The KNOWREF Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution

Ali Emami*[1], Paul Trichelair*[1], Adam Trischler[2], Kaheer Suleman[2],
Hannes Schulz[2], and Jackie Chi Kit Cheung[1]

[1]School of Computer Science, Mila/McGill University
[2]Microsoft Research Montreal
{*ali.emami, paul.trichelair*}*@mail.mcgill.ca*
{*adam.trischler, kasulema, hannes.schulz*}*@microsoft.com*
jcheung@cs.mcgill.ca

Abstract

We introduce a new benchmark for coreference resolution and NLI, KNOWREF, that targets common-sense understanding and world knowledge. Previous coreference resolution tasks can largely be solved by exploiting the number and gender of the antecedents, or have been handcrafted and do not reflect the diversity of naturally occurring text. We present a corpus of over 8,000 annotated text passages with ambiguous pronominal anaphora. These instances are both challenging and realistic. We show that various coreference systems, whether rule-based, feature-rich, or neural, perform significantly worse on the task than humans, who display high inter-annotator agreement. To explain this performance gap, we show empirically that state-of-the art models often fail to capture context, instead relying on the gender or number of candidate antecedents to make a decision. We then use problem-specific insights to propose a data-augmentation trick called *antecedent switching* to alleviate this tendency in models. Finally, we show that antecedent switching yields promising results on other tasks as well: we use it to achieve state-of-the-art results on the GAP coreference task.

1 Introduction

Coreference resolution is one of the best known tasks in Natural Language Processing (NLP). Despite a large body of work in the area over the last few decades (Morton, 2000; Bean and Riloff, 2004; McCallum and Wellner, 2005; Rahman and Ng, 2009), the task remains challenging. Many resolution decisions require extensive world knowledge and understanding common points of reference (Pradhan et al., 2011). In the case of pronominal anaphora resolution, these forms of "common sense" become much more important when cues like gender and number do not by themselves indicate the correct resolution (Trichelair et al., 2018).

To date, most existing methods for coreference resolution (Raghunathan et al., 2010; Lee et al., 2011; Durrett et al., 2013; Lee et al., 2017, 2018) have been evaluated on a few popular datasets, including the CoNLL 2011 and 2012 shared coreference resolution tasks (Pradhan et al., 2011, 2012). These datasets were proposed as the first comprehensively tagged and large-scale corpora for coreference resolution, to spur progress in state-of-the-art techniques. According to Durrett and Klein (2013), this progress would contribute in the "uphill battle" of modelling not just syntax and discourse, but also semantic compatibility based on world knowledge and context.

Despite improvements in benchmark dataset performance, the question of what exactly current systems learn or exploit remains open, particularly with recent neural coreference resolution models. Lee et al. (2017) note that their model does "little in the uphill battle of making coreference decisions that require world knowledge," and highlight a few examples in the CoNLL 2012 task that rely on more complex understanding or inference. Because these cases are infrequent in the data, systems can perform very well on the CoNLL tasks according to standard metrics by exploiting surface cues. High-performing models have also been observed to rely on social stereotypes present in the data, which could unfairly impact their decisions for some demographics (Zhao et al., 2018).

There is a recent trend, therefore, to develop more challenging and diverse coreference tasks. Perhaps the most popular of these is the Winograd Schema Challenge (WSC), which has emerged as an alternative to the Turing test (Levesque et al., 2011). The WSC task is carefully controlled such that heuristics involving syntactic salience, the number and gender of the antecedents, or other

*equal contribution

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3952–3961
Florence, Italy, July 28 - August 2, 2019.

obvious syntactic/semantic cues are ineffective. Previous approaches to common sense reasoning, based on logical formalisms (Bailey et al., 2015) or deep neural models (Liu et al., 2016), have solved only restricted subsets of the WSC with high precision. These shortcomings can in part be attributed to the limited size of the corpus (273 instances), which is a side effect of its hand-crafted nature. Webster et al. (2018) recently presented a corpus called GAP that consists of about 4,000 unique binary coreference instances from English Wikipedia. This corpus is intended to address gender bias and the mentioned size limitations of the WSC. We believe that gender bias in coreference resolution is part and parcel of a more general problem: current models are unable to abstract away from the entities in the sentence to take advantage of the wider context to make a coreference decision.

