>
<td>8.018</td>
<td>15</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>17</td>
<td>Zhao said China should lower the cost of exports and control the export of goods that incur too much loss.</td>
<td>20</td>
<td>China should also earn more foreign exchange from tourists and contracted labour abroad</td>
<td>0.984</td>
<td>9.633</td>
<td>13</td>
<td>2</td>
<td>0.47</td>
</tr>
<tr>
<td>18</td>
<td>The value of syndicated loans in Asia Pacific so far this year has reached billion, an 8% increase from the same period last year</td>
<td>24</td>
<td>Although it is difficult to forecast the extent of the profit improvement this year, the gain should be significant, he added</td>
<td>0.900</td>
<td>7.944</td>
<td>20</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>19</td>
<td>The value of syndicated loans in Asia Pacific so far this year has reached billion, an 8% increase from the same period last year</td>
<td>24</td>
<td>This year began well, with the performance in January and February at least equal to the same period last year</td>
<td>0.912</td>
<td>8.165</td>
<td>20</td>
<td>1</td>
<td>0.24</td>
</tr>
<tr>
<td>20</td>
<td>The value of the dollar's net long position rose to $20.08 billion in the week ended September 3, from $15.82 billion the previous week.</td>
<td>24</td>
<td>M-1 money supply for the week ended February 23, reported today, rose 1.9 billion dlrs to 738.5 billion.</td>
<td>0.960</td>
<td>9.123</td>
<td>18</td>
<td>1</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>The price of gold on the spot market slid 4.2 per cent yesterday, compounding recent falls to hit a low of $1,223.54 a troy ounce.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Speculative demand, which influences the gold price on futures markets, has also risen.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Countries would coordinate counter-cyclical regulatory measures, thereby preventing capital from skirting them by moving across borders.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>This inflow of funds into the stock exchange, occurring also in other countries, may continue.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The minister said some of the existing 30 Sayam Bank branches will be merged with their Krung Thai counterparts, while others will continue operating but under Krung Thai's name.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Banks’ capital simply refers to their equity funding – how much of their liabilities are owned by shareholders rather than being owed to creditors as some form of debt.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Banks may henceforth use these as reference rates in their lending, it said.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>We explained that the process of restructuring the economy away from its dependence on exports toward a balance between domestic and external demand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A Fed spokesman said that all of the borrowings were made yesterday by fewer than half the banks.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The value of credits given cannot exceed either one-fifth of the combined capital and reserves of the bank itself or two-fifths of the value of the stake owned in the bank.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Markets are still very much in danger.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Markets are still very much in danger.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>There are no other markets.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>But imports do not seem to be picking up much because the Japanese economy remains sluggish.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The ban had threatened food shortages of imported wheat, fresh vegetables and medicines. Inflation in food items escalated further. U.S. gold futures dropped as much as 4.6 percent on the decision. Transactions of paper gold also fell up to 75 pct of the volume prior to imposition of the tax. The government now imposes a 15 pct tax on physical trades exceeding 37,500 yen for gold jewellery and coins and a 2.5 yen tax per 10000 yen on futures transactions. Higher U.S. exports narrowed the country's current account deficit in the second quarter. The current account balance of payments deficit was seen at 934 mln dlrs in 1987. We are beginning to see momentum pick up in the US. It said recovery will begin during the second quarter. Growth is likely to remain modest by historical standards over the medium term due to fiscal consolidation. EPA officials told reporters the underlying trend of the economy is firm but growth is slow due to sluggish exports. Britain’s trade deficit shrank from 4 percent of GDP in 2007 to around 1% of GDP by early 2011. A deeper concern is that Britain has become too dependent on moribund rich-world markets. The US government has painted a bearish outlook for the prices of corn, wheat and soyabean. Pioneer Sugar was expected to recommend acceptance of the bid, through which CSR would benefit from the bottoming out of a cyclical downturn in sugar prices. Anything we do, we will be washed frustration Over Japan's large trade surplus. To do something about Japan's huge trade surplus. US manufacturing employment recovered to nearly 12 million in mid-2012 but since then has stagnated. If the dollar resumes its slide, the EMS could be in for more turbulence. Britain’s trade deficit shrank from 4 percent of GDP in 2007 to around 1% of GDP by early 2011.
### Appendix B.3 Data of ESA matching experiment (SciAm)

