WikiCREM: A Large Unsupervised Corpus for Coreference Resolution

Vid Kocijan[1] **Oana-Maria Camburu**[1,2] **Ana-Maria Creţu**[3] **Yordan Yordanov**[1]
Phil Blunsom[1,4] **Thomas Lukasiewicz**[1,2]
[1]University of Oxford, UK [2]Alan Turing Institute, London, UK
[3]Imperial College London, UK [4]DeepMind, London, UK
firstname.lastname@cs.ox.ac.uk, a.cretu@imperial.ac.uk

Abstract

Pronoun resolution is a major area of natural language understanding. However, large-scale training sets are still scarce, since manually labelling data is costly. In this work, we introduce WIKICREM (Wikipedia CoREferences Masked) a large-scale, yet accurate dataset of pronoun disambiguation instances. We use a language-model-based approach for pronoun resolution in combination with our WIKICREM dataset. We compare a series of models on a collection of diverse and challenging coreference resolution problems, where we match or outperform previous state-of-the-art approaches on 6 out of 7 datasets, such as GAP, DPR, WNLI, PDP, WINOBIAS, and WINOGENDER. We release our model to be used off-the-shelf for solving pronoun disambiguation.

1 Introduction

Pronoun resolution, also called coreference or anaphora resolution, is a natural language processing (NLP) task, which aims to link the pronouns with their referents. This task is of crucial importance in various other NLP tasks, such as information extraction (Nakayama, 2019) and machine translation (Guillou, 2012). Due to its importance, pronoun resolution has seen a series of different approaches, such as rule-based systems (Lee et al., 2013) and end-to-end-trained neural models (Lee et al., 2017; Liu et al., 2019). However, the recently released dataset GAP (Webster et al., 2018) shows that most of these solutions perform worse than naïve baselines when the answer cannot be deduced from the syntax. Addressing this drawback is difficult, partially due to the lack of large-scale challenging datasets needed to train the data-hungry neural models.

As observed by Trinh and Le (2018), language models are a natural approach to pronoun resolution, by selecting the replacement for a pronoun that forms the sentence with highest probability. Additionally, language models have the advantage of being pre-trained on a large collection of unstructured text and then fine-tuned on a specific task using much less training data. This procedure has obtained state-of-the-art results on a series of natural language understanding tasks (Devlin et al., 2018).

In this work, we address the lack of large training sets for pronoun disambiguation by introducing a large dataset that can be easily extended. To generate this dataset, we find passages of text where a personal name appears at least twice and mask one of its non-first occurrences. To make the disambiguation task more challenging, we also ensure that at least one other distinct personal name is present in the text in a position before the masked occurrence. We instantiate our method on English Wikipedia and generate the Wikipedia CoREferences Masked (WIKICREM) dataset with 2.4M examples, which we make publicly available for further usage[1]. We show its value by using it to fine-tune the BERT language model (Devlin et al., 2018) for pronoun resolution.

To show the usefulness of our dataset, we train several models that cover three real-world scenarios: (1) when the target data distribution is completely unknown, (2) when training data from the target distribution is available, and (3) the transductive scenario, where the unlabeled test data is available at the training time. We show that fine-tuning BERT with WIKICREM consistently improves the model in each of the three scenarios, when evaluated on a collection of 7 datasets. For example, we outperform the state-of-

[1]The code can be found at https://github.com/vid-koci/bert-commonsense.
The dataset and the models can be obtained from https://ora.ox.ac.uk/objects/uuid:c83e94bb-7584-41a1-aef9-85b0e764d9e3

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4303–4312,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

the-art approaches on GAP (Webster et al., 2018), DPR (Rahman and Ng, 2012), and PDP (Davis et al., 2017) by 5.9%, 8.4%, and 12.7%, respectively. Additionally, models trained with WIKICREM show increased performance and reduced bias on the gender diagnostic datasets WINOGENDER (Rudinger et al., 2018) and WINOBIAS (Zhao et al., 2018).