To tackle this issue, we present a coreference resolution corpus called KNOWREF that specifically targets the ability of systems to reason about a situation described in the context.[1] We designed this task to be challenging, large-scale, and based on natural text. The main contributions of this paper are as follows:

1. We develop mechanisms by which we construct a human-labeled corpus of 8,724 Winograd-like text samples whose resolution requires significant common sense and background knowledge. As an example:

   Marcus is undoubtedly faster than Jarrett right now but in [his] prime the gap wasn't all that big. (answer: Jarrett)

2. We propose a task-specific metric called *consistency* that measures the extent to which a model uses the full context (as opposed to a surface cue) to make a coreference decision. We use this metric to analyze the behavior of state-of-the-art methods and demonstrate that they generally under-utilize context information.

3. We find that a fine-tuned version of the recent large-scale language model, BERT (Devlin et al., 2018), performs significantly better than other methods on KNOWREF, although with substantial room for improvement to match human performance.

4. We demonstrate the benefits of a data-augmentation technique called *antecedent switching* in expanding our corpus, further deterring models from exploiting surface cues, as well as in transferring to models trained on other co-reference tasks like GAP, leading to state-of-the-art results.

[1]The corpus, the code to scrape the sentences from the source texts, as well as the code to reproduce all of our experimental results are available at https://github.com/aemami1/KnowRef.

2 Related Work

2.1 General coreference resolution

Automated techniques for standard coreference resolution — that is, the task of correctly partitioning the entities and events that occur in a document into resolution classes — date back to decision trees and hand-written rules (Hobbs, 1977; McCarthy, 1995). The earliest evaluation corpora were the Message Understanding Conferences (MUC) (Grishman and Sundheim, 1996) and the ACE (Doddington et al., 2004). These focused on noun phrases tagged with coreference information, but were limited in either size or annotation coverage.

The datasets of Pradhan et al. (2011, 2012) from the CoNLL-2011 and CoNLL-2012 Shared Tasks were proposed as large-scale corpora with high inter-annotator agreement. They were constructed by restricting the data to coreference phenomena with highly consistent annotations, and were packaged with a standard evaluation framework to facilitate performance comparisons.

The quality of these tasks led to their widespread use and the emergence of many resolution systems, ranging from hand-engineered methods to deep-learning approaches. The multi-pass sieve system of Raghunathan et al. (2010) is fully deterministic and makes use of mention attributes like gender and number; it maintained the best results on the CoNLL 2011 task for a number of years (Lee et al., 2011). Later, lexical learning approaches emerged as the new state of the art (Durrett and Klein, 2013), followed more recently by neural models (Wiseman et al., 2016; Clark and Manning, 2016). The current state-of-the-art result on the CoNLL 2012 task is by an end-to-end neural model from Lee et al. (2018) that does not rely on a syntactic parser or a hand-engineered mention detector.

2.2 Gender bias in general coreference resolution

Zhao et al. (2018) observed that state-of-the-art

methods for coreference resolution become gender-biased, exploiting various stereotypes that leak from society into data. They devise a dataset of 3,160 manually written sentences called *WinoBias* that serves both as a gender-bias test for coreference resolution models and as a training set to counter stereotypes in existing corpora (i.e., the two CoNLL tasks). The following example is representative:

(1) The physician hired the secretary because he was overwhelmed with clients.

(2) The physician hired the secretary because she was overwhelmed with clients.

Experiments conducted on various models demonstrated that an end-to-end neural model (Lee et al., 2017) maintains its performance without the gender bias when trained partially on both the previous datasets and on *WinoBias*.

A concurrent work by Rudinger et al. (2018) also proposed an empirical study of the biases in coreference resolution systems. In contrast to Zhao et al. (2018), who attribute the bias in part to the datasets, they conjecture that the gender bias comes primarily from the models themselves. Based on statistics from the Bureau of Labor, they show that various systems all exhibit significant gender bias.

