

The Hard-CoRe Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution

Ali Emami^{*1}, Paul Trichelair^{*1}, Adam Trischler², Kaheer Suleman², Hannes Schulz², and Jackie Chi Kit Cheung¹

¹School of Computer Science, Mila/McGill University

²Microsoft 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 task for coreference resolution, Hard-CoRe, that targets common-sense reasoning and world knowledge. Previous coreference resolution tasks have been overly vulnerable to systems that simply exploit the number and gender of the antecedents, or have been hand-crafted and do not reflect the diversity of sentences in naturally occurring text. With these limitations in mind, we present a resolution task that is both challenging and realistic. We demonstrate that various coreference systems, whether rule-based, feature-rich, graphical, or neural-based, perform at random or slightly above-random on the task, whereas human performance is very strong with high inter-annotator agreement. To explain this performance gap, we show empirically that state-of-the-art models often fail to capture context and rely only on the antecedents to make a decision.

Introduction

Coreference resolution is one of the best known tasks in the Natural Language Processing (NLP) community. Despite a large body of work in the area throughout the last few decades (Morton 2000; Bean and Riloff 2004; McCallum and Wellner 2005; Rahman and Ng 2009), the task remains challenging, as 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 requirements may become more important when cues such as gender or plurality do not by themselves indicate the correct resolution (Levesque, Davis, and Morgenstern 2011).

To date, most existing methods for coreference resolution (Raghunathan et al. 2010; Lee et al. 2011; Durrett, Hall, and Klein 2013; Lee et al. 2017) have been evaluated on a few popular datasets, including the CoNLL 2011 and 2012 shared coreference resolution tasks (Pradhan et al. 2011; Pradhan et al. 2012). These datasets were initially proposed as the first comprehensively tagged and large-scale corpora for coreference resolution, in the hope that they would contribute to improvements in state-of-the-art techniques. These improvements would ideally yield systems that rely on more than just shallow semantic features like

number, gender, and semantic class, and more on world knowledge and the context of the mention. According to (Durrett and Klein 2013), these systems would contribute in the “uphill battle” to modelling not just syntax and discourse, but also semantic compatibility.

Despite the general success of these tasks, the question of what exactly current systems learn or exploit remains open, particularly with recent neural coreference resolution models that achieve high performance. For example, (Lee et al. 2017) astutely note that their state-of-the-art model does “little in the uphill battle of making coreference decisions require world knowledge”, listing a few examples in the CoNLL 2012 task that require external sources of knowledge with more complex inference. However, since these cases were infrequent in the tasks themselves, the best systems may still exploit the same cues whose avoidance originally inspired the CoNLL tasks. In addition, these systems have been observed to rely on societal stereotypes present in coreference datasets, which could significantly impact their performance for some demographics (Zhao et al. 2018).

There has been a recent trend to develop tasks that feature more challenging coreference problems. The most popular of these is the Winograd Schema Challenge (WSC), which has emerged as a popular alternative to the Turing test as a means to measure progress towards human-like artificial intelligence (Levesque, Davis, and Morgenstern 2011).

The WSC task is carefully controlled such that heuristics involving syntactic salience, the number and gender of the antecedent, or other simple syntactic and semantic cues become ineffective. Previous approaches to common sense reasoning, for instance 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. Others have developed systems for relaxed common sense datasets with looser constraints (Rahman and Ng 2012; Peng, Khashabi, and Roth 2015). These shortcomings can at least in part be attributed to the limited size of the WSC corpus, which is a natural side-effect of it being expertly hand-crafted.

In general, endeavors to tackle coreference resolution reveal that, as much as the innovative strides in models and techniques are crucial to progress, the nature and de-

^{*}Equal contribution.

mands of the tasks themselves guide and possibly constrain progress. As much as datasets like WSC are limited in size or naturalness, large-scale datasets like CoNLL are limited in their challenge level, exhibiting vulnerability to work-around tricks that both distinguish and confound neural models.

Accordingly, we propose the following desiderata for coreference resolution tasks. Such tasks should be

- **Challenging:** in the two CoNLL shared tasks, most instances can be resolved using syntactic salience, the number and gender of the antecedent, or other simple syntactic and semantic cues.
- **Large-scale:** WSC is small (fewer than 1000 instances), resulting in difficulties in evaluation (i.e, obtaining a substantial train-test split) and for neural models that require much more training data.
- **Based on natural text:** Sentences in, e.g., WSC are written from scratch by experts, requiring substantial time and effort to produce and inducing a bias specific to the person/s who wrote them. It is important that sentences occur in text written for other purposes.