2 Related Work

There are several large and commonly used benchmarks for coreference resolution, such as (Pradhan et al., 2012; Schäfer et al., 2012; Ghaddar and Langlais, 2016). However, Webster et al. (2018) argue that a high performance on these datasets does not correlate with a high accuracy in practice, because examples where the answer cannot be deduced from the syntax (we refer to them as *hard pronoun resolution*) are underrepresented. Therefore, several hard pronoun resolution datasets have been introduced (Webster et al., 2018; Rahman and Ng, 2012; Rudinger et al., 2018; Davis et al., 2017; Zhao et al., 2018; Emami et al., 2019). However, they are all relatively small, often created only as a test set.

Therefore, most of the pronoun resolution models that address hard pronoun resolution rely on little (Liu et al., 2019) or no training data, via unsupervised pre-training (Trinh and Le, 2018; Radford et al., 2019). Another approach involves using external knowledge bases (Emami et al., 2018; Fähndrich et al., 2018), however, the accuracy of these models still lags behind that of the aforementioned pre-trained models.

A similar approach to ours for unsupervised data generation and language-model-based evaluation has been recently presented in our previous work (Kocijan et al., 2019). We generated MASKEDWIKI, a large unsupervised dataset created by searching for repeated occurrences of nouns. However, training on MASKEDWIKI on its own is not always enough and sometimes makes a difference only in combination with additional training on the DPR dataset (called WSCR) (Rahman and Ng, 2012). In contrast, WIKICREM brings a much more consistent improvement over a wider range of datasets, strongly improving models' performance even when they are not fine-tuned on additional data. As opposed to our previous work (Kocijan et al., 2019), we evaluate models on a larger collection of test sets, showing the usefulness of WIKICREM beyond the Winograd Schema Challenge.

Moreover, a manual comparison of WIKICREM and MASKEDWIKI (Kocijan et al., 2019) shows a significant difference in the quality of the examples. We annotated 100 random examples from MASKEDWIKI and WIKICREM. In MASKEDWIKI, we looked for examples where masked nouns can be replaced with a pronoun, and only in 7 examples, we obtained a natural-sounding and grammatically correct sentence. In contrast, we estimated that 63% of the annotated examples in WIKICREM form a natural-sounding sentence when the appropriate pronoun is inserted, showing that WIKICREM consists of examples that are much closer to the target data. We highlight that pronouns are not actually inserted into the sentences and thus none of the examples sound unnatural. This analysis was performed to show that WIKICREM consists of examples with data distribution closer to the target tasks than MASKEDWIKI.

3 The WIKICREM Dataset

In this section, we describe how we obtained WIKICREM. Starting from English Wikipedia[2], we search for sentences and pairs of sentences with the following properties: at least two distinct personal names appear in the text, and one of them is repeated. We do not use pieces of text with more than two sentences to collect concise examples only. Personal names in the text are called "candidates". One non-first occurrence of the repeated candidate is masked, and the goal is to predict the masked name, given the correct and one incorrect candidate. In case of more than one incorrect candidate in the sentence, several datapoints are constructed, one for each incorrect candidate.

We ensure that the alternative candidate appears before the masked-out name in the text, in order to avoid trivial examples. Thus, the example is retained in the dataset if:

(a) the repeated name appears after both candidates, all in a single sentence; or

(b) both candidates appear in a single sentence, and the repeated name appears in a sentence directly following.

Examples where one of the candidates appears in the same sentence as the repeated name, while the

[2] https://dumps.wikimedia.org/enwiki/ dump id: enwiki-20181201

other candidate does not, are discarded, as they are often too trivial.

We illustrate the procedure with the following example:

*When asked about **Adams'** report, **Powell** found many of the statements to be inaccurate, including a claim that **Adams** first surveyed an area that was surveyed in 1857 by Joseph C.*

The second occurrence of "Adams" is masked. The goal is to determine which of the two candidates ("Adams","Powell") has been masked out. The masking process resembles replacing a name with a pronoun, but the pronoun is not inserted to keep the process fully unsupervised and error-free.

We used the Spacy Named Entity Recognition library[3] to find the occurrences of names in the text. The resulting dataset consists of $2,438,897$ samples. $10,000$ examples are held out to serve as the validation set. Two examples from our dataset can be found on Figure 1.

Gina arrives and she is furious with Denise for not protecting Jody from Kingsley, as `[MASK]` *was meant to be the parent.*
Candidates: Gina, **Denise**

When Ashley falls pregnant with Victor's child, Nikki is diagnosed with cancer, causing Victor to leave `[MASK]`*, who secretly has an abortion.*
Candidates: Ashley, Nikki

Figure 1: WIKICREM examples. Correct answers are given in bold.