<table>
<thead>
<tr>
<th>No.</th>
<th>Query</th>
<th>WN</th>
<th>Translation Suggestion</th>
<th>Original cosine value</th>
<th>ESA Similarity Score</th>
<th>WN2</th>
<th>Type</th>
<th>LS Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Even at low doses, the radiation can damage genes that may increase the patient's risk of developing cancer later.</td>
<td>19</td>
<td>Eventually the cancer-related genes are hit.</td>
<td>0.997</td>
<td>9.922</td>
<td>6</td>
<td>1</td>
<td>0.19</td>
</tr>
<tr>
<td>2</td>
<td>Their findings, published in Nature, tie up loose ends that other theories cannot account for.</td>
<td>15</td>
<td>The early instability theory still has some loose ends.</td>
<td>0.925</td>
<td>8.421</td>
<td>9</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>3</td>
<td>A solar flare erupted from the sun in an impressive display captured by NASA camera, but scientists say the medium-sized event will have a minimal impact, if any, on Earth.</td>
<td>30</td>
<td>The sun attracts Earth by warping the spacetime around it.</td>
<td>0.893</td>
<td>7.812</td>
<td>10</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>humans are increasingly affected by distractions and have more trouble switching between tasks</td>
<td>13</td>
<td>The human brain has the capability to rewire itself to some extent.</td>
<td>0.877</td>
<td>7.529</td>
<td>12</td>
<td>1</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>the power of the devices is staggeringly unbelievable</td>
<td>8</td>
<td>It is a classic reductionist approach, and it can be very powerful.</td>
<td>0.943</td>
<td>8.762</td>
<td>12</td>
<td>1</td>
<td>0.44</td>
</tr>
<tr>
<td>6</td>
<td>Other experiments typically examine the effect of exercise on the hippocampus, the brain's primary memory center, but the object-recognition task involves activity in the perirhinal cortex</td>
<td>26</td>
<td>HIPPOCAMPAL Mediates learning and memory formation, intertwined functions that are impaired in schizophrenia.</td>
<td>0.814</td>
<td>6.522</td>
<td>13</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>7</td>
<td>Even at low doses, the radiation can damage genes that may increase the patient's risk of developing cancer later</td>
<td>19</td>
<td>Aneuploidy could lead to genomic instability early on and later mutate known cancer genes.</td>
<td>0.997</td>
<td>9.925</td>
<td>14</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>Dark energy will continue to push galaxies ever faster away until they fade completely from view.</td>
<td>16</td>
<td>Or the universe might even recollapse if dark energy density falls to a negative value.</td>
<td>0.952</td>
<td>8.948</td>
<td>15</td>
<td>2</td>
<td>0.47</td>
</tr>
<tr>
<td>9</td>
<td>The alternative is that the remains of the star are injecting energy into the debris, and this could happen if these supernovae produced a rapidly spinning magnetar</td>
<td>27</td>
<td>Though exceedingly faint, the AXP fades in and out with the x-ray period of the neutron star</td>
<td>0.885</td>
<td>7.675</td>
<td>17</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>10</td>
<td>The alternative is that the remains of the star are injecting energy into the debris, and this could happen if these supernovae produced a rapidly spinning magnetar</td>
<td>27</td>
<td>If both SGRs and AXPs are magnetars, then magnetars plausibly constitute a substantial fraction of all neutron stars</td>
<td>0.936</td>
<td>8.623</td>
<td>18</td>
<td>2</td>
<td>0.33</td>
</tr>
<tr>
<td>11</td>
<td>It not only promises to revolutionize semiconductor, but could also lead to breakthroughs in fundamental quantum physics research</td>
<td>18</td>
<td>In dealing with quantum physics, it is essential to specify precisely what physical quantities are to be measured</td>
<td>0.939</td>
<td>8.680</td>
<td>18</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>12</td>
<td>They also showed improved performance in tests for memory and sustained attention.</td>
<td>12</td>
<td>Researchers have demonstrated that adults with cerebellar damage show delays and tend to make mistakes in spatial reasoning tests</td>
<td>0.867</td>
<td>7.365</td>
<td>19</td>
<td>1</td>
<td>0.34</td>
</tr>
<tr>
<td>13</td>
<td>Scientists have been expecting an increase in solar activity because the sun is moving into a more volatile period of an 11-year cycle in which its magnetic field reverses its orientation</td>
<td>31</td>
<td>Deep inside the sun, these two rates are similar, and the magnetic field is able to organize itself on large scale</td>
<td>0.829</td>
<td>6.741</td>
<td>21</td>
<td>2</td>
<td>0.41</td>
</tr>
<tr>
<td>14</td>
<td>Dark energy will continue to push galaxies ever faster away until they fade completely from view.</td>