This work on gender stereotypes provides some insight into the behavior of current models. In the example above, if *she* is predicted incorrectly to refer to *the secretary*, it is likely because the model learned a representation for the secretary profession that encodes gender information. Current models do not capture the context nor the relation between *was overwhelmed* and *hired* that lead to the correct resolution. The subject of our work is to investigate the potential for models to capture contextual relationships instead of cues from, e.g., gender stereotypes. Unlike WinoBias, our task is composed of passages that occur naturally in text and it is several times larger.

2.3 Difficult cases in coreference resolution

As the creators of the CoNLL tasks note, most coreference techniques rely primarily on surface-level features, like the proximity between mentions, or shallow semantic features like number, gender, named entities, semantic class, etc., rather than knowledge and context.

To address this, Levesque et al. (2011) manually constructed a dataset of challenging pronoun disambiguation problems called the Winograd Schema Challenge. The goal was that any successful system would necessarily use common-sense knowledge. Although the WSC is an important step in evaluating systems en route to human-like language understanding, its size and other characteristics are a bottleneck for progress in pronoun disambiguation (Trichelair et al., 2018). A Winograd-like expanded corpus was proposed by Rahman and Ng (2012) to address the WSC's size limitations; however, systems that perform well on the expanded dataset do not transfer successfully to the original WSC (Rahman and Ng, 2012; Peng et al., 2015), likely due to loosened constraints in the former.

The task that we propose distinguishes itself from the WSC by building on sentences that occur in natural text. This yields highly diverse problem instances. It is particularly important that, as well as being challenging, tasks are representative of natural text, so that improvements are more likely to transfer to the full coreference setting.

Recently, Webster et al. (2018) presented a corpus called GAP that consists of 4,454[2] unique binary coreference instances from English Wikipedia. It is meant to address gender bias and the described size limitation of the WSC. For instance, it exposes the unbalanced performance of current state-of-the-art resolvers, which more accurately resolve masculine pronouns than feminine pronouns. As for the difficulty of the task, the models tested on GAP were not trained directly on the corpus, which does not give a clear picture of the task's difficulty. A simple heuristic called *Parallelism+URL*, which is based on using the syntactic distance between antecedents and the target pronoun, is so far the strongest GAP baseline, at above 70% accuracy. This suggests that GAP is vulnerable to exploits that circumvent a need for knowledge, albeit not the gender and number cues that coreference resolvers have exploited before. Finally, our corpus construction process differs from that of GAP's by more strictly requiring that the sentences are in WSC-format, that is, there are exactly two named entities that occur strictly before the pronoun and only one of which may co-refer with the pronoun (in GAP, the pronoun may occur between and before the named entities and may in fact co-refer with both named entities). In addition, our corpus construction process exploits the fact that the named entities can be replaced with any name in or-

[2] In GAP, one unique coreference instance corresponds to two pronoun-name pairs, for which they report 8,908 pairs.