We note that our dataset contains hard examples. To resolve the first example, one needs to understand that Denise was assigned a task and "meant to be the parent" thus refers to her. To resolve the second example, one needs to understand that having an abortion can only happen if one falls pregnant first. Since both candidates have feminine names, the answer cannot be deduced just on the common co-occurrence of female names and the word "abortion".

We highlight that our example generation method, while having the advantage of being unsupervised, also does not give incorrect signals, since we know the ground truth reference.

Even though WIKICREM and GAP both use text from English Wikipedia, they produce differing examples, because their generating processes differ. In GAP, passages with pronouns are collected and the pronouns are manually annotated, while WIKICREM is generated by masking names that appear in the text. Even if the same text is used, different masking process will result in different inputs and outputs, making the examples different under the transductive hypothesis.

WIKICREM statistics. We analyze our dataset for gender bias. We use the Gender guesser library[4] to determine the gender of the candidates. To mimic the analysis of pronoun genders performed in the related works (Webster et al., 2018; Rudinger et al., 2018; Zhao et al., 2018), we observe the gender of the correct candidates only. There were 0.8M "male" or "mostly_male" names and 0.42M "female" or "mostly_female" names, the rest were classified as "unknown". The ratio between female and male candidates is thus estimated around 0.53 in favour of male candidates. We will see that this gender imbalance does not have any negative impact on bias, as shown in Section 6.2.

However, our unsupervised generating procedure sometimes yields examples where the correct answer cannot be deduced given the available information, we refer to these as *unsolvable examples*. To estimate the percentage of unsolvable examples, we manually annotated 100 randomly selected examples from the WIKICREM dataset. In order to prevent guessing, the candidates were not visible to the annotators. For each example, we asked them to state whether it was solvable or not, and to answer the solvable examples. In 100 examples, we found 18 unsolvable examples and achieved 95.1% accuracy on the rest, showing that the annotation error rate is tolerable. These annotations can be found in Appendix A.

However, as shown in Section 6.2, training on WIKICREM alone does not match the performance of training on the data from the target distribution. The data distribution of WIKICREM differs from the data distribution of the datasets for evaluation. If we replace the `[MASK]` token with a pronoun instead of the correct candidate, the resulting sentence sometimes sounds unnatural and would not occur in a human-written text. On the annotated 100 examples, we estimated the percentage of natural-sounding sentences to be 63%. While the these sentences are not incorrect, the

[3] https://spacy.io/usage/linguistic-features#named-entities

[4] https://pypi.org/project/gender-guesser/

distribution of the training data differ from the target data.

4 Model

We use a simple language-model-based approach to anaphora resolution to show the value of the introduced dataset. In this section, we first introduce BERT (Devlin et al., 2018), a language model that we use throughout this work. In the second part, we describe the utilization of BERT and the fine-tuning procedures employed.

4.1 BERT

The Bidirectional Encoder Representations from Transformers (BERT) language model is based on the transformer architecture (Vaswani et al., 2017). We choose this model due to its strong language-modeling abilities and high performance on several NLU tasks (Devlin et al., 2018).

BERT is initially trained on two tasks: next sentence prediction and masked token prediction. In the next sentence prediction task, the model is given two sentences and is asked to predict whether the second sentence follows the first one. In the masked token prediction task, the model is given text with approximately 15% of the input tokens masked, and it is asked to predict these tokens. The details of the pre-training procedure can be found in Devlin et al. (2018).

In this work, we only focus on the masked token prediction. We use the PyTorch implementation of BERT[5] and the pre-trained weights for BERT-large released by Devlin et al. (2018).

4.2 Pronoun Resolution with BERT

This section introduces the procedure for pronoun resolution used throughout this work. Let S be the sentence with a pronoun that has to be resolved. Let $\mathbf{a}$ be a candidate for pronoun resolution. The pronoun in S is replaced with a `[MASK]` token and used as the input to the model to compute the log-probability $\log \mathbb{P}(\mathbf{a}|S)$. If $\mathbf{a}$ consists of more than one token, the same number of `[MASK]` tokens is inserted into S, and the log-probability $\log \mathbb{P}(\mathbf{a}|S)$ is computed as the average of log-probabilities of all tokens in $\mathbf{a}$.