<td>16</td>
<td>But if the dark energy density decreases and matter becomes dominant again, our cosmic horizon will grow, revealing more of the universe.</td>
<td>0.923</td>
<td>8.376</td>
<td>22</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>15</td>
<td>But large doses of resveratrol are required to show any effect, so chemical mimics like SRT-1720 were developed to activate sirtuin at much lower doses.</td>
<td>25</td>
<td>Instead we had a nearly working prototype to show,</td>
<td>0.927</td>
<td>8.461</td>
<td>9</td>
<td>2</td>
<td>0.41</td>
</tr>
<tr>
<td>16</td>
<td>Sirtuins have proved to be highly interesting proteins</td>
<td>8</td>
<td>Such proteins remain highly valued because of their accuracy</td>
<td>0.875</td>
<td>7.505</td>
<td>9</td>
<td>1</td>
<td>0.34</td>
</tr>
<tr>
<td>17</td>
<td>The power of the devices is staggeringly unbelievable</td>
<td>8</td>
<td>It is an immensely powerful creative process</td>
<td>0.964</td>
<td>9.197</td>
<td>7</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Other experiments typically examine the effect of exercise on the hippocampus, the brain’s primary memory center, but the object-recognition task involves activity in the perirhinal cortex</td>
<td>26</td>
<td>In contrast, higher synesthetes show less activation at these earlier levels</td>
<td>0.976</td>
<td>9.455</td>
<td>11</td>
<td>2</td>
<td>0.21</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>19</td>
<td>Other experiments typically examine the effect of exercise on the hippocampus, the brain’s primary memory center, but the object-recognition task involves activity in the perirhinal cortex</td>
<td>26</td>
<td>HIPPOCAMPAL Mediates learning and memory formation, intertwined functions that are impaired in schizophrenia</td>
<td>0.814</td>
<td>6.518</td>
<td>13</td>
<td>2</td>
<td>0.34</td>
</tr>
<tr>
<td>20</td>
<td>Obese mice are a standard research tool, but experts differ as to how relevant they are to humans.</td>
<td>18</td>
<td>Self-reported ancestry will continue to be a potentially useful diagnostic tool for physicians.</td>
<td>0.885</td>
<td>7.678</td>
<td>13</td>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>21</td>
<td>Chemical mimics were developed to activate sirtuin at much lower doses.</td>
<td>11</td>
<td>Such activities can lead to poor compliance with treatment, increasing propensities toward violence.</td>
<td>0.897</td>
<td>7.897</td>
<td>13</td>
<td>2</td>
<td>0.32</td>
</tr>
<tr>
<td>22</td>
<td>Most orbit insertions are tricky because a speeding craft has to slow down or risk overshooting its target.</td>
<td>18</td>
<td>Until recently, astronomers fully expected to see gravity slowing down the expansion of the cosmos.</td>
<td>0.899</td>
<td>7.920</td>
<td>15</td>
<td>1</td>
<td>0.38</td>
</tr>
<tr>
<td>23</td>
<td>A tiny second moon may once have orbited Earth.</td>
<td>9</td>
<td>It slows the ascent of baseballs and holds the moon in orbit around the earth.</td>
<td>0.891</td>
<td>7.780</td>
<td>15</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>24</td>
<td>Sirtuins have proved to be highly interesting proteins.</td>
<td>8</td>
<td>It very likely has some benefits, especially from a high intake of fruits and vegetables.</td>
<td>0.875</td>
<td>7.498</td>
<td>15</td>
<td>1</td>
<td>0.46</td>
</tr>
<tr>
<td>25</td>
<td>The researchers intend to look for potential subsurface gaseous signs of life with an innovative drill they have developed</td>
<td>19</td>
<td>And so if you wanted an explanation for life, it had to be about the molecular basis for life</td>
<td>0.952</td>
<td>8.962</td>
<td>19</td>
<td>1</td>
<td>0.47</td>
</tr>
<tr>
<td>26</td>
<td>By studying the level of firing of individual neurons in the brain’s posterior cingulate cortex,</td>
<td>15</td>
<td>Perhaps ATP, which is released along with neurotransmitters when axons fire, was somehow escaping along the axon.</td>
<td>0.859</td>
<td>7.225</td>
<td>18</td>
<td>1</td>
<td>0.31</td>
</tr>
<tr>
<td>27</td>
<td>A solar flare erupted from the sun in an impressive display captured by NASA camera, but scientists say the medium-sized event will have a minimal impact, if any, on Earth.</td>
<td>30</td>
<td>In one scenario, the neutrinos would oscillate during their eight-minute journey through the vacuum of space from the sun to the earth.</td>
<td>0.893</td>
<td>7.809</td>
<td>22</td>
<td>2</td>
<td>0.18</td>
</tr>
</tbody>
</table>
It is often depicted as an atomic-scale chicken wire made of carbon atoms and their bonds. Chipmakers induce strain in silicon by bonding it to another crystalline material—in this case, a silicon-germanium blend—for which the lattice spacing is greater.