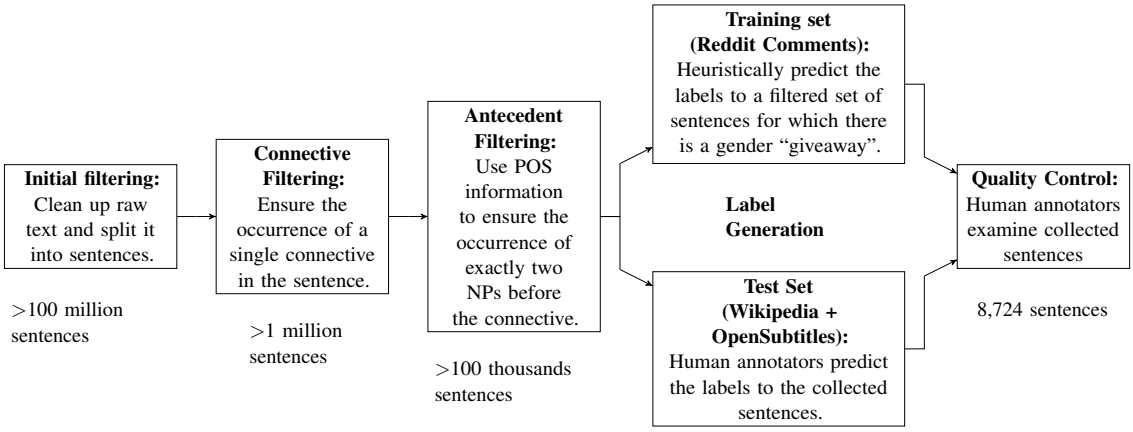

Figure 1: The corpus construction process for KNOWREF

der to increase the task difficulty by automatically removing gender giveaways as well as to significantly increase the size of the corpus by switching the named entities to create a new task instance.

As such, our paper seeks to explore a wider problem of which gender bias may be one facet: current models do not effectively abstract away from the entities (and instead rely on exploits using gender or plurality) to make the coreference resolution. By developing a benchmark task consisting strictly of sentences for which such cues are ineffective, we seek to challenge and potentially improve current coreference resolution models. In addition, based on our new benchmark, KNOWREF, we introduce a data-augmentation mechanism, called *antecedent switching*, to encourage models to perform this abstraction.

3 The KNOWREF Coreference Task

We develop a coreference task called KNOWREF that features 8,724 difficult pronoun disambiguation problems. Each instance is a short passage containing a target pronoun that must be correctly resolved to one of two possible antecedents.

Formally, each problem instance can be described as a tuple $P = \{S, C_1, C_2, T, K\}$, where S is the sentence, C_1 and C_2 are the candidate antecedents, T is the target pronoun to be resolved to one of C_1 and C_2, and K indicates the correct antecedent. Note that C_1, C_2, T and K appear in S. KNOWREF provides $\{S, C_1, C_2, T\}$ as input for models, which must predict K (e.g., as the output of a binary classification over C_1, C_2). A representative sentence S is the following.

(3) {Paul} helped {Lionel} hide when [he] was pursued by the authorities.

Here, C_1 = Paul, C_2 = Lionel, T = he, and $K = C_2$ = Lionel.

We control the text so as not to give away the pronoun's correct antecedent in surface-level cues involving syntactic salience or the number and gender of the antecedent. Successful systems must instead make use of the context, which may require world knowledge and common-sense inferences; i.e., that someone who is being helped to hide may be one who is being pursued by the authorities.

In the following section, we describe the methodology used to construct our corpus, provide a glimpse of a few of its instances and their resolution rationales, outline the task's evaluation criteria, and describe its characteristics.

3.1 Corpus construction

To construct KNOWREF, we scrape text samples from a large collection of documents: the combination of 2018 English Wikipedia, OpenSubtitles, and Reddit comments dating from 2006–2018. We filter this text through a multi-stage process to ensure quality and diversity as depicted in Figure 1, and described in more detail below.

3.1.1 Initial Filtering

After removing markup, non-ASCII characters, parenthetical expressions, headings and lists, we split the text into sentences. We keep sentences of token length between 9 and 33 words after naïve tokenization, which start with an upper case letter, and which contain no math.

3.1.2 Connective Filtering

Our first substantial filtering step uses regular expressions to ensure that each passed sentence con-

KNOWREF Example 1:	{Radu} appeared to be killed by {Brother Paulo}, but [he] reappears a short while later injured, but alive. (K = Radu)
Original sentence:	Radu appeared to be killed by Sister Paula, but he reappears a short while later injured, but alive.
KNOWREF Example 2:	{Wanda} tries to apologize to {Rose}, but [she] refuses to accept. (K = Rose)
Original sentence:	Warren tries to apologize to Rose, but she refuses to accept.
KNOWREF Example 3:	{Tom} arrives to where {Alex} was tied, but [he] has come free of his lead. (K = Alex)
Original sentence:	Tom arrives to where Vanessa was tied, but she has come free of her lead.