The candidate-finding procedures are dataset-specific and are described in Section 6. Given a sentence S and several candidates $\mathbf{a}_1, \ldots, \mathbf{a}_n$, we select the candidate $\mathbf{a}_i$ with the largest $\log \mathbb{P}(\mathbf{a}_i|S)$.

[5] https://github.com/huggingface/pytorch-pretrained-BERT

4.3 Training

When training the model, the setup is similar to testing. We are given a sentence with a name or a pronoun masked out, together with two candidates. The goal is to determine which of the candidates is a better fit. Let $\mathbf{a}$ be the correct candidate, and $\mathbf{b}$ be an incorrect candidate. Following our previous work (Kocijan et al., 2019) we minimize the negative log-likelihood of the correct candidate, while additionally imposing a max-margin between the log-likelihood of the correct and incorrect terms. We observe that this combined loss consistently yields better results on validation sets of all experiments than negative log-likelihood or max-margin loss on their own.

$$\mathcal{L} = -\log \mathbb{P}(\mathbf{a}|S) + \\ +\alpha \cdot \max(0, \log \mathbb{P}(\mathbf{b}|S) - \log \mathbb{P}(\mathbf{a}|S) + \beta), \quad (1)$$

where α and β are hyperparameters controlling the influence of the max-margin loss term and the margin between the log-likelihood of the correct and incorrect candidates, respectively.

The hyperparameter settings for fine-tuning BERT on WIKICREM were the same as by Devlin et al. (2018), except for the learning rate and introduced constants α and β. For our hyperparameter search, we used learning rate $lr \in \{3 \cdot 10^{-5}, 1 \cdot 10^{-5}, 5 \cdot 10^{-6}, 3 \cdot 10^{-6}\}$ and hyperparameters $\alpha \in \{5, 10, 20\}$, $\beta \in \{0.1, 0.2, 0.4\}$ with grid search. The hyperparameter search is performed on a subset of WIKICREM with 10^5 datapoints to reduce the searching time. We compare the influence of hyperparameters on the validation set of WIKICREM dataset. The best validation score was achieved with $lr = 1 \cdot 10^{-5}$, $\alpha = 10$, and $\beta = 0.2$. We used batches of size 64.

Since WIKICREM is large and one epoch takes around two days even when parallelized on 8 Tesla P100 GPUs, we only fine-tune BERT on WIKICREM for a single epoch. We note that better results may be achieved with further fine-tuning and improved hyperparameter search.

Fine-tuning on other datasets is performed in the same way as training except for two differences. Firstly, in fine-tuning, the model is trained for 30 epochs due to the smaller size of datasets. Secondly, we do not sub-sample the training set for hyperparameter search. We validate the model

after every epoch, retaining the model that performs best on the WIKICREM validation set.

5 Evaluation Datasets

We now introduce the 7 datasets that were used to evaluate the models. We decide not to use the CONLL2012 and WINOCOREF (Pradhan et al., 2012; Peng et al., 2015) datasets, because they contain more general coreference examples than just pronouns. We did not evaluate on the KNOWREF dataset (Emami et al., 2019), since it was not yet publicly available at the time of writing.

GAP. GAP (Webster et al., 2018) is a collection of $4,454$ passages from Wikipedia containing ambiguous pronouns. It focuses on the resolution of personal pronouns referring to human names and has a $1:1$ ratio between masculine and feminine pronouns. In addition to the overall performance on the dataset, each model is evaluated also on its performance on the masculine subset (F_1^M), feminine subset (F_1^F), and its gender bias ($\frac{F_1^F}{F_1^M}$). The best performance was exhibited by the Referential Reader (Liu et al., 2019), a GRU-based model with additional external memory cells.

For each example, two candidates are given with the goal of determining whether they are the referent. In approximately 10% of the training examples, none of the candidates are correct. When training on the GAP dataset, we discard such examples from the training set. We do not discard any examples from the validation or test set.

When testing the model, we use the Spacy NER library to find all candidates in the sentence. Since the GAP dataset mainly contains examples with human names, we only retain named entities with the tag PERSON. We observe that in 18.5% of the test samples, the Spacy NER library fails to extract the candidate in question, making the answer for that candidate "FALSE" by default, putting our models at disadvantage. Because of this, 7.25% of answers are always false negatives, and 11.25% are always true negatives, regardless of the model. Taking this into account, we compute that the maximal F_1-score achievable by our models is capped at 91.1%.

