Comprehensive Survey on Coreference Resolution

David Ko
University of California, Berkeley
Info 159, Natural Langauge Processing, Spring 2023

Abstract

[1] In this comprehensive survey, I aim to discuss the developments of coreference resolution throughout the late 1900s until current. The survey also provides insight into concepts that surround coreference resolution and discusses different coreference models throughout the last decade with an attempt to simplify these concepts. I provide brief descriptions and discuss the main significance of many different approaches to existing coreference resolution models.

1 Introduction

Coreference resolution serves as a method in which you identify all words in a text that refer to the same entity. It is an extremely valuable tool in natural language processing, as it assists with tasks such as sentiment analysis and information retrieval. The task of coreference resolution allows us to create a machine learning model that identifies what entity these pronouns refer to in order to increase the model's understanding of the information provided. That being said, researchers have had various approaches to conference resolution; many of the models that are used for such tasks are based on neural networks.

To preface the models discussed in the survey, it is important to understand model evaluation techniques. Coreference models are typically measured with an F-1 score. How the F1 score works is that it combines precision and recall into a formula: 2 * (precision * recall) / (precision + recall), where precision is the proportion of true positive results within predicted positive results, and recall is the proportion of true positive results within where precision is the fraction of true positive results among all predicted positive results, and recall is the fraction of predicted positive results among all actual true results (Goutte, 2005). The F-measure score ranges from 0 to 1, and the goal when evaluating models is to achieve a higher score.

2 Early Stages of Coreference Resolution

One of the early surveys discussing coreference resolution came from Bonnie Webber in 1978. In his journal, "Description Formation and Discourse Model Synthesis", he uses examples of text from children's books to illustrate the difficult nature of handling ambiguity in the task of coreference resolution. To address this, Webber discusses certain discourse models that could be used to accomplish such tasks, and what parameters are required for such models to work. Webber's work introduced the task of coreference resolution and lays the groundwork for how to make advancements for the future (Webber, 1978).

Moreover, in 1979, researcher Jerry R. Hobbs published a paper on Coherence and Coreference, where he assigned formal definitions to multiple coherence relations. Coherence relations are relationships among a group of words that link clauses or sentences. According to Hobb's survey, he compiles all the different terms that have been used informally to describe coherence relations, an example being "rhetorical predicates" (Grimes, 1975). Understanding conjunctive relations is crucial to the task of coreference resolution, as these relations are the foundation for how language is constructed. In his survey, he goes on to discuss how there is a system for making inferences in NLP, which he dissects into 4 categories: data, representation, operations, and control. Data is the information given to the model, representation is the format in which information is stored, operations refers to the steps needed to work on the data, and control refers to the order of the steps in order to maximize performance (Hobbs, 1979). Hobb's survey provides a structure for all coreference models after, as these 4 categories serve as the building blocks for all existing coreference resolution models.

[1] 2180 words

3 Mention Detection

Amongst other topics, an important topic to preface coreference resolution models is the concept of mention detection, a core component in coreference resolution that many models in this survey use. As other researchers have put it, mention detection could be thought of as an "important preprocessing step for annotation and interpretation" for coreference resolution"(Yu, 2020). To put simply, mention detection is the process in which the model extracts "all possible mentions from a given text" (Lata, 2022). Mention detection is worth mentioning in the task of coreference resolution, as it is widely used in many coreference resolution models. Extracting all pronouns and entities within a document allows researchers to focus more on other features in their own model.

4 Neural Models: The Entity-Grid Model

One notable example of a neural network model that uses mention detection is the "entity-grid" model introduced in 2005, where each document is presented as a grid. Each row has a mention of an entity and each column has the entity's features such as age or gender. This statistical model was based on Centering Theory (introduced by researcher Grosz) and built upon by researchers Regina Barzilay and Mirella Lapata (Grosz, 1995). To put it more concretely, the model analyzes what grammatical role a certain word is in a sentence. As a simple example, the model would determine whether the word "Microsoft" is a subject or a noun, and also identify what relationship "Microsoft" would have with other words in the sentence. Below is the table provided by Barzilay and Lapata's survey.

Table 1
A fragment of the entity grid. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	S	O	S	X	O	–	–	–	–	–	–	–	–	–	–	1
2	–	–	O	–	–	X	S	O	–	–	–	–	–	–	–	2
3	–	–	S	O	–	–	–	–	S	O	O	–	–	–	–	3
4	–	–	S	–	–	–	–	–	–	–	–	S	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	S	O	–	5
6	–	X	S	–	–	–	–	–	–	–	–	–	–	–	O	6