Table 1: Examples of KNOWREF instances.

tains connectives.[3] We use a regular expression to ensure that there is only one connective cluster (e.g. ", and though"), and that there are at least two non-stopwords before this connective and a pronoun after it. As a final check, we ensure that no pronoun occurs before the connective, which tends to remove sentences which are not self-contained.

3.1.3 Antecedent Filtering

On the remaining set of sentences, we use Stanford's Maxent tagger (Toutanova et al., 2003) to infer a flat part-of-speech (POS) labelling. Using the inferred POS tags, we ensure that there are exactly two noun phrases (NPs) before the connective that do not re-occur after it (a re-occurrence after the connective means that the pronoun likely refers to the non-repeated noun phrase).

The mentioned checks resulted in roughly 100,000 sentences across all three corpora. At least some of these remaining sentences have similar properties to Winograd schema sentences; that is, the two noun phrases (NPs) and the pronoun share the same type. From here, we keep only sentences where the type indicates that both NPs correspond to persons, which further filters the remaining sentences. We do this because NPs that denote people are often named entities or can easily be replaced by named entities without loss of information. We targeted these instances also because we investigate how resolution systems use gender cues and most gendered pronouns occur with person-type NPs.

3.1.4 Label Generation

We generate our training and test sets from distinct sources of text using two different methods.

Training set: We automatically collect 70,000 sentences from Reddit that have passed the filters described above, and filter these down to roughly 7,500 sentences for which the antecedents are named entities of different genders. We use a Python library[4] to infer the genders, based on a list of 40,000 names categorized as female or male compiled by Jörg Michael. Given the pronoun and the distinct predicted genders for the antecedents, we can infer the label for the pronoun's correct resolution with high accuracy and without the need for expensive human annotation. After assigning this label, we remove the gender giveaway by replacing one of the named entities so that both entities and the pronoun all match in gender (e.g., in a sentence with "James", "Jessica", and "she" as the NPs and pronoun, we replace "James" with "Jane"). These sentences form our training set. To assess its quality, we gave an annotator a random sample of 100 training instances with their heuristically determined labels. The annotator then evaluated each sentence as "correctly labelled", "incorrectly labelled", or "unresolvable" if neither of the two candidates were more suitable than the other to corefer with the pronoun.[5] In total, 86% of the instances were deemed to be labelled correctly, 11% incorrectly labelled, and 3% were not resolvable, implying that our automatic selection heuristic is strong but imperfect.

Test set: Human annotators examined all collected sentences for quality control. We also use a source for the test sentences that is distinct from that of the training set, directing our pipeline to collect sentences from Wikipedia and OpenSubtitles rather than Reddit. This is to ensure that stylistic cues common in the training source cannot be exploited by models at test time. In total, roughly 10,000 candidate sentences were extracted initially. As before, we automatically remove gender giveaways by replacing the named entities with names of the same gender, rendering the pronoun ambiguous. Then, six human annotators predicted which antecedent was the correct coreferent of the pronoun for a sample of 2,000 candidate sentences, or

[3] comma, semicolon, *or*, *since*, *but*, *because*, *although*, etc.

[4] https://pypi.org/project/SexMachine/

[5] The details and result of this quality-testing study will also be made public along with the code and dataset.

Sentence Characteristic	% of Data
Masculine target pronouns	52.7
Feminine target pronouns	47.3
First Antecedent Correct	50.7
Second Antecedent Correct	49.2

Table 2: Characteristics of the dataset, in terms of pronoun distribution and correct label.

they labeled the sentence with "neither" (in the case where neither antecedent feasibly corefers with the pronoun) or "unclear" (if the sentence was not intelligible). Sentences that have a strong agreement from 5 or more annotators on a single antecedent (and which are not labeled as "neither" or "unclear") are kept for testing. This yielded 1,269 test sentences. We measured high inter-annotator agreement on the test set with a Fleiss' Kappa of $\kappa = 0.78$.

Our pipeline thus yields a total of 8,724 sentences (7,455 training and 1,269 test) whose pronoun disambiguation should not be clear from shallow features like gender, number, and semantic type – they should instead require varying degrees of external knowledge. These sentences constitute the KNOWREF corpus. Examples of some instances are given in Table 1. As these examples reveal, each instance may require a unique bit of common sense knowledge to resolve.