We highlight that, when evaluating our models, we are stricter than previous approaches (Liu et al., 2019; Webster et al., 2018). While they count the answer as "correct" if the model returns a substring of the correct answer, we only accept the full answer. The aforementioned models return the exact location of the correct candidate in the input sentence, while our approach does not. This strictness is necessary, because a substring of a correct answer could be a substring of several answers at once, making it ambiguous.

WSC. The Winograd Schema Challenge (Levesque et al., 2011) is a hard pronoun resolution challenge inspired by the example from Winograd (1972):

The city councilmen refused the demonstrators a permit because ***they*** *[feared/advocated] violence.*
Question: Who [feared/advocated] violence?
Answer: the city councilmen / the demonstrators

A change of a single word in the sentence changes the referent of the pronoun, making it very hard to resolve. An example of a Winograd Schema must meet the following criteria (Levesque et al., 2011):

1. Two entities appear in the text.
2. A pronoun or a possessive adjective appears in the sentence and refers to one of the entities. It would be grammatically correct if it referred to the other entity.
3. The goal is to find the referent of the pronoun or possessive adjective.
4. The text contains a "special word". When switched for the "alternative word", the sentence remains grammatically correct, but the referent of the pronoun changes.

The Winograd Schema Challenge is specifically made up from challenging examples that require commonsense reasoning for resolution and should not be solvable with statistical analysis of co-occurence and association.

We evaluate the models on the collection of 273 problems used for the 2016 Winograd Schema Challenge (Davis et al., 2017), also known as WSC273. The best known approach to this problem uses the BERT language model, fine-tuned on the DPR dataset (Kocijan et al., 2019).

DPR. The Definite Pronoun Resolution (DPR) corpus (Rahman and Ng, 2012) is a collection of problems that resemble the Winograd Schema Challenge. The criteria for this dataset have been relaxed, and it contains examples that might not require commonsense reasoning or examples where the "special word" is actually a whole phrase. We remove 6 examples in the DPR training set that overlap with the WSC dataset. The

dataset was constructed manually and consists of 1316 training and 564 test samples after we removed the overlapping examples. The best result on the dataset was reported by Peng et al. (2015) using external knowledge sources and integer linear programming.

PDP. The Pronoun Disambiguation Problem (PDP) is a small collection of 60 problems that was used as the first round of the Winograd Schema Challenge in 2016 (Davis et al., 2017). Unlike WSC, the examples do not contain a "special word", however, they do require commonsense reasoning to be answered. The examples were manually collected from books. Despite its small size, there have been several attempts at solving this challenge (Fähndrich et al., 2018; Trinh and Le, 2018), the best result being held by the Marker Passing algorithm (Fähndrich et al., 2018).

WNLI. The Winograd Natural Language Inference (WNLI) is an inference task inspired by the Winograd Schema Challenge and is one of the 9 tasks on the GLUE benchmark (Wang et al., 2019). WNLI examples are obtained by rephrasing Winograd Schemas. The Winograd Schema is given as the "premise". A "hypothesis" is constructed by repeating the part of the premise with the pronoun and replacing the pronoun with one of the candidates. The goal is to classify whether the hypothesis follows from the premise.

A WNLI example obtained by rephrasing one of the WSC examples looks like this:

Premise: *The city councilmen refused the demonstrators a permit because they feared violence.*
Hypothesis: *The demonstrators feared violence.*
Answer: true / **false**

The WNLI dataset is constructed manually. Since the WNLI training and validation sets overlap with WSC, we use the WNLI test set only. The test set of WNLI comes from a separate source and does not overlap with any other dataset.

The currently best approach transforms examples back into the Winograd Schemas and solves them as a coreference problem (Kocijan et al., 2019). Following our previous work (Kocijan et al., 2019), we reverse the process of example generation in the same way. We automatically detect which part of the premise has been copied to construct the hypothesis. This locates the pronoun that has to be resolved, and the candidate in question. All other nouns in the premise are treated as alternative candidates. We find nouns in the premise with the Stanford POS tagger (Manning et al., 2014).