We present a coreference resolution dataset that aims to fulfill these desiderata. The main contributions of this paper can be summarized as follows:

1. We construct a preliminary dataset of 1275 samples comprising a binary-choice pronoun disambiguation task. These samples require significant common sense and background knowledge to solve. We generate samples semi-automatically (which makes the dataset easily scalable), automatically scraping candidate sentences from a single corpus, automatically altering the antecedents in the sentences to be identical as far as gender, number, and semantic class, and validating the resulting sentences manually (if they are easy to resolve with common sense).
2. We demonstrate the difficulty of our preliminary dataset by showing the failure of five state-of-the-art systems to outperform a random baseline, while measuring high human performance.
3. We analyze the behavior of the state-of-the-art methods and empirically demonstrate that they generally avoid using the context to make a decision.

Related Work

Standard coreference resolution

Early automated techniques for standard coreference resolution, that is, the task of correctly partitioning the entity and events that occur in a document into resolution classes, date back decision trees and hand-written rules (Hobbs 1977; McCarthy and Lehnert 1995). The earliest corpora developed for standard coreference resolution date back to the Message Understanding Conferences (MUC) (Consortium and others 2001; Chinchor and Sundheim 2003) and the ACE (Doddington et al. 2004). These targeted noun phrases that were tagged with coreference information, but were

either limited in size or had annotation that was restricted to a small subset of entities.

In light of these hindrances, the datasets of (Pradhan et al. 2011; Pradhan et al. 2012) were proposed as large-scale corpora that boasted high inter-annotator agreement, obtained by restricting the coreference to phenomena with annotations of high consistency, as well as being accompanied by a standard evaluation scenario to capture various ranges of performance on competing systems.

Largely due to the recognition and quality of these two tasks, many coreference resolution systems emerged, ranging from entirely hand-engineered systems to advanced learning approaches. The multi-pass sieve-based system of (Raghunathan et al. 2010) is fully deterministic and makes use of mention attributes like gender and number, boasting the best results on the CoNLL 2011 task for a number of years (Lee et al. 2011). Later, highly lexical learning approaches took over as the new state-of-the-art (Durrett and Klein 2013), followed by more recent neural models (Wiseman, Rush, and Shieber 2016; Clark and Manning 2016). The current state of the art on the CoNLL 2012 task is the first end-to-end coreference resolution model that does not rely on a syntactic parser or hand-engineered mention detector (Lee et al. 2017).

Gender bias in standard coreference resolution

(Zhao et al. 2018) observed that state-of-the-art methods for the standard coreference resolution are gender-biased, exploiting various stereotypes that appear in society. They study two hypotheses to explain this bias: the training data and the external information sources (word embedding). Accordingly, they devise a new dataset called *WinoBias* that serves both as gender-bias test for coreference resolution models and which can also be used as a training set to compensate for the lack of diversity in previous corpora (i.e the 2 CoNLL tasks). This new dataset contains 3160 sentences and covers cases that require an understanding of both semantics and syntax. The sentences contain entities corresponding to people referred to by occupation. 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.

As illustrated, the goal is to obtain a balanced dataset that would ultimately remove gender stereotypes. Experiments were conducted using three different models: the Stanford coreference resolution system (Raghunathan et al. 2010), the coreference resolution system of (Durrett and Klein 2013), and the recent system that tackles this task end-to-end (Lee et al. 2017). The results show that the end-to-end neural model maintains the same performance without the gender bias when trained partially on both the previous datasets and *WinoBias*.

A concurrent work, (Rudinger et al. 2018), also proposed an empirical study of the biases existent in current coreference resolution systems. As in the previous paper,

they focus on gender bias by occupation. In contrast to (Zhao et al. 2018), however, who attribute the bias in part to the datasets, they conjecture that the gender bias comes primarily from the models themselves. They study three different methods: the rule-based Stanford coreference model (Raghunathan et al. 2010), statistical methods (Björkelund and Kuhn 2014; Durrett and Klein 2013) with hand-crafted features and a neural model (Clark and Manning 2016). Based on statistics from the Bureau of Labor, they show that the three models contains gender bias that are present in today’s society.