In the first example, the common understanding that death (by way of killing) causes a disappearance helps us to conclude that Radu, the victim of murder, is the one to who reappears.

In the next example, human readers recognize that to *accept* is something one does with an *apology*. Therefore, *she* refers to the one that accepts the apology, i.e., Rose.

For the third example, an understanding that *being tied* is related to being deprived of freedom leads us to conclude that Alex has come free.

3.2 Task Characteristics

In Table 2, we report several statistical characteristics of the data. These suggest a near-equal distribution of feminine and masculine target pronouns (*he/him/his* vs. *she/her*) as well as an equal distribution of the two labels, which keeps chance-based performance at 50% expected accuracy.

3.3 Evaluation

Our task requires a model to choose between two candidates, but classical coreference models build clusters of expressions that refer to the same entity. With respect to our setting, several errors can be made by these existing models: predicting that the two entities and the pronoun share a similar cluster (*Both Antecedents Predicted*), that none of the two candidates shares a cluster with the pronoun (*No Decision*), or creating a cluster that contains the pronoun with the wrong candidate (*Incorrect Decision*). To obtain a score specific to our task, we compute a *Task-Specific Accuracy* which discards all of the cases in which the model makes no decision relevant to the target pronoun or chooses both entities as co-referring to the target pronoun.

4 Experiments and Results

In this section, we compare the performance of five representative coreference systems on our task: Stanford's rule-based system (Raghunathan et al., 2010) (**Rule**), Stanford's statistical system (Clark and Manning, 2015) (**Stat**), Clark and Manning (2016)'s deep reinforcement learning system (**Deep-RL**), Martschat and Strube (2015)'s latent tree model (**Latent**), and Lee et al. (2018)'s end-to-end neural system (**E2E**). We also report the accuracy of the state-of-the-art model, **E2E**, after retraining on KNOWREF and on KNOWREF+CoNLL.

Additionally, we develop a task-specific model for KNOWREF: a discriminatively trained fine-tuned instance of Bidirectional Encoder Representations from Transformers (**BERT**) (Devlin et al., 2018). We train our task-specific **BERT** according to recent work on language models (LMs) for the WSC (Trinh and Le, 2018). We first construct a modified version of the data wherein we duplicate each sentence, replacing the pronoun with one of the two antecedents in each copy. The task, akin to NLI, is then to predict which of the two modified sentences is most probable. To compute probabilities, we add a softmax layer with task-specific parameter vector $v \in \mathcal{R}^H$. Denote by $h_{S1} \in \mathcal{R}^H$ (respectively h_{S2}) the final hidden state for the sentence copy with the pronoun replaced by the first antecedent (respectively the second). Then the probability assigned to the first antecedent is

$$P_1 = \frac{e^{v^\top h_{S1}}}{e^{v^\top h_{S2}} + e^{v^\top h_{S2}}}. \quad (1)$$

Model	Both Antecedents Predicted	No Decision	Incorrect Decision	Correct Decision	Task-Specific Accuracy
Random	–	–	–	–	0.50
Human[5]	–	–	–	–	0.92
Rule	0.001	0.12	0.43	0.45	0.52
Stat	0.006	0.09	0.45	0.45	0.50
Deep-RL	0.001	0.09	0.46	0.45	0.49
Latent	0.000	0.12	0.41	0.47	0.54
E2E (CoNLL only)	0.01	0.42	0.23	0.35	0.60
E2E (KNOWREF)	0.000	0.26	0.31	0.43	0.58
E2E (KNOWREF+CoNLL)	0.000	0.19	**0.28**	0.52	**0.65**
BERT (KNOWREF)	0.000	0.000	0.39	**0.61**	0.61

Table 3: Coverage and performance of various representative systems on the KNOWREF Test set.

The probability assigned to the second antecedent is $P_2 = 1 - P_1$. We use $H = 768$ hidden units in our BERT implementation and learn v by minimizing the binary cross entropy with the ground-truth antecedent labels (in one-hot format).

