A Survey on The State of the Art in Coreference Resolution

Rachel Lin
University of California, Berkeley
raelin@berkeley.edu

Abstract

Coreference resolution has seen significant advances in recent years, solidifying its role as a fundamental task in natural language processing (NLP). This paper aims to provide an overview of this topic, covering the major developments that have led to today's state-of-the-art coreference resolution models. It will explore the datasets and evaluation metrics widely used in the field, along with various methods applied in coreference resolution. Additionally, it will examine the limitations and challenges that persist, highlighting areas for future research and improvement. By the end of this paper, readers should have a solid understanding of the core concepts, current trends, and ongoing issues in coreference resolution. [1]

1 Introduction

Coreference resolution is the task of identifying which entity a mention (noun, pronoun, or phrase) refers to in a text. It helps connect different elements within a document, enabling a deeper understanding of the content. For instance, in the following sentences, "John went to the farm to buy milk. He was extremely pleased with his purchase," coreference resolution links 'John' as 'he,' indicating that John was pleased with his milk purchase. Coreference resolution is crucial in several natural language processing tasks, such as question answering, text summarization, and machine translation (Wu et al., 2020; Wilkens et al., 2020; Yehudai et al., 2023).

2 Datasets and Evaluation Metrics

2.1 Datasets

The OntoNotes 5.0 corpus, encompassing texts in English, Chinese, and Arabic, has become the standard benchmark for evaluating coreference resolution models. It provides a large-scale corpus of conference annotations, derived from various sources such as magazines, news, web blogs, and conversational speech. Building on the OntoNotes 5.0 corpus, the CoNLL 2011/2012 shared task provides a standardized framework to evaluate the performance of new coreference resolution models. The CoNLL 2011/2012 shared task establishes a consistent method for evaluating models by providing a standard set of training, validation, and test data, along with a set of evaluation parameters (Pradhan et al., 2011, 2012).

Although most modern systems use the OntoNotes dataset, several efforts have been made to improve coreference resolution performance by using more domain-specific corpora (Lee et al., 2017, 2018; Wiseman et al., 2015; Kantor and Globerson, 2019). This emphasis on domain-specific corpora stems from the observation that OntoNotes has been unable to generalize to new domains (Moosavi and Strube, 2017, 2018; Subramanian and Roth, 2019). For instance, to meet the demand for better performance in English literature, Bamman et al. (2020) introduced a new conference resolution dataset specifically designed for English literary texts. LitBank is a dataset created specifically for English literature, offering annotated coreferences from a variety of literary sources. The coreference annotations in this dataset are tailored to the distinctive style and structure of literary texts, which differ from the types of texts included in the OntoNotes dataset. Webster et al. (2018) offer another example of resolving ambiguous gendered pronouns in English texts to combat the gender bias found in existing corpora. The Gendered Ambiguous Pronouns (GAP) dataset offers a solution to existing corpora favoring masculine entities by presenting a more balanced representation of gendered pronouns.

[1] Word count: 1985

2.2 Evaluation Metrics

The CoNLL 2011/2012 shared tasks use an average F1 score of three metrics: *MUC*, B^3, and *CEAF*, each of which employs precision and recall to calculate a score. However, research by Moosavi and Strube (2016) has identified shortcomings in these evaluation metrics. *MUC* is a link-based metric that computes the number of missing or extra links between the gold-standard entities (key) and the system's output (response). However, it does not penalize over-clustered mentions or assign different weights to more popular mentions, making it difficult to capture the overall coherence relationship of the text (Duron-Tejedor et al., 2023). On the other hand, B^3 and *CEAF* are both mention-based metrics. B^3 calculates precision and recall based on individual mentions rather than links (Cai and Strube, 2010). B^3's focus on mentions allows it to overcome some of *MUC*'s shortcomings, but its limitation in handling repeated mentions in the response entities can lead to inflated precision or recall scores (Moosavi and Strube, 2016). Lastly, *CEAF* determines an optimal mapping between system-predicted clusters with gold-standard clusters based on a similarity measure. Although CEAF may be more robust to errors caused by over-clustering, it overlooks correct decisions involving unaligned response entities (Moosavi and Strube, 2016). Due to these limitations, Moosavi and Strube (2016) proposed a new metric called LEA (Link-based Entity-Aware). LEA considers link accuracy, entity importance, and entity-response matches, addressing the limitations of existing metrics and offering a more reliable method to evaluate coreference resolution.