The work on gender stereotypes already provides some insight into the behavior of current models. In the example above, if *she* is predicted as being *the secretary*, it is because the model learned a representation for this profession that contains gender information and based its prediction on that. The tested models do not capture the context and the relation between *was overwhelmed* and *hired*. Our work does not focus specifically on gender stereotypes. Unlike WinoBias, our task is composed of sentences that occur naturally in text. Moreover, we devise an experiment to show empirically that current state-of-the-art models often rely on the entity itself and ignore the context when making a decision.

“Difficult” coreference resolution

As the creators of the CoNLL tasks themselves note, most techniques rely primarily on surface-level features, such as the proximity between mentions, as well as shallow semantic features such as number, gender, named entities, semantic class, etc. They posit that a reduction of this knowledge gap is essential for the advancement of coreference resolution systems and that an important step towards that includes their large, standardized corpora (Pradhan et al. 2011).

A manually constructed dataset of challenging pronoun disambiguation problems called the Winograd Schema Challenge was proposed by (Levesque, Davis, and Morgenstern 2011). It was designed with the goal that any successful system would necessarily have reduced this knowledge gap. Although the WSC has shown itself to be an important means to evaluate systems en route to achieving human-like intelligence, as far as in the task of pronoun disambiguation, the size and quality of the dataset has become a crucial bottleneck for progress. A Winograd-like expanded corpus was proposed by (Rahman and Ng 2012) in order to partially address the WSC’s size limitation. However, systems that perform well on the expanded dataset (Rahman and Ng 2012; Peng, Khashabi, and Roth 2015) do not report similar results on the original WSC, likely due to loosened constraints in the former.

The task that we propose distinguishes itself from the WSC (written from scratch by few experts) by being constructed from sentences that occur naturally in text. This contributes to much more diverse and pervasive problem instances. It is particularly important that, as well as being challenging, these tasks contain cases that are widely representative, so that improvements are more likely to transfer to the full coreference resolution task. The process by which

we construct the dataset is also largely automatic, and so is amenable to significant scaling.

Hard-CoRe Coreference Task

We have developed a task called Hard-CoRe (a play on the expression “Hard Co-Reference”) that features 1275 difficult pronoun disambiguation sentences. Specifically, each task instance is a short passage containing a target pronoun that must be correctly resolved to one of two possible antecedents. As an example:

- (3) Yifei goes about to woo Kishori, so that [she] can get the letters possessed by her. Who is [she]? (Answer: Yifei)

Problem instances are controlled to not give away the pronoun’s antecedent through cues involving syntactic salience, the number and gender of the antecedent, or other simple syntactic and semantic cues. Systems must make common sense inferences; i.e., that someone who woos another may be doing so to acquire something.

In the following section, we will describe the methodology used to construct the task dataset, provide a glimpse of a few of its instances and resolution rationale, outline the evaluation criteria used, and describe the task characteristics.

Dataset Creation

We create the dataset using a semi-automatic process that performs the following steps:

First, a collection of documents (in this case, the 2015 English Wikipedia) is processed to remove markup, non-ASCII characters, parenthetical expressions, headings and lists. Next, we pass each sentence through a pipeline that filters out sentences that either do not fit the structure presented in the previous section or where the pronoun could be easily resolved based on surface features such as entity type. Finally, the remaining sentences are manually inspected to ensure that a system can not rely on shallow syntactic cues such as gender and number and that they are not ambiguous. The details of the pipeline are described below.

In a first step, we use regular expressions to ensure the presence of connectives ¹. We keep sentences which have between 9 and 33 words after naïve tokenization, start with an upper case letter and contain no math. 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 the connective. As a final check, we ensure that no pronoun occurs before the connective, which tends to remove sentences which are not self-contained.

On the remaining set, we use the fast Stanford’s Maxent tagger (Toutanova et al. 2003) to infer a flat part-of-speech (POS) tag labelling. Using the POS tags, we ensure that there are two nouns before the connective which do not re-occur after the connective. A noun re-occurrence after the connective indicates that the pronoun likely refers to the non-repeated noun.

¹comma, semicolon, *or*, *since*, *but*, *because*, *although*, etc.

