# A Survey on the Development and Applications of Coreference Resolution

**Bryce Wong**
University of California, Berkeley
brywong@berkeley.edu

## Abstract

Within the past decade, the task of coreference resolution has become popular in addressing a variety of NLP-related procedures. This paper aims to summarize the development of the state-of-the-art model for coreference resolution. In addition, commentary will be provided on the applications of coreference resolution regarding different NLP tasks and some concerns regarding modern coreference resolution models.[1]

## 1 Introduction

Coreference resolution is the task of identifying tokens within a document that refer to the same entity. For example, if I were given the sentence: "Alice told me that she lost her keys while I was shopping," the entity "Alice" would be linked with "she" (5th word) and "her" (7th word) whereas the entity "I" would be linked with "me" (3rd word). Although the task itself is quite simple, there is a great deal of complexity that comes from encoding referential data in each document. However, finding an efficient solution to this problem can result in major improvements across many pre-existing NLP tasks.

## 2 Coreference Resolution Models

The following section discusses the history behind the development of today's modern state-of-the-art model. Additionally, some insights are provided into other model variants that view coreference resolution from a unique perspective.

### 2.1 The State-of-the-art Model

The first state-of-the-art model was proposed in 2017 by a research group from the University of Washington in conjunction with Facebook AI. Their approach involved analyzing all spans (chunks of tokens) within each document and training a model to generate embeddings that capture contextual information between neighboring spans. The embeddings were designed such that the model, given a specific span, could predict with high probability other spans that precede it. This was the first approach that did not rely on syntactic parsing, or utilizing the grammatical structure of the document for coreference resolution (Lee et al., 2017).

Many developments have been made since this initial model's release. One of the first major changes was altering the model to include ELMo and GloVe embeddings in 2018, converting each token within a document into its corresponding vector embedding before passing the input into the original model (Peters et al., 2018). These embeddings were effective since they were generated from models trained on large text datasets with the intent of capturing relationships between different words and entities. In 2019, the development of BERT led to the creation of a BERT-based coreference resolution model, where the pre-trained embeddings were replaced by a BERT transformer that was more effective at identifying related entities within a document (Joshi et al., 2019).

Despite the effectiveness of these models, the runtime can be quite slow due to processing a large amount of spans per document. As such, the current state-of-the-art model utilizes word-level coreference resolution, identifying coreference between individual words first, then generalizing these results into coreference between an entire span of tokens, reducing the number of spans from $O(n^4)$ to $O(n^2)$, where $n$ is the length of the document. (Dobrovolskii, 2021). In addition to the reduced runtime, the reduced memory requirement makes it easier to implement coreference resolution in other models, thus serving as a flexible NLP tool with limited drawbacks.

[1]Word Count: 1931

### 2.2 Model Variations

In the past few years, there have also been attempts to explore different models that are less reliant on pre-existing constructions. For example, CorefQA was proposed in 2020, where coreference resolution tasks were redefined as a series of question-answer prompts. After providing CorefQA with a document to analyze, the user would ask the model questions about different entities and the specific tokens that refer to each one (Wu et al., 2020b).

Additionally, the F-COREF model was proposed in 2022 to promote faster coreference resolution, significantly decreasing the amount of training time required for state-of-the-art models in exchange for a small accuracy decrease (Otmazgin et al., 2022a). Most of the simplifications came from their implementation of knowledge distillation: training a smaller model to mimic the performance of a larger state-of-the-art model (in this context, the LingMess model, Otmazgin et al., 2022b).

Some models have also attempted to remove the span construction altogether. For example, a research group from Tel Aviv University altered current coreference resolution models by propagating contextual span information within a document to the document's start and end tokens, thus making the model much more memory-efficient (Kirstain et al., 2021).

## 3 Evaluation Metrics

Historically, the standard dataset used to analyze model performance for coreference resolution tasks is the OntoNotes 5.0 corpus. This collection contains millions of entries from a variety of text genres, such as news articles, online blogs, and telephone conversations. Additionally, the corpus contains entries in three different languages: English, Chinese, and Arabic. The corpus is applicable for a variety of NLP tasks, and the provided dataset contains millions of coreference annotations over the dataset (Pradhan et al., 2013).