3 Related Work

Traditional works on coreference resolution for English often employ a mention-ranking approach, selecting the highest-ranking antecedent of each mention span, a chunk of tokens (Durrett and Klein, 2013; Rahman and Ng, 2011; Wiseman et al., 2015). While this method is scalable and straightforward to train, it introduces a computational challenge: it would be impractical to evaluate all possible pairs of spans as most mentions do not have any coreferent links to earlier mentions (Lee et al., 2017; Wiseman et al., 2015). To address the limitations of the traditional approach, Clark and Manning (2016) introduced a neural network-based coreference system that operated by merging coreference clusters. This approach yielded improved performance on the CoNLL 2012 test data for both English and Chinese compared to the traditional methods.

In 2017, Lee et al. (2017) proposed the first state-of-the-art coreference resolution model that outperformed all the previous models that relied on a syntactic parser, including neural-based ones, on the OntoNotes 5.0 shared task. Their end-to-end neural model featured a span-ranking system that eliminated the need for a syntactic parser. By representing spans as vector embeddings, the model iteratively updated these representations while simultaneously learning how to cluster them.

Several models have surpassed the performance of Lee et al.'s model since its introduction in 2017. One approach to achieving this was using more advanced embedding techniques to generate better span representations. Peters et al. (2018) enhanced the performance of Lee et al.'s coreference resolution model through ELMo, an embedding technique that represents each word token as a function across multiple layers of a bidirectional language model (biLM). The biLM was pre-trained on a large text corpus, allowing it to encode syntactic information in a rich word representation, leading to better performance. Joshi et al. (2020) presented SpanBERT, an extension of the BERT model that emphasizes capturing relationships between spans of text. SpanBERT's pretraining approach concentrates on spans rather than individual tokens, enabling the model to learn about relationships among words within a span. It also considered span boundaries, placing additional focus on tokens at the edges of spans, which led to better performance for coreference resolution.

A drawback of pretrained language models is their slow runtime of due to the significant memory footprint associated with span representations (Kirstain et al., 2021). One solution to this problem was to eliminate the reliance on span representations, which commonly leads to a model complexity of $O(n^4)$ (n is the number of input tokens in a document). Kirstain et al. (2021) suggested a start-to-end coreference resolution model which uses span endpoints. By focusing on the start and end points of spans and retaining only the top πn scored

mentions, they were able to restrict the model to a quadratic complexity. Similarly, Dobrovolskii (2021) reduced the complexity to $O(n^2)$ by considering coreference links between words instead of spans and using RoBERTa. Both approaches are highly efficient and deliver performance that competes with current models on the OntoNotes benchmark.

4 Methods

This section introduces a few extensible approaches for coreference resolution, designed to address coreference in various natural language processing tasks

4.1 Question Answering

Wu et al. (2020) formalized coreference resolution as a question-answering task, an extension of machine reading comprehension. In their proposed model, CorefQA, the approach generates a query based on the context surrounding a mention, then extracts the text spans with the highest probability of being a coreferent answer. This approach addresses the limitation in many coreference models, known as mention proposal, where mentions left out during the mention proposal stage often cannot be recovered later (Bohnet et al., 2023; Wu et al., 2020). By turning coreference resolution into a question-answering problem, CorefQA allows for retrieving left-out mentions and provides a deeper examination of the connections between mentions and their contexts. This leads to a more flexible and comprehensive approach for coreference resolution, improving the system's capacity to find and link related text spans. Their approach has demonstrated improved performance over previous models on the CoNLL 2012 and GAP benchmarks.

4.2 Text Simplification

Wilkens et al. (2020) integrated automatic coreference resolution into an automatic text simplification system to assist people with language disabilities. The goal was to solve tasks by simplifying complex language structures, such as replacing complex words with simpler ones. By integrating coreference resolution, their system ensured referential clarity even as complex terms or mentions were substituted. For example, "the wolf" might be simplified to "a wolf," and "this hyena" could become "the hyena." This approach addresses the challenge of simplifying language without compromising the clarity of references, making it easier for individuals with language disabilities to understand and engage with the text.