WINOGENDER. WINOGENDER (Rudinger et al., 2018) is a dataset that follows the WSC format and is aimed to measure gender bias. One of the candidates is always an occupation, while the other is a participant, both selected to be gender neutral. Examples intentionally contain occupations with strong imbalance in the gender ratio. Participant can be replaced with the neutral "someone", and three different pronouns (he/she/they) can be used. The aim of this dataset is to measure how the change of the pronoun gender affects the accuracy of the model.

Our models mask the pronoun and are thus not affected by the pronoun gender. They exhibit no bias on this dataset by design. We mainly use this dataset to measure the accuracy of different models on the entire dataset. According to Rudinger et al. (2018), the best performance is exhibited by Durrett and Klein (2013) when used on the male subset of the dataset. We use this result as the baseline.

WINOBIAS. Similarly to the WINOGENDER dataset, WINOBIAS (Zhao et al., 2018) is a WSC-inspired dataset that measures gender bias in the coreference resolution algorithms. Similarly to WINOGENDER, it contains instances of occupations with high gender imbalance. It contains 3,160 examples of Winograd Schemas, equally split into validation and test set. The test set examples are split into 2 types, where examples of type 1 are "harder" and should not be solvable using the analysis of co-occurrence, and examples of type 2 are easier. Additionally, each of these subsets is split into anti-stereotypical and pro-stereotypical subsets, depending on whether the gender of the pronoun matches the most common gender in the occupation. The difference in performance between pro- and anti-stereotypical examples shows how biased the model is. The best performance is exhibited by Lee et al. (2017) and Durrett and Klein (2013), as reported by Zhao et al. (2018).

6 Evaluation

We quantify the impact of WIKICREM on the introduced datasets.

6.1 Experiments

We train several different models to evaluate the contribution of the WIKICREM dataset in different real-world scenarios. In **Scenario A**, no information of the target distribution is available. In **Scenario B**, the distribution of the target data is known and a sample of training data from the target distribution is available. Finally, **Scenario C** is the transductive scenario where the unlabeled test samples are known in advance. All evaluations on the GAP test-set are considered to be Scenario C, because BERT has been pre-trained on the English Wikipedia and has thus seen the text in the GAP dataset at the pre-training time.

We describe the evaluated models below.

BERT. This model, pretrained by Devlin et al. (2018), is the starting point for all models and serves as the soft baseline for Scenario A.

BERT_WIKIRAND. This model serves as an additional baseline for Scenario A and aims to eliminate external factors that might have worked against the performance of BERT. To eliminate the effect of sentence lengths, loss function, and the percentage of masked tokens during the training time, we generate the RANDOMWIKI dataset. It consists of random passages from Wikipedia and has the same sentence-length distribution and number of datapoints as WIKICREM. However, the masked-out word from the sentence is selected randomly, while the alternative candidate is selected randomly from the vocabulary. BERT is then trained on this dataset in the same way as BERT_WIKICREM, as described in Section 4.3.

BERT_WIKICREM. BERT, additionally trained on WIKICREM. Its evaluation on non-GAP datasets serves as the evaluation of WIKICREM under Scenario A.

BERT_DPR. BERT, fine-tuned on DPR. We hold out 10% of the DPR train set (131 examples) to use them as the validation set. All datasets, other than GAP, were inspired by the Winograd Schema Challenge and come from a similar distribution. We use this model as the baseline for Scenario B.

BERT_WIKICREM_DPR. This model is obtained by fine-tuning BERT_WIKICREM on DPR using the same split as for BERT_DPR. It serves as the evaluation of WIKICREM under Scenario B.

BERT_GAP_DPR. This model serves as an additional comparison to the BERT_WIKICREM_DPR model. It is obtained by fine-tuning BERT_GAP on the DPR dataset.

BERT_GAP. This model is obtained by fine-tuning BERT on the GAP dataset. It serves as the baseline for Scenario C, as explained at the beginning of Section 6.1.

BERT_WIKICREM_GAP. This model serves as the evaluation of WIKICREM for Scenario C and is obtained by fine-tuning BERT_WIKICREM on GAP.

BERT_ALL. This model is obtained by fine-tuning BERT on all the available data from the target datasets at once. Combined GAP-train and DPR-train data are used for training. The model is validated on the GAP-validation set and the WINOBIAS-validation set separately. Scores on both sets are then averaged to obtain the validation performance. Since both training sets and both validation sets have roughly the same size, both tasks are represented equally.

