An Annotated Bibliography

Wanyue Xiao

April 9, 2022

References

[1] K. Clark and C. D. Manning, "Improving coreference resolution by learning entity-level distributed representations," *arXiv preprint arXiv:1606.01323*, 2016. [Online]. Available: https://arxiv.org/pdf/1606.01323.pdf

The authors proposed a system aiming to solve the long-lasting problem caused by incorporating entity-level information in co-reference resolution. The system consists of four parts: a mention-pair encoder, a cluster-pair encoder, a cluster-ranking model, and a mention-ranking model. The mention-pair encoder is a feed-forward neural network that is capable of producing distributed representations of pairs of mentions (e.g., context embeddings) of pairs of co-reference clusters (e.g., additional mention features, document genre, distance features). By using the distributed representation as the input, the cluster-pair encoder produced a cluster-level representation through the pooling technique. Next, the mention-ranking model was used to score each mention whose parameters were used to initialize the cluster-ranking model. In the end, the cluster-ranking model was trained with a learning-to-search algorithm to calculate the score for each cluster. The system achieved F1 scores of 65.29 for English 63.66 for Chinese on CoNLL 2012 Shared Task dataset.

Even though the system only involved some basic multi-layer feed-forward neural network models, it was trained with an inspiring Max-Margin loss function, which punished the model based on different types of errors. However, one problem occurred. Since the model scored the pair of current mentions with each of its previous antecedent candidates, its complexity grows exponentially with the increase of sentence length.

[2] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "Spanbert: Improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64–77, 2020. [Online]. Available: https://arxiv.org/pdf/1907.10529.pdf

Joshi et al. proposed the SpanBERT, a pre-training model built on a well-tuned replica of BERT, to improve the performance of the existing pre-trained model with spans representation and prediction. The key idea is that, instead of randomly masking tokens in a sentence, the model masks contiguous random spans. Next, the model predicts the entire masked span's content by using a novel span-boundary representation, which is a fixed-length representation of a span gained from the representations of the span position and observed tokens near the boundaries.

The experimental result was promising. The SpanBERT yielded a better result in F1 (79.6%) on the OntoNotes benchmark than BERT-Large and Google BERT. In this paper, one can learn that the span masking scheme consistently outperforms random token masking. It also improves BERT performance by introducing the new concept of span boundary objective. Additionally, the result proved that the model has better performance without involving the next sentence prediction.

[3] M. Joshi, O. Levy, D. S. Weld, and L. Zettlemoyer, "Bert for coreference resolution: Baselines and analysis," *arXiv preprint arXiv:1908.09091*, 2019. [Online]. Available: https://arxiv.org/pdf/1908.09091.pdf

Since the Bert was invented in 2018, it has become one of the top candidate models for all sorts of Natural Language Processing tasks. Joshi et al. commenced to use the BERT model in coreference resolution in 2019, achieving great improvements in F1 on the OntoNotes (+3.9) and GAP benchmark (+11.5), respectively. Their research had also made a thorough performance comparison among ELMo, BERT-base, and BERT-large.

The proposed model was based on the higher-order coreference model published by Lee et al. (2018) and made two alterations to the model architecture. The first alteration was the encoder. The authors replaced the entire LSTM-based encoder with the BERT transformer and used the concatenated vector of the first and the last word pieces as span representations. The second difference was the measures of splitting documents into segments, given that the BERT model was trained on sequences of at most 512-word pieces. The authors experimented with two splitting variants (e.g., independent and overlap variants). The independent variant used each non-overlapping segment as an independent instance for BERT. Hence, the information capacity of each instance was subject to the content included in the current segment. However, this measure compromised the model's ability to represent words located at the segment edge. The latter split each document into overlapping segments by making a T-sized segment after every T/2 token, solving the problem caused by the independent splitting measurement.

The experiment showed that there were no patent differences between ELMo and BERT-base models. The BERT-large was superior to BERT-based in particular cases, including related entities resolution, pronoun resolution, and lexical matching (such as race track and track). Both the BERT-large and the BERT-base models showed their incompetency in dealing with mention paraphrasing resolution, conversation resolution, and some more complicated cases. The authors also pointed out BERT's inability on longer sequences. Therefore, coreference resolution on longer documents remained a potential challenge in the future.