Figure 1: Barzilay and Lapata's Entity Entity Grid

Since Microsoft is a company, the entity grid model would classify it as type "organization". The model would then conduct "a separate clustering mechanism then coordinates the possibly contradictory pairwise classifications and constructs a partition on the set of NPs... [employing] Ng and Cardie's coreference resolution system. The system [would] decide whether two NPs are coreferent by exploiting a wealth of lexical, grammatical, semantic, and positional features." (Barzila, 2005) . Essentially, the model would pair the word "Microsoft" with all other words in the sentence and output a percentage of how co-referent "Microsoft" is with other pronouns in the given excerpt. Calculating the percentage of coreference builds off of another coreference resolution system by Ng and Cardie that, to put simply, uses machine learning models like decision trees and best-first clustering to make such calculations (Cardie, 2002). What makes the grid entity model stand out from its predecessors is the ability to account for "local coherence", which is a phenomenon in which a sentence is not grammatically correct by global standards, but is still understood by certain groups of people (Nguyen, 2017). In general, the entity-grid model serves as one of the foundational models for coreference resolution. Over time, researchers added different features to this entity-grid model, a notable one by Elsner and Charniak, who added "discourse prominence and named entity type and coreference features to distinguish between important and unimportant entities" and increased the model's F1 performance to 80 percent (Elsner, 2011). More recently in 2018, this entity-grid model was used for "written asynchronous conversations such as forums and emails (Joty, 2018).

5 End-To-End Coreference Model

The End-to-end Neural Coreference Resolution model was a major breakthrough in the task of coreference resolution. It predicts coreference links while skipping the sub-task of mention detection. What makes this coreference resolution model different from previous models is that it "directly considers all spans in a document as potential mentions and learns distributions over possible antecedents for each" (Lee, 2017). So, rather than identifying which words are potential mentions for a certain entity, the end-to-end system instead classifies every word as a mention, pairs groups of words together, and outputs a "mention score", where a higher score means there is a high chance that the group of words is co-referent. Throughout the process of the proposed end-to-end model, there are intermediary steps that are used to gain the "mention score" such as bidirectional LSTM, span head,

and span representation as seen in the table below.

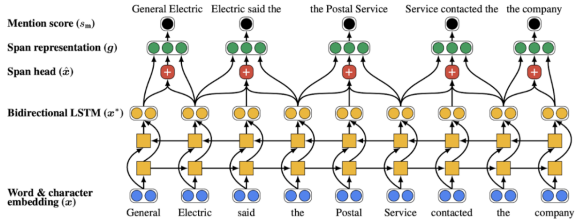

Figure 2: End-To-End model proposed by Lee, He, Lewis, and Zettlemoyer in 2017

This model would prove to outperform all previous models that existed before using metrics such as the F1 score.

6 C2F Coreference Model and Global Decisions

In 2018, authors of the same survey published a modified version of their already existing end-to-end model, only this one would be called "c2f-coref". This model served as another breakthrough in the field of coreference resolution in that it enhanced the model's performance by deriving more complex relationships between entities, also known as "higher-order inference"(Lee, 2018). The model accomplished this by forming a structure that allowed previous coreference values to impact future coreference decisions (Liu and Cambria, 2023).

Also building upon the end-to-end model, researchers also proposed a model that would learn coreference at the document level and take global decisions. At a high level, this model would iterate through a document and rank the pairs of each word together, then rank from 0-2 which pairs are co-referent, with 0 being no coreference, 1 being mention links, and 2 being co-referent. Testing this model with the CoNLL 2012 corpus data set would prove to demonstrate improvements over previous models (Miculicich, 2022).

7 Word-Level Coreference Resolution

Even though the end-to-end model has a core component of utilizing spans, researchers like Dobrovolskii (2021) introduced a word-level coreference resolution model, where each word was iterated over rather than each span as a means to reduce model complexity. The model, using a bi-linear function, would combine contextualized representations of each word to generate a coreference score. Overall, this reduced the model complexity to $O(n^2)$, whereas the initial end-to-end model had a complexity of $O(n^4)$ (Dobrovolskii, 2021).

8 Transformer-Based Models in Coreference Resolution

Following this survey, Another breakthrough discovery occurred when the BERT model was introduced in 2019 and used for the task of coreference resolution. This model utilized the same components of the end-to-end model but substituted the bidirectional LSTM encoder with a BERT encoder. In the BERT model, the documents given would be divided into two subcategories: the independent method and the overlap method (Joshi, 2019). The independent method would divide the words in the document into segments, but these segments had no overlap. As one can assume, the overlap method divided these words into segments with overlap; the end-to-end model proposed in 2017 utilized overlapping. These segments, also known as spans, would be inputted into the BERT encoder, which would then output the "mention score" (McCallum, 2004). Replacing the LSTM encoder in the end-to-end coreference model with the BERT encoder significantly improved the F1 score (80) of the model (Joshi, 2019). Following this model, researchers also modified the BERT model to create a SpanBERT model, which was a model with a more advanced encoder (Joshi, 2020). Interestingly enough, however, research done in 2020 by Xu and Choi demonstrated that the higher order inference concept introduced in the c2f model did not improve the SpanBERT model's performance with an F1 score of 80 on the CoNLL 2012 data set in English, implying that there is still more progress to be made (Xu, 2020).