BERT_WIKICREM_ALL. This model is obtained in the same way as the BERT_ALL model, but starting from BERT_WIKICREM instead.

6.2 Results

The results of the evaluation of the models on the test sets are shown in Table 1. We notice that additional training on WIKICREM consistently improves the performance of the models in all scenarios and on most tests. Due to the small size of some test sets, some of the results are subject to deviation. This especially applies to PDP (60 test samples) and WNLI (145 test samples).

We observe that BERT_WIKIRAND generally performs worse than BERT, with GAP and PDP being notable exceptions. This shows that BERT is a strong baseline and that improved performance of BERT_WIKICREM is not a consequence of training on shorter sentences or with different loss function. BERT_WIKICREM consistently outperforms both baselines on all tests, showing that WIKICREM can be used as a standalone dataset.

We observe that training on the data from the target distribution improves the performance the most. Models trained on GAP-train usually show more than a 20% increase in their F_1-score on GAP-test. Still, BERT_WIKICREM_GAP shows

	Transductive scenario							
	GAP F_1	GAP F_1^F	GAP F_1^M	Bias $\frac{F_1^F}{F_1^M}$	DPR	WSC	WNLI	
SOTA	72.1%	71.4%	72.8%	0.98	76.4%	**72.5%**	**74.7%**	
BERT	50.0%	47.2%	52.7%	0.90	59.8%	61.9%	65.8%	no
BERT_WIKIRAND	55.1%	51.8%	58.2%	0.89	59.2%	59.3%	65.8%	train
BERT_WIKICREM	**59.0%**	**57.5%**	**60.5%**	**0.95**	**67.4%**	**63.4%**	**67.1%**	data
BERT_GAP	75.2%	75.1%	75.3%	**1.00**	66.8%	63.0%	68.5%	
BERT_WIKICREM_GAP	**77.4%**	**78.4%**	**76.4%**	1.03	**71.1%**	**64.1%**	**70.5%**	
BERT_DPR	60.9%	61.3%	60.6%	1.01	**83.3%**	67.0%	71.9%	existing
BERT_GAP_DPR	**70.0%**	**70.4%**	**69.5%**	1.01	79.4%	65.6%	72.6%	train
BERT_WIKICREM_DPR	64.2%	64.2%	64.1%	**1.00**	80.0%	**71.8%**	**74.7%**	data
BERT_ALL	76.0%	77.4%	74.7%	**1.04**	80.1%	**70.0%**	74.0%	
BERT_WIKICREM_ALL	**78.0%**	**79.4%**	**76.7%**	**1.04**	**84.8%**	**70.0%**	**74.7%**	

	WB T1-a	WB T1-p	WB T2-a	WB T2-p	WINOGENDER	PDP	
SOTA	60.6%	74.9%	78.0%	88.6%	50.9%	74.0%	
BERT	61.3%	60.3%	76.2%	75.8%	59.2%	71.7%	no
BERT_WIKIRAND	53.5%	52.5%	64.6%	65.2%	57.9%	73.3%	train
BERT_WIKICREM	**65.2%**	**64.9%**	**95.7%**	**94.9%**	**66.7%**	**76.7%**	data
BERT_GAP	64.6%	63.8%	88.1%	87.9%	67.5%	**85.0%**	
BERT_WIKICREM_GAP	**71.2%**	**70.5%**	**97.2%**	**98.2%**	**75.4%**	83.3%	
BERT_DPR	**78.0%**	**78.2%**	85.6%	86.4%	79.2%	81.7%	existing
BERT_GAP_DPR	77.8%	76.5%	**89.6%**	**89.1%**	75.8%	**86.7%**	train
BERT_WIKICREM_DPR	76.0%	76.3%	81.3%	80.3%	**82.1%**	76.7%	data
BERT_ALL	**77.8%**	**77.2%**	94.7%	94.9%	**78.8%**	81.7%	
BERT_WIKICREM_ALL	76.8%	75.8%	**98.7%**	**99.0%**	76.7%	**86.7%**	

Table 1: Evaluation of trained models on all test sets. GAP and WINOBIAS (abbreviated WB) are additionally split into subsets, as introduced in Section 5. Double lines in the table separate results from three different scenarios: when no training data is available, when additional training data exists, and the transductive scenario. The table is further split into sections separated with single horizontal lines. Each section contains a model that has been trained on WIKICREM and models that have not been. The best result in each section is in bold. The best overall result is underlined. Scores on GAP are measured as F_1-scores, while the performance on other datasets is given in accuracy. The source of each SOTA is listed in Section 5.

a consistent improvement over BERT_GAP on all subsets of the GAP test set. This confirms that WIKICREM works not just as a standalone dataset, but also as an additional pre-training in the transductive scenario.