[4] K. Lee, L. He, M. Lewis, and L. Zettlemoyer, "End-to-end neural coreference resolution," *arXiv preprint arXiv:1707.07045*, 2017. [Online]. Available: https://arxiv.org/pdf/1707.07045.pdf

Before 2017, the majority of co-reference resolution tasks took the results of a synaptic parser or hand-engineered mention decoder as input. In this research, Lee et al. introduced the first end-to-end coreference resolution model in this research, outperforming all existential works published before 2017 without adding additional features. The core concept of this model

was to optimize the marginal likelihood of antecedents from gold co-reference clusters. The experiment could be spliced into two steps. The first step was to compute embedding representations of spans for all potential mentions. Specifically, the procedure treated all spans in a document as potential mentions. Next, a BiLSTM model with a head attention mechanism was used to compute an embedding vector for each span based on various features (e.g., pre-trained and character embeddings, Start and End index features). The following step was to identify whether the span is a mention or not based on the result obtained from the BiLSTM model. The second step was to identify if a mention is co-referent with its previous antecedent by learning the distribution over potential antecedents for each span. The model automatically pruned spans with a lower mention score in this step. The model significantly outperforms other state-of-the-art models by 1.5 F1 on the OntoNotes benchmark dataset.
As the paper indicated, it was the first time an end-to-end model has been created for the co-reference resolution. The experiment could learn that pruning is beneficial for co-reference resolution. The limitations were also conspicuous. The authors also pointed out that this model has limited capability to make co-reference decisions requiring world knowledge or external sources of information. Besides, given that the model was based on word embeddings, the model tended to make wrong predictions for phrases with similar embeddings.

[5] A. Rahman and V. Ng, "Coreference resolution with world knowledge," *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 814–824, 2011. [Online]. Available: https://aclanthology.org/P11-1082.pdf

World knowledge is external knowledge extracted from three data sources, including a web-based encyclopedia, unannotated data, and coreference-annotated data (Rahman Ng, 2011). Prior to this paper, research proved that word knowledge was beneficial to the resolver consisting of a weak coreference model and a small feature set. However, this statement reminded unverified cases with strong coreference models with rich feature sets. To thoroughly examine the usefulness of rich

> feature sets with world knowledge on coreference resolution, the authors conducted an experiment by incorporating the world knowledge into one traditional mention-pair model and a cluster-ranking model trained with 39 morphosyntactic features. The utility of the new feature set has been evaluated by the two proposed models on the coreference-annotated documents using both the ACE and the OntoNotes annotation scheme.
> The results showed that each type of feature improves the performance of both baseline models, implying that all feature types (such as FrameNet, nouns pairs and verb pairs generated from coreference-annotated data, and appositives extracted from the unannotated data) are truly beneficial for the coreference resolution. Therefore, the experiment proved that incorporating different world knowledge into a coreference resolution model helps it discover potential coreference links and make coreference decisions. The best result is obtained when all these features sources are applied to the baseline model.

Evolving from the rule-based resolution to the statistical and machine learning-based resolution, the field of coreference resolution underwent a dramatic shift. During this process, countless algorithms and evaluation metrics sprung out, paving the path for the inception of deep learning. The introduction of deep learning in NLP further enables models to reduce their dependency on hand-crafted features.

After reading these five papers summarized above, one can summarize that the coreference resolution models should use morphological and semantic information and world knowledge resources to achieve better results. Besides, entity-level features, such as position, document genre, and speaker features, are also constructive for making coreference resolution decisions.

Lee et al. introduced the first End-to-End neural conference model in 2017, employing a span-ranking model for the task of coreference resolution. The most conspicuous strength of the initial model was its interpretability since it utilized the head-finding attention mechanism to display spans that could bring the most contribution. The weaknesses were also palpable. The issue of a surfeit of false positives might occur once the model overused word embeddings for similarity matching. Furthermore, the model did not incorporate world knowledge, which proved its beneficial contribution to coreference resolution. Joshi et al. further used the BERT models. They evaluated the performance of different types of SOTA models on various categories of coreference resolution, showing that BERT-large's capability in processing distinct related entities resolution, pronouns resolution, and lexical matching resolution. However, the exposure limit of this experiment implied future research in pertaining models capable of more effectively encoding document-level context using sparse representations.