HardCoRe Example 1:	<u>Radu</u> appeared to be killed by <u>Brother Paulo</u> , but [he] reappears a short while later injured, but alive. (Answer: Radu)
Original sentence:	Radu appeared to be killed by Sister Paula, but he reappears a short while later injured, but alive.
HardCoRe Example 2:	<u>Wanda</u> tries to apologize to <u>Rose</u> , but [she] refuses to accept. (Answer: Rose)
Original sentence:	Warren tries to apologize to Rose, but she refuses to accept.
HardCoRe Example 3:	<u>Tom</u> arrives to where <u>Alex</u> was tied, but [he] has come free of his lead. (Answer: Alex)
Original sentence:	Tom arrives to where Vanessa was tied, but she has come free of her lead.
HardCoRe Example 4:	<u>Carl</u> then realizes that <u>Peter</u> is innocent if [he] genuinely believed him to be guilty. (Answer: Peter)
Original sentence:	Carla then realizes that Peter is innocent if he genuinely believed her to be guilty.

Table 1: Examples of Hard-CoRe instances.

In the final, most expensive check, we use Stanford CoreNLP parser (Manning et al. 2014) to determine number, type and gender of the noun phrases (NPs). We ensure that there are exactly two NPs before the connective, and that we can use the pronoun to distinguish between them (e.g. “they” indicates plural, “he/him” indicates male person). The mentioned checks resulted in 72,243 sentences. 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 kept only sentences where the type indicates that both NPs correspond to persons, which further filtered the remaining sentences to a total of roughly 5931 sentences. We do this because NPs which denote people are often named entities or can easily be replaced by named entities without loss of information. In addition, since our hypotheses relate largely to how current resolution systems use gender cues, the bulk of the gendered pronouns can be found when they refer to such NPs and so we sought to efficiently target these instances.

For the remaining 5931 sentences, we examined/adjusted them in a series of five steps to control their quality and difficulty:

1. We examine the resolution provided by (Raghunathan et al. 2010)’s system (via CoreNLP) for the sentence. The sentence, often structured so that only one of the entities shares the same gender as the pronoun, is correctly resolved by the resolution system anyways, but we check that the resolution is correct for each sentence regardless. The resolutions serve as the gold label for our task.
2. Then, we address the gender giveaways by replacing one of the named entities so that both entities and the pronoun match in terms of gender (e.g., in a sentence with “James”, “Jessica”, and “she” as the NPs and pronoun, we replace “James” with “Jane”).
3. Next, we ensure there are no structural giveaways (e.g., replacing “Bob defeats and stakes Malik, but fails to kill [him]”, with “Bob defeats and stakes Malik, but *he* fails to kill [him]”)
4. We remove the sentence if, after the above three checks/modifications, the resolution is not straightforward or logical.

After applying the above rules, there remain 1275 sen-

tences whose pronoun disambiguation cannot involve shallow semantic features like gender, number, and type, but instead requires varying degrees of external knowledge. These sentences constitute the HardCoRE dataset. Examples of some of its instances are provided in Table 1. As the examples reveal, each passage may require a unique instance of common sense knowledge to resolve.

In the first example, our understanding that death (by way of killing, for example) causes a disappearance helps us conclude that Radu, the victim of the Brother Paulo’s murder, is the one that reappearance most plausibly refers to.

In the next example, we quickly infer that *accept* is related to *accepting the apology*. Therefore, *she* refers to the one that is receiving apologies, which is Rose.

For the third example, our understanding that *being tied* is related to being deprived of freedom allows us to conclude that it is Alex who has come free.

In the fourth example, we first see that Peter is realized to be innocent by Carl and that the second clause is there to justify the realization. In reading the second clause, a candidate person (either Peter or Carl) is said to have genuinely believed the other to be guilty. An important trait of an innocent person can be their genuine belief that they are indeed innocent, which, in turn, may also result in their genuine belief that someone else is guilty. Accordingly, it fits better that it was Peter who had this genuine belief, and this is what gives Carl the justification he needs to realize Peter’s innocence.

Task Characteristics

Sentence Characteristic	% of Data
Masculine target pronouns	51.5
Feminine target pronouns	44.5
Non-Gendered target pronouns	4.0
First Antecedent Correct	50.7
Second Antecedent Correct	49.2

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

In Table 2, we report the characteristics of the data, which suggests a roughly equal distribution of feminine and masculine target pronouns (*he/him/his* vs. *she/her*) as well as

Model	Both Antecedents Predicted	No Decision	Incorrect Decision	Correct Decision	Task-Specific Accuracy
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	0.130	0.10	0.39	0.38	0.50
Random	–	–	–	–	0.50
Human ¹	–	–	–	–	0.92 (0.93 agreement)

Table 3: Coverage and performance of various representative systems on the Hard-CoRe Task.

an equal distribution of the two labels, which keeps random baseline performance at 50% expected accuracy. The significantly lower proportion of non-gendered target pronouns may be due to our final filtering step, where we kept only sentences in which the antecedents correspond to persons.