Human Performance: We determined human performance on KNOWREF by collecting the predictions of six native English speakers on a randomly generated sub-sample of 100 problem instances; we consider correct those predictions that agreed with the majority decision and matched the ground-truth label derived from the original sentence. We report the performance of the five coreference systems and the human baseline in Table 3.

The human performance of 0.92 attests to the task's viability. The performance of the automatic systems pretrained on CoNLL, at random or slightly above random, demonstrates that state-of-the-art coreference resolution systems are unable to solve the task. This suggests the existence in the wild of difficult but realistic coreference problems that may be under-represented in CoNLL.

After training on KNOWREF, **E2E** improves by more than 5% in task-specific accuracy. We can infer from this result that the model can make some use of context to make predictions if trained appropriately, but that the CoNLL shared tasks may not contain enough of such instances for models to generalize from them. Finally our task-specific model reaches an accuracy of at best 65%, far below human performance despite having access to the two candidates.

4.1 Analysis by Switching Entities

Inspired by Trichelair et al. (2018), we propose to use a task-specific metric, *consistency*, to measure the ability of a model to use context in its coreference prediction, as opposed to relying on gender and number cues related to the entities. Accounting for this is critical, as we desire models that can capture social, situational, or physical awareness.

To measure consistency in the KNOWREF corpus, we duplicate the data set but switch the candidate antecedents each time they appear in a sentence. This changes the correct resolution. If a coreference model relies on knowledge and contextual understanding, its prediction should change as well, thus it could be called *consistent* in its decision process. If, however, its decision is influenced solely by the antecedent, its output would stay the same despite the change in context induced by switching. We define the *consistency* score as the percentage of predictions that change from the original instances to the switched instances. An example of a switching is:

(4) **Original**: {Alex} tells {Paulo}, but [he] does not believe him.
Switched: {Paulo} tells {Alex}, but [he] does not believe him.

The correct answer switches from K = Paulo to K = Alex.

[5]This is an estimate based on a subsample of the data.

Model	Consistency
Rule	0%
Stat	76%
Deep-RL	66%
Latent	78%
E2E	62%
E2E (KNOWREF)	66%
E2E (KNOWREF+CoNLL)	67%
BERT (KNOWREF)	69%

Table 4: The sensitivity of various systems to the instance antecedents, according to the number of changed decisions when the antecedents are switched. Higher is better.

Table 4 shows the consistency scores of the various baseline models evaluated on the original and switched duplicates of KNOWREF. The rule-based system (Raghunathan et al., 2010) always resolves to the same entity, suggesting that context is ignored. Indeed, the mechanisms underlying this model mostly rely on a gender and number dictionary (Bergsma and Lin, 2006). This dictionary informs a count-based approach that assigns a masculine, feminine, neutral, and plural score to each word. If the pronoun is *his*, the candidate with the higher masculine score is likely to be linked to the pronoun.

The other models, Stat, Deep-RL, E2E, Latent and BERT are much more robust to the switching procedure, demonstrating that the resolution partially relies on context cues. Regarding **E2E**, we can observe that training the model on KNOWREF forces the model to rely more on the context, leading to an improvement of 5%. It further demonstrates the usefulness of the corpus to obtain a better representation of the context.

4.2 Data Augmentation by Switching

Inspired by the switching experiment, we propose to extend the KNOWREF training set by switching every entity pair (thereby doubling the number of instances). We hypothesize that this data augmentation trick could force the model to abstract away from the entities to the context in order to boost performance, since it encounters the same contextual scenario in the doubled sentences.

Training on the augmented data, we observe an improvement of 10% for fine-tuned BERT (Table 5), yielding a task-specific accuracy of 71% on the KNOWREF test set. The improvement in accuracy is marginal for **E2E**, but we observe a large gain in consistency. We suspected that the data augmentation trick might also be useful in mitigating a model's gender bias, by encouraging the model to rely more on the context than on gendered entity names. To test this hypothesis, we train the same model with and without the data augmentation trick on the recently released GAP corpus (Webster et al., 2018).

