Coreference Resolution with Entity Equalization

Ben Kantor
Tel Aviv University
benkantor@mail.tau.ac.il

Amir Globerson
Tel Aviv University
amir.globerson@gmail.com

Abstract

A key challenge in coreference resolution is to capture properties of entity clusters, and use those in the resolution process. Here we provide a simple and effective approach for achieving this, via an "Entity Equalization" mechanism. The Equalization approach represents each mention in a cluster via an approximation of the sum of all mentions in the cluster. We show how this can be done in a fully differentiable end-to-end manner, thus enabling high-order inferences in the resolution process. Our approach, which also employs BERT embeddings, results in new state-of-the-art results on the CoNLL-2012 coreference resolution task, improving average F1 by 3.6%.[1]

1 Introduction

Coreference resolution is the task of grouping mentions into entities. A key challenge in this task is that information about an entity is spread across multiple mentions. Thus, deciding whether to assign a given mention to a candidate entity could require entity-level information that needs to be aggregated from all mentions.

Most coreference resolution systems rely on pairwise scoring of entity mentions (Clark and Manning, 2016; Lee et al., 2017; Denis and Baldridge, 2008; Rahman and Ng, 2009; Durrett et al., 2013; Chang et al., 2013; Wiseman et al., 2016; Martschat and Strube, 2015). As such they are prone to missing global entity information.

The problem of entity-level representation (also referred to as high-order coreference models) has attracted considerable interest recently, with methods ranging from imitation learning (Clark and Manning, 2015) to iterative refinement (Lee et al., 2018). Specifically, Lee et al. (2018) tackled this problem by iteratively averaging the antecedents of each mention to create mention representations that are more "global" (i.e., reflect information about the entity to which the mention refers).

Here we propose an approach that provides an entity-level representation in a simple and intuitive manner, and also facilitates end-to-end optimization. Our "Entity Equalization" approach posits that each entity should be represented via the sum of its corresponding mention representations. It is not immediately obvious how to perform this equalization, which relies on the entity-to-mention mapping, but we provide a natural smoothed representation of this mapping, and demonstrate how to use it for equalization.

Now that each mention contains information about all its corresponding entities, we can use a standard pairwise scoring model, and this model will be able to use global entity-level information.

Similar to recent coreference models, our approach uses contextual embeddings as input mention representations. While previous approaches employed the ELMo model (Lee et al., 2018), we propose to use BERT embeddings (Devlin et al., 2018), motivated by the impressive empirical performance of BERT on other tasks. It is challenging to apply BERT to the coreference resolution setting because BERT is limited to a fixed sequence length which is shorter than most coreference resolution documents. We show that this can be done by using BERT in a fully convolutional manner. Our work is the first to use BERT for the task of coreference resolution, and we demonstrate that this results in significant improvement over current state-of-the-art.

In summary, our contributions are: a. A simple and intuitive approach for entity-level representation via the notion of Entity-Equalization. b. The first use of BERT embeddings in coreference-resolution. c. New state-of-the-art performance on

[1] Our code is publicly available at https://github.com/kkjawz/coref-ee

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 673–677
Florence, Italy, July 28 - August 2, 2019.

the CoNLL-2012 coreference resolution task, improving over previous F1 performance by 3.6%.

2 Background

Following Lee et al. (2017), we cast the coreference resolution task as finding a set of antecedent assignments y_i for each span i in the document. The set of possible values for each y_i is $\mathcal{Y}(i) = \{\epsilon, 1, \ldots, i-1\}$, a dummy antecedent ϵ and all preceding spans. Non-dummy antecedents represent coreference links between i and y_i, whereas ϵ indicates that the span is either not a mention, or is a first mention in a newly formed cluster. Whenever a new cluster is formed it receives a new index, and every mention with $y_i \neq \epsilon$ receives the index of its antecedents. Thus the process results in clusters of coreferent entities.