9 Start-To-End Coreference Model

To improve on the end-to-end model's space complexity, the start-to-end model (2021) was introduced which "would only use the information on the start and end tokens of the span in order to calculate the "mention score" and antecedent score" (Liu and Cambria, 2023). Introduced by researcher Kirstain, this model would calculate the "mention score" by eliminating some intermediary steps within span-level representation – this is what led to improvement in run-time and space complexity. What span level representation did in the previous models was split the words into segments and vectorize the segments of words (Poerner, 2020). The

s2e model did the same, only instead of vectorizing every word within a span, it would vectorize the start and end words of the span and use that information to calculate the "mention score"(Kirstain, 2021).

A year later (2022), we see that the s2e model is improved upon by other researchers who incorporate the idea of centering theory (the theory also seen in the entity-grid model); simply put, the model gains the information from the start, and end token (like normal) but also includes the value of the center transition relationships between sentences. This model outperforms Kirstain's model with an F1 score of 80.9 and serves as an efficient model for "resolving pronouns in long documents" (Chai, 2022).

10 Faster Coreference Model

In 2022, researchers created a Python package called "fastcoref", which would provide an embedded model known as the "F-coref", a model still in development that aims to improve run-time by "a combination of distillation of a compact model from the LingMess model, and an efficient batching implementation called leftover batching" (Otmazgin, 2022). As we see throughout this comprehensive survey, the issue of run-time can always be improved upon in the task of coreference resolution.

11 Introducing Visualization in Coreference Resolution

Perhaps a model that stands out from every other model discussed and requires more research is the incorporation of visual dialog into the task of coreference resolution. With images added to the task coreference resolution, identifying coreference links becomes more challenging; however, the survey provides a "multi-layer transformer model" that incorporates the image into the task of coreference resolution. Testing this model on the Visdial v 1.0 data set, this promising model could serve as a baseline for other researchers to understand and improve upon (Li, 2021).

12 Conclusion

Upon analysis of the development of coreference resolution, we see that there is still ongoing research on improving the existing coreference models. The end-to-end model in 2017 would serve as one of the larger advancements in coreference resolution, as many models after would utilize its components.

References

M. Lapata Barzila. 2005. Modeling local coherence: An entity-based approach. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

Ng Cardie. 2002. Improving machine learning approaches to coreference resolution. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

Michael Strube Chai. 2022. Incorporating centering theory into neural coreference resolution. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies:2996–3002.

Dobrovolskii. 2021. Word-level coreference resolution. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing:7670–7675.

Eugene Charniak Elsner. 2011. Disentangling chat with local coherence models. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Eric Gaussier Goutte. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *Advances in Information Retrieval*, 3408.

Grimes. 1975. The thread of discourse. *The Hague: Mouton*.

Scott Weinstein Grosz, Aravind K. Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, Volume 21, Number 2, June 1995.

Hobbs. 1979. Coherence and coreference. *Cognitive Science Volume 3, Issue 1*.

Daniel Weld Joshi. 2019. Bert for coreference resolution: Baselines and analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP):5803–5808.

Omer Levy Joshi. 2020. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, Volume 8:64–77.

Dat Tien Nguyen Joty, Muhammad Tasnim Mohiuddin. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers):558–568.

Levy Kirstain, O. Ram. 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.

Kamlesh Dutta Lata, Pardeep Singh. 2022. Mention detection in coreference resolution: survey. *Applied Intelligence*, Volume 52:9816–9860.

Luke Zettlemoyer Lee. 2017. End-to-end neural coreference resolution. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing:188–197.

Luke Zettlemoyer Lee. 2018. Higher-order coreference resolution with coarse-to-fine inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers):687–692.

Marie-Francine Moens Li. 2021. Modeling coreference relations in visual dialog. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.

Anh Tuan Luu Liu, Rui Mao and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*.

Ben Wellner McCallum. 2004. Conditional models of identity uncertainty with application to noun coreference. *Advances in Neural Information Processing Systems 17 (NIPS 2004)*.

James Henderson Miculicich. 2022. Graph refinement for coreference resolution. *Findings of the Association for Computational Linguistics: ACL 2022*, page 2732–2742.

Shafiq Joty Nguyen. 2017. A neural local coherence model. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Yoav Goldberg Otmazgin, Arie Cattan. 2022. F-coref: Fast, accurate and easy to use coreference resolution. 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:48–56.

Hinrich Schütze Poerner, Ulli Waltinger. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. Corporate Technology Machine Intelligence (MIC-DE).

Webber. 1978. Description formation and discourse model synthesis. *American Journal of Computational Linguistics*.

Jinho D. Choi Xu. 2020. Revealing the myth of higher-order inference in coreference resolution. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP):8527–8533.

B.Bohnet Yu. 2020. Neural mention detection. *Proceedings of the Twelfth Language Resources and Evaluation Conference*.