Despite its widespread use, the dataset itself is also somewhat limited due to the lack of variety in its domain sources. As such, there have been attempts at increasing the scope of current coreference resolution datasets in order to make coreference models more generalizable. One instance of this is the creation of a coreference resolution dataset with an English literature corpus spearheaded by the School of Information at the University of California, Berkeley (Bamman et al., 2020). Additionally, there are efforts in translating pre-existing English datasets to include other languages besides Chinese and Arabic. Wino-X is one example of this: their dataset includes German, Russian, and French translations of coreference resolution entries that are derived from the WinoGrande dataset (Emelin and Sennrich, 2021; Sakaguchi et al., 2019).

## 4 Applications

In this section, some applications of coreference resolution are provided in the context of other NLP tasks. Note that the following list is not exhaustive; however, it gives a good overview of how coreference resolution is utilized today.

### 4.1 Machine Translation

Machine translation is the task of translating an input document into another language. Since the grammar between two languages can differ drastically, it is useful to utilize coreference resolution to identify helpful tokens (i.e. pronouns) that refer to key entities within an input text. Some modern machine translation systems utilize a graph-based encoder that contextualize relationships between tokens over the entire document – this data representation makes it easier to directly translate individual tokens and reconstruct the translated tokens into a valid syntactical structure (Ohtani et al., 2019). This point is especially important when considering languages with gendered grammar systems (such as French or Spanish), where certain words have either masculine or feminine determinants associated with them (Yehudai et al., 2023).

### 4.2 Machine Reading Comprehension

Machine reading comprehension, also commonly known as question-answering, is an NLP task where the model is given a passage to read and understand. After, the model should be able to answer any potential question a user may ask about the passage. Although state-of-the-art machine reading comprehension models are effective at answering general questions (usually related to the content in the input passage), these models typically struggle with questions about coreference resolution (i.e. which tokens correspond to a given entity in the passage). To remedy this issue, some researchers have worked on including more coreference resolution prompts in machine reading comprehension datasets, as well as designing an automated

pipeline for converting pre-existing coreference resolution annotations into question-answer prompts (Wu et al., 2020a).

### 4.3 Text Simplification

Text simplification aims to transform an input document into a form that makes it easier for individuals to read. These tasks are typically targeted towards a specific demographic group, such as individuals who are dyslexic or children who lack higher-education reading skills. In the past, text simplification was achieved solely through analyzing the syntactic and lexical structure of the input documents. By building "coreference chains" that link tokens to the specific entities that they are describing, it is possible to design an algorithm that can modify the structure of these chains to yield a simplified sentence that is easier to process and understand (Wilkens et al., 2020).

### 4.4 Event Coreference Resolution

Event coreference resolution is the task of analyzing which fragments of text refer to the same real-world event within a single document or collection of documents. This application of coreference resolution is most useful in the context of analyzing news articles, as identifying different parts of a document that refer to the same event can be useful in extracting the essence of the text (Lu et al., 2020). Note that there are a lack of annotated datasets that contain event coreference resolution annotations; as such, there have also been attempts at automating the collection of "coreferential event pairs," or pairs of documents that refer to the same event (Choubey and Huang, 2021).

### 4.5 Dialogue Processing

Coreference resolution also has utility in dialogue processing: analyzing a stream of documents that contain conversations between two or more individuals. Note that the model will be analyzing dialogue in real-time, meaning that the model will sequentially execute coreference resolution between the current document it receives and all of the previous documents it has seen during the conversation (Xu and Choi, 2022). This type of NLP task can be useful in the context of character linking, or identifying parts of a dialogue that refer to actual individuals in the real world (Bai et al., 2021).

## 5 Concerns

Although coreference resolution has many practical uses in a variety of NLP tasks, there are still a few concerns that impact the efficacy of these models. A few of these issues are summarized in this section.