Model	Accuracy	Δ	Consistency
BERT (KNOWREF)	71%	+10%	89%
E2E (KNOWREF)	61%	+3%	71%
E2E (KNOWREF+CoNLL)	66%	+1%	75%

Table 5: Accuracy on the KNOWREF test set for each model after augmenting the training set, as well as the difference from the result without data augmentation.

Model	$\frac{F_1^F}{F_1^M}$	F_1
Parallelism[6]	0.93	66.9
Parallelism+URL[6]	0.95	70.6
BERT (GAP)	1.02	69.2
BERT (GAP) + Data Aug.	**1.00**	**71.1**

Table 6: Performance on the GAP test set

BERT fine-tuned on GAP achieves a state of the art F_1 of 71.1 after data augmentation (Table 6). Not only does the augmentation improve the overall performance (+1.9) but it further balances the predictions' female:male ratio to 1:1.

4.3 Error Analysis

We show examples of BERT's performance (trained on KNOWREF) on our test set in Table 7. This includes instances on which it succeeds and fails for both original and switched sentences. In general, it is not clear why certain instances are more difficult for BERT to resolve, although training BERT on the augmented, switched corpus significantly reduces the frequency of inconsistent resolutions (from 31% to 11%).

These examples illustrate how challenging certain real-world situations can be for models to un-

[6]Scores reported in the original paper (Webster et al., 2018)

Sentence Type	Sentence	Answer
Original	Kara is in love with Tanya but she is too shy to tell [her].	Tanya ✓
Switched	Tanya is in love with Kara but she is too shy to tell [her].	Kara ✓ (consistently correct)
Original	Peter had not realised how old Henry was until [he] sees his daughter.	Henry ✗
Switched	Henry had not realised how old Peter was until [he] sees his daughter.	Peter ✗ (consistently incorrect)
Original	Poulidor was no match for Merckx, although [he] offered much resistance .	Poulidor ✓
Switched	Merckx was no match for Poulidor, although [he] offered much resistance .	Poulidor ✗ (inconsistent)

Table 7: Examples of various success/failure cases of BERT on the KNOWREF test set

derstand, compared to humans who can reason about them with ease.

5 Conclusion

We present a new corpus and task, KNOWREF, for coreference resolution. Our corpus contains difficult problem instances that require a significant degree of common sense and world knowledge for accurate coreference link prediction, and is larger than previous similar datasets. Using a task-specific metric, consistency, we demonstrate that training coreference models on KNOWREF improves their ability to build better representations of the context. We also show that progress in this capability is linked to reducing gender bias, with our proposed model setting the state of the art on GAP.

In the future, we wish to study the use of KNOWREF to improve performance on general coreference resolution tasks (e.g., the CoNLL 2012 Shared Tasks). We also plan to develop new models on KNOWREF and transfer them to difficult common sense reasoning tasks.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by Microsoft Research. Jackie Chi Kit Cheung is supported by the Canada CIFAR AI Chair program.

References

Dan Bailey, Amelia Harrison, Yuliya Lierler, Vladimir Lifschitz, and Julian Michael. 2015. The winograd schema challenge and reasoning about correlation. In *Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

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

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*.

Greg Durrett, David Hall, and Dan Klein. 2013. Decentralized entity-level modeling for coreference resolution. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

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*.

Ralph Grishman and Beth Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

Jerry R Hobbs. 1977. Pronoun resolution. *ACM SIGART Bulletin*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the fifteenth conference on computational natural language learning: Shared task*.

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*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association of Computational Linguistics*.

Andrew McCallum and Ben Wellner. 2005. Conditional models of identity uncertainty with application to noun coreference. In *Advances in neural information processing systems*.

JF McCarthy. 1995. Using decision trees for coreference resolution. In *Proc. 14th International Joint Conf. on Artificial Intelligence*.

Thomas S Morton. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. *Urbana*.

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*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*.

Paul Trichelair, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman, and Fernando Diaz. 2018. On the evaluation of common-sense reasoning in natural language understanding. *The NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of NAACL-HLT*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL-HLT*.