Evaluation

Our task requires the model to choose between two candidates, but classical coreference models build clusters of expressions that refers to the same entity. With respect to our setting, several errors can be made by these existing models: the two entities and the pronoun share a similar cluster (*Both Antecedents Predicted*), none of the two candidates shares a cluster with the pronoun (*No Decision*) or a cluster with the wrong candidate and the pronoun is created (*Incorrect Decision*). To obtain a score specific to our task, where only one candidate can be linked to the pronoun, we compute a *Task Specific Accuracy* that counts the number of correct decisions when only one of the two candidates is chosen.

Given that there may exist a degree of subjectivity to the answers, we also determine a human performance baseline on the task. This equates accuracy to the percentage of problem instances for which a majority of people make the correct prediction. We explain this procedure in detail in the next section.

Experiments and Results

In this section, we compare the performance of five representative 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. 2017)’s end-to-end neural system (**E2E**). We also report human performance (**Human**).

Human Performance: We determined the human performance 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 gold label

derived from the original sentence. To determine the agreement statistics, only predictions that amounted to a strong majority decision (at least 5/6 agreement) among all six speakers were considered to agree.

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 viability of the task. The performance of the automatic systems, random or slightly above random, demonstrates that state-of-the-art coreference resolution systems are unable to solve the task. Moreover, the end-to-end neural model struggles to distinguish the two candidates in 13% of the sentences by merging them into the same cluster.

Analysis by Switching Entities

We hypothesize that the current systems mostly rely on gender and number cues that are related to the entity itself and not the context. This causes performance to suffer when sentences contain gender-unbiased entities. We propose an experiment to validate our hypothesis. We use the current dataset and switch the two candidates every-time they appear in the sentence. If the coreference models rely on the context, the predicted candidate should change as well. If it relies on the entity itself, the output should stay the same.

An example of a switched sentence :

- (4) Original sentence: Alex tells Paulo , but [he] does not believe him. (Answer: Paulo)
Switched sentence: Paulo tells Alex, but [he] does not believe him. (Answer : Alex)

Table 4 shows the proportion of predicted candidates that remain the same after switching. The rule-based system (Raghunathan et al. 2010) always resolves the same entity, suggesting that the decision is only based on the entity itself and that the context is entirely ignored. The hard decision behind this model mostly relies on a gender and number dictionary (Bergsma and Lin 2006). This dictionary is 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.

¹This is an estimate based on a subsample of the data.

Model	Fraction of unchanged decisions from antecedent-switching
Rule	1.00
Stat	0.24
Deep-RL	0.34
Latent	0.22
E2E	0.29

Table 4: The sensitivity of various representative systems on the antecedent of a sentence according to the number of unchanged decisions when the antecedents are switched.

The other models, Stat, Deep-RL, E2E, and Latent, are much more robust to this experiment, demonstrating their use of context cues to make their decision. Yet, it is surprising that the model proposed by (Martschat and Strube 2015), which relies only on the previous and next token of each candidate as a context feature, outperforms the end-to-end model. The latter uses a Bidirectional LSTM to build a context-dependent representation of each candidate. It is likely that the model adapts to the training set, OntoNotes 5.0, by focusing on gender/number cues which represent the majority of pronoun disambiguation.

While an ideal coreference model would make use of common-sense reasoning and context cues, we show that current state-of-the-art methods for coreference resolution fail to perform well on the proposed task. The additional experiment demonstrates that the decision is often motivated by the entity itself and not by the context, especially when using rule-based methods. Yet, the performance of the method proposed by (Martschat and Strube 2015) is above random and suggests that there is space for improvement.

Conclusion

We presented a coreference task that features high human performance and inter-agreement but random or slightly above-random performance for various state-of-the-art resolution systems, including rule-based, graphical, and neural models. In addition, we demonstrated that the antecedents of a sentence seem to have the greatest impact on the resolution decision of these models, a behaviour that suggests their reliance on shallow semantic features like number and gender. In turn, models that exploit these features may perform very well on certain coreference tasks, but not on those for which a significant degree of common sense and world knowledge is required.