But it turns out that these unconventional genes do give rise to active RNAs, through which they profoundly alter the behavior of normal genes.

Shannon's entropy does not enlighten us about the value of information, which is highly dependent on context.

No evidence of indication or of whether gyroscope was rotating.

Parents should certainly discuss risk with their provider, but not refuse care that may save and extend their child's life.

Yet they have exactly the values that sustain life.

The agreement indicates that astronomers finally have a consistent account of 14 billion years of cosmic evolution.

Finally, the spectrum of the light seems to indicate the light is coming from material that is extremely hydrogen-poor.
<table>
<thead>
<tr>
<th>Page</th>
<th>Text</th>
<th>Word</th>
<th>Sentence</th>
<th>Context</th>
<th>Character</th>
<th>Character</th>
<th>Character</th>
<th>Character</th>
<th>Character</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>The alternative is that the remains of the star are injecting energy into the debris.</td>
<td>15</td>
<td>This pressure is what prevents the star from collapsing under its own weight.</td>
<td>0.921</td>
<td>8.340</td>
<td>13</td>
<td>1</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>The alternative is that the remains of the star are injecting energy into the debris.</td>
<td>15</td>
<td>But why would a neutron star behave like this?</td>
<td>0.961</td>
<td>9.145</td>
<td>9</td>
<td>2</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Scientists have been expecting an increase in solar activity because the sun is moving into a more volatile period of an 11-year cycle in which its magnetic field reverses its orientation</td>
<td>31</td>
<td>Such behavior occurs in systems as diverse as avalanches on sandpiles and magnetic flares on the sun.</td>
<td>0.873</td>
<td>7.471</td>
<td>17</td>
<td>2</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


Bajaj, B. (2003) A Terminological Approach to the Investigation of Temporal Ordering Relations in English and German Aircraft Accident Reports. PhD. University of Surrey.


Lagoudaki, M. (2009) *Expanding the possibilities of translation memory systems: from the translator's wishlist to the developer's design.* PhD. Imperial College.


Li, Y., McLean, D., Bandar, Z., O'shea, J. D. , Crockett, K. (2006) Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering. 18 (8), pp.1138-1150.


Sorg, P. (2010) Research-ESA (2.0) [Source code].