### 5.1 Domain Generalization

One main issue with coreference resolution is that models typically struggle with generalizing to different domains of data that they have not been exposed to while training. This issue is not important for narrow coreference tasks that focus on specific subsets of data; however, this is an issue for general coreference models intended to work on all types of documents and scenarios. Although an effective solution to this problem is consolidating multiple pre-existing coreference resolution datasets into one large dataset for a model to train on, it does seem like the main bottleneck comes from the lack of annotated data from a variety of domains (Toshniwal et al., 2021).

### 5.2 Language Variety

Building off of the previous point, there seems to be a lack of coreference resolution datasets that generalize to different languages, both verbal and nonverbal. However, there are many researchers that are currently working on generating new data sources to fill the gaps. As previously discussed in the paper, there have been some attempts to generate datasets in other languages like German or French (Wino-X, Emelin and Sennrich, 2021). Additionally, some strides have been made in interpreting coreference resolution for sign language, although most of the difficulty stems from extracting features from different hand gestures (Yin et al., 2021). There has also been some experimentation in utilizing machine translation models to automate the translation of coreference resolution datasets from English to other languages; however, the quality of modern machine translation tools limits this pipeline significantly (Bitew et al., 2021).

### 5.3 Gender Bias

Current state-of-the-art models can also exhibit some gender bias in the annotations they make due to the lack of gender diversity in coreference resolution datasets. For example, some models may make assumptions about the target entities and their professions: for example, the model may

associate "physician" with "male" and "secretary" as "female" regardless of other provided context in the document (Zhao et al., 2018). In fact, some studies have shown that BERT models trained on gender-balanced datasets with an equal number of male/female entries can perform better than some state-of-the-art coreference resolution models in the context of gender pronoun resolution (Chada, 2019).

### 5.4 Identity Bias

Additionally, the lack of diverse identities (queer/non-binary) within coreference resolution datasets causes these models to perform inaccurately on documents involving LGBTQ+ individuals. For example, many models fail to link the pronouns "they" and "them" with singular entities that may not identify with binary pronouns (Dev et al., 2021). However, some progress has been made by curating more gender-inclusive datasets with non-binary pronouns and augmenting pre-existing coreference resolution datasets to include more neutral pronouns (Uppunda et al., 2021).

## 6 Conclusion

Coreference resolution has become a crucial NLP task within the past couple of years. Given its variety of practical applications and recent improvements in the speed and memory-usage of current state-of-the-art models, it is not surprising to see coreference resolution being implemented in many modern NLP pipelines today. However, it is important to be weary of the limitations and biases that are present within coreference resolution datasets, as it is critical to address these concerns now before they grow out of control.

## References


Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint coreference resolution and character linking for multiparty conversation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 539–548, Online. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Lazy low-resource coreference resolution: a study on leveraging black-box translation tools. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 57–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rakesh Chada. 2019. Gendered pronoun resolution using BERT and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133, Florence, Italy. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Automatic data acquisition for event coreference resolution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1185–1196, Online. Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019. Bert for coreference resolution: Baselines and analysis.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. 2020. End-to-end neural event coreference resolution.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50, Hong Kong, China. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022a. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022b. Lingmess: Linguistically informed multi expert scorers for coreference resolution.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution.

Ankith Uppunda, Susan Cochran, Jacob Foster, Alina Arseniev-Koehler, Vickie Mays, and Kai-Wei Chang. 2021. Adapting coreference resolution for processing violent death narratives. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4553–4559, Online. Association for Computational Linguistics.

Rodrigo Wilkens, Bruno Oberle, and Amalia Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100, Marseille, France. European Language Resources Association.

Mingzhu Wu, Nafise Sadat Moosavi, Dan Roth, and Iryna Gurevych. 2020a. Coreference reasoning in machine reading comprehension.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020b. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2022. Online coreference resolution for dialogue processing: Improving mention-linking on real-time conversations.

Asaf Yehudai, Arie Cattan, Omri Abend, and Gabriel Stanovsky. 2023. Evaluating and improving the coreference capabilities of machine translation models.

Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021. Signed coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4950–4961, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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