In the future, we plan to develop a larger corpus of at least 100,000 problem instances, by using additional text-data sources (besides Wikipedia) and by developing a crowd-sourcing protocol to accept or reject candidate sentences processed by our the automatic modifications and checks. This would represent a corpus amenable to powerful neural models, which could serve as an important training set for models that have thus far performed poorly on difficult coreference problems like the Winograd Schema Challenge.

References

- [Bailey et al. 2015] Bailey, D.; Harrison, A.; Lierler, Y.; Lifschitz, V.; and Michael, J. 2015. The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.
- [Bean and Riloff 2004] Bean, D., and Riloff, E. 2004. Un-supervised 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*.
- [Bergsma and Lin 2006] Bergsma, S., and Lin, D. 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*, 33–40. Association for Computational Linguistics.
- [Björkelund and Kuhn 2014] Björkelund, A., and Kuhn, J. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 47–57.
- [Chinchor and Sundheim 2003] Chinchor, N., and Sundheim, B. 2003. Message understanding conference (muc) 6. *LDC2003T13*.
- [Clark and Manning 2015] Clark, K., and Manning, C. D. 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 (Volume 1: Long Papers)*, volume 1, 1405–1415.
- [Clark and Manning 2016] Clark, K., and Manning, C. D. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.
- [Consortium and others 2001] Consortium, L. D., et al. 2001. Message understanding conference (muc) 7. *LDC2001T02. FTP FILE. Philadelphia: Linguistic Data Consortium*.
- [Doddington et al. 2004] Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, 1.
- [Durrett and Klein 2013] Durrett, G., and Klein, D. 2013. Easy victories and uphill battles in coreference resolution. In *EMNLP*, 1971–1982.
- [Durrett, Hall, and Klein 2013] Durrett, G.; Hall, D.; and Klein, D. 2013. Decentralized entity-level modeling for coreference resolution. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 114–124.
- [Hobbs 1977] Hobbs, J. R. 1977. Pronoun resolution. *ACM SIGART Bulletin* (61):28–28.

- [Lee et al. 2011] Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; and Jurafsky, D. 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*, 28–34. Association for Computational Linguistics.
- [Lee et al. 2017] Lee, K.; He, L.; Lewis, M.; and Zettlemoyer, L. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.
- [Levesque, Davis, and Morgenstern 2011] Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, 47.
- [Liu et al. 2016] Liu, Q.; Jiang, H.; Evdokimov, A.; Ling, Z.-H.; Zhu, X.; Wei, S.; and Hu, Y. 2016. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.
- [Manning et al. 2014] Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.
- [Martschat and Strube 2015] Martschat, S., and Strube, M. 2015. Latent structures for coreference resolution. *Transactions of the Association of Computational Linguistics* 3(1):405–418.
- [McCallum and Wellner 2005] McCallum, A., and Wellner, B. 2005. Conditional models of identity uncertainty with application to noun coreference. In *Advances in neural information processing systems*, 905–912.
- [McCarthy and Lehnert 1995] McCarthy, J. F., and Lehnert, W. G. 1995. Using decision trees for coreference resolution. *arXiv preprint cmp-lg/9505043*.
- [Morton 2000] Morton, T. S. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 173–180. Association for Computational Linguistics.
- [Peng, Khashabi, and Roth 2015] Peng, H.; Khashabi, D.; and Roth, D. 2015. Solving hard coreference problems. *Urbana* 51:61801.
- [Pradhan et al. 2011] Pradhan, S.; Ramshaw, L.; Marcus, M.; Palmer, M.; Weischedel, R.; and Xue, N. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 1–27. Association for Computational Linguistics.
- [Pradhan et al. 2012] Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, 1–40. Association for Computational Linguistics.
- [Raghunathan et al. 2010] Raghunathan, K.; Lee, H.; Rangarajan, S.; Chambers, N.; Surdeanu, M.; Jurafsky, D.; and Manning, C. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 492–501. Association for Computational Linguistics.
- [Rahman and Ng 2009] Rahman, A., and Ng, V. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 968–977. Association for Computational Linguistics.
- [Rahman and Ng 2012] Rahman, A., and Ng, V. 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*, 777–789. Association for Computational Linguistics.
- [Rudinger et al. 2018] Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.
- [Toutanova et al. 2003] Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 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-Volume 1*, 173–180. Association for Computational Linguistics.
- [Wiseman, Rush, and Shieber 2016] Wiseman, S.; Rush, A. M.; and Shieber, S. M. 2016. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.
- [Zhao et al. 2018] Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. *arXiv preprint arXiv:1804.06876*.