Similarly, BERT_WIKICREM_DPR outperforms BERT_DPR on the majority of tasks, showing the applicability of WIKICREM to the scenario where additional training data is available. However, good results of BERT_GAP_DPR show that additional training on a manually constructed dataset, such as GAP, can yield similar results as additional training on WIKICREM. The reason behind this difference is the impact of the data distribution. GAP, DPR, and WIKICREM contain data that follows different distributions which strongly impacts the trained models. This can be seen when we fine-tune BERT_GAP on DPR to obtain BERT_GAP_DPR, as the model's performance on GAP-test drops by 8.2%. WIKICREM's data distribution strongly differs from the test sets' as described in Section 3.

However, the best results are achieved when all available data is combined, as shown by the models BERT_ALL and BERT_WIKICREM_ALL. BERT_WIKICREM_ALL achieves the highest performance on GAP, DPR, WNLI, and WINOBIAS among the models, and sets the new state-of-the-art result on GAP, DPR, and WINOBIAS. The new state-of-the-art result on the WINOGENDER dataset is achieved by the BERT_WIKICREM_DPR model, while BERT_WIKICREM_ALL and BERT_GAP_DPR set the new state-of-the-art on the PDP dataset.

7 Conclusions and Future Work

In this work, we introduced WIKICREM, a large dataset of training instances for pronoun resolution. We use our dataset to fine-tune the BERT language model. Our results match or outperform state-of-the-art models on 6 out of 7 evaluated datasets.

The employed data-generating procedure can be further applied to other large sources of text to generate more training sets for pronoun resolution. In addition, both variety and size of the generated datasets can be increased if we do not restrict ourselves to personal names. We hope that the community will make use of our released WIKICREM dataset to further improve the pronoun resolution task.

Acknowledgments

This work was supported by the Alan Turing Institute under the UK EPSRC grant EP/N510129/1, by the EPSRC grant EP/R013667/1, by the EPSRC studentship OUCS/EPSRC-NPIF/VK/1123106, by the JP Morgan PhD Fellowship 2019-2020, and by an EPSRC Vacation Bursary. We also acknowledge the use of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1).

References

Ernest Davis, Leora Morgenstern, and Charles L. Ortiz. 2017. The first Winograd Schema Challenge at IJCAI-16. *AI Magazine*, 38(3):97–98.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. *Computing Research Repository*, arXiv:1810.01375.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KNOWREF Coreference Corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *ACL 2019*.

Johannes Fähndrich, Sabine Weber, and Hannes Kanthak. 2018. A marker passing approach to Winograd schemas. In *Semantic Technology*, pages 165–181, Cham. Springer International Publishing.

Abbas Ghaddar and Philippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for Winograd Schema Challenge. *Computing Research Repository*, arXiv:1905.06290.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd Schema Challenge. *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 46.

Fei Liu, Luke S. Zettlemoyer, and Jacob Eisenstein. 2019. The referential reader: A recurrent entity network for anaphora resolution. *Computing Research Repository*, abs/1902.01541.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Kotaro Nakayama. 2019. Wikipedia mining for triple extraction enhanced by co-reference resolution.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *HLT-NAACL*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd Schema Challenge. In *Proceedings of EMNLP*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Computing Research Repository*, abs/1804.09301.

Ulrich Schäfer, Christian Spurk, and Jörg Steffen. 2012. A fully coreference-annotated corpus of scholarly papers from the ACL Anthology. *Proceedings of COLING 2012: Posters*, pages 1059–1070.

T. H. Trinh and Q. V. Le. 2018. A Simple Method for Commonsense Reasoning. *Computing Research Repository*, arXiv:1806.02847.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, arXiv:1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the ACL*.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., Orlando, FL, USA.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.