4.3 Machine Translation

Yehudai et al. (2023) have explored coreference resolution in machine translation to maintain grammatical structure and ensure coherence when translating between languages. This is important for languages such as French and Spanish that have grammatical noun genders, which require consistency between gendered pronouns and nouns they refer to. Their approach aimed to eliminate the need for target-side annotations as a prerequisite for accurate translation, thus simplifying the translation process. While their model's performance has not reached the performance of state-of-the-art coreference resolvers, their work can reduce common errors related to gender mismatches in machine translation.

5 Limitations

Coreference resolution has become a fundamental task in NLP, but several factors limit its performance. This section summarizes a few of the key issues impacting the performance of coreference resolution systems, along with current research efforts aimed at addressing these challenges.

5.1 OntoNotes Domain Generalizability

As mentioned earlier, the OntoNotes 5.0 corpus is widely used to assess the performance of coreference resolution models, but it has several limitations. Since it is a collection of documents from limited domains, mainly from the 2000s, it might not represent the broader range of texts in more contemporary sources (Xia and Van Durme, 2021). In addition, the way OntoNotes annotations are structured can lead to issues. The corpus does not include singleton annotations—entities mentioned only once in a document. It also splits longer documents into smaller, independent parts for ease of annotation (Pradhan et al., 2012), disrupting the natural flow of longer texts and breaking up coreference chains. These factors contribute to the observation that models trained on OntoNotes 5.0 often struggle to generalize to new domains (Moosavi and Strube, 2017; Moosavi and Strube, 2018; Subramanian and Roth, 2019). To address these challenges, researchers have been developing new annotated datasets designed to better suit

specific domains, leading to improved performance compared to training on OntoNotes (Bamman et al., 2020; Webster et al., 2018).

5.2 Multilingual Expansion

Beyond the broad domain restrictions in current datasets, many do not support a wide range of languages. For example, OntoNotes only supports English, Chinese, and Arabic (Pradhan et al., 2011; Pradhan et al., 2012). While much of the ongoing work focuses on expanding resources for English language tasks, there have also been advances in developing coreference datasets for other languages such as German, French, and Russian (Emelin and Sennrich, 2021). Furthermore, efforts have been made to use cross-lingual coreference models that build on multilingual embeddings and language-independent features, providing a more universal approach to coreference resolution across various languages (Kundu et al., 2018).

5.3 Multilingual Expansion

Gender bias arises in coreference resolution models trained on corpora that favor masculine entities (Webster et al., 2018). Studies have shown that this bias reflects societal stereotypes due to imbalanced training data. In the OntoNotes dataset, over 80% of gender pronouns such as "he" and "she" refer to male entities, leading to a skewed representation of gender. As a result, coreference models trained on such data often perpetuate social stereotypes, with models more likely to associate terms such as "manager" or “programmer” with male pronouns (Rudinger et al., 2018; Zhao et al., 2018). To address this bias, Zhao et al. (2018) constructed an additional training corpus where male entities were swapped for female entities to counteract these stereotypes. This approach led to increased performance with the OntoNotes dataset, suggesting that balancing the representation of genders can significantly improve coreference resolution accuracy.

6 Conclusion

Coreference resolution has become a core component in natural language processing, enabling tasks like text summarization and machine translation. Despite considerable advancements, the field still faces several limitations that should be addressed. The OntoNotes dataset, the leading benchmark for coreference resolution, has outdated content and an imbalance between gender pronouns, impacting its ability to generalize to new domains. To improve coreference resolution, efforts should focus on expanding the dataset to include newer and more diverse sources while balancing gender representation to create a more equitable training dataset. These updates are crucial for developing fairer and more accurate models and would enhance the reliability of coreference resolution across a broader range of natural language processing applications.

References

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.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 28–36.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. 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.

Ana-Isabel Duron-Tejedor, Pascal Amsili, and Thierry Poibeau. 2023. How to evaluate coreference in literary texts?

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

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, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

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.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 395–400, Melbourne, Australia. Association for Computational Linguistics.

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

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, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2018. Using linguistic features to improve the generalization capability of neural coreference resolvers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Brussels, Belgium. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

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*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

A. Rahman and V. Ng. 2011. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research*, 40:469–521.

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, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Sanjay Subramanian and Dan Roth. 2019. Improving generalization in coreference resolution via adversarial training. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. 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), pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. 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.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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