2.1 Baseline Model

We briefly describe the baseline model (Lee et al., 2018) which we will later augment with Entity-Equalization and BERT features. Let $s(i, j)$ denote a pairwise score between two spans i and j. Next, for each span i define the distribution $P(y_i)$ over antecedents:

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y \in \mathcal{Y}(i)} e^{s(i,y)}} .$$

The score is a function of the span representations defined as follows. For each span i let $\boldsymbol{g}_i \in \mathbb{R}^d$ denote its corresponding representation vector (see Lee et al. (2018) for more details about the model architecture). Lee et al. (2017) computes the antecedent score $s(i, j) = f_s(\boldsymbol{g}_i, \boldsymbol{g}_j)$ as a pairwise function of the span representations, i.e. not directly incorporating any information about the entities to which they might belong. Lee et al. (2018) improved upon this model by "refining" the span representations as follows. The expected antecedent representation $\boldsymbol{a}_i$ of each span i is computed by using the current antecedent distribution $P(y_i)$ as an attention mechanism:

$$\boldsymbol{a}_i = \sum_{y_i \in \mathcal{Y}(i)} P(y_i) \cdot \boldsymbol{g}_{y_i} \tag{1}$$

The current span representation $\boldsymbol{g}_i$ is then updated via interpolation with its expected antecedent representation $\boldsymbol{a}_i$:

$$\boldsymbol{g}'_i = \boldsymbol{f}_i \circ \boldsymbol{g}_i + (\mathbf{1} - \boldsymbol{f}_i) \circ \boldsymbol{a}_i \tag{2}$$

where $\boldsymbol{f}_i = f_f(\boldsymbol{g}_i, \boldsymbol{a}_i)$ is a learned gate vector. Thus, the refined representation $\boldsymbol{g}'_i$ is an element-wise weighted average of the current span representation and its direct antecedents. Using this representation the refined antecedent distribution can be calculated as follows:

$$P'(y_i) = \frac{e^{s(\boldsymbol{g}'_i, \boldsymbol{g}'_{y_i})}}{\sum_{y \in \mathcal{Y}(i)} e^{s(\boldsymbol{g}'_i, \boldsymbol{g}'_y)}}$$

3 Entity Equalization

The idea behind the refinement procedure in Lee et al. (2018) was to create features that are closer to cluster-level representations and hence more "global". This was partially achieved by considering not only the current span but also its antecedents. We would like take this idea one step further and create refined span representations that contain information about the entire cluster to which it belongs. One way to achieve this is by simply representing each mention via the sum of the mentions currently in its coreference cluster. Formally, let $C(i)$ be a coreference cluster (as defined by the antecedent distribution $P(y_i)$) such that $i \in C(i)$, and replace Equation 1 with:

$$\boldsymbol{a}_i = \sum_{j \in C(i)} \boldsymbol{g}_j \tag{3}$$

As a result each span will now contain information about all of its current coreference cluster, effectively equalizing the representations of different spans belonging to the same cluster.

However, note that it is not clear how to train such a procedure end-to-end because the clustering process is not differentiable. To overcome this problem, we use a differentiable relaxation of the clustering process (Le and Titov, 2017) and use the resulting soft clustering matrix to create a fully differentiable cluster representation. We call this refinement procedure Entity Equalization and provide a detailed description in the next section.

To illustrate the difference between Entity Equalization and antecedent averaging, consider the following example: *"[John] went to the park and [he] got tired. [John] decided to go back home."* Now assume that the model outputs the following antecedent distribution $P(y_i)$:

	John_1	he	John_2
John_1	1	0	0
he	1	0	0
John_2	1	0	0

there is only one coreference cluster induced by this antecedent matrix, $C = \{John_1, he, John_2\}$. A cluster representation for $John_2$ would be the sum of the representations of all three mentions. However, using antecedent averaging, the representation of *John*$_2$ will be a weighted average of the representations of *John*$_2$ and *John*$_1$, because only *John*$_1$ is an antecedent of *John*$_2$.

3.1 Implementing Equalization

In order to achieve differentiable cluster representations, we need a differentiable soft-clustering process. Le and Titov (2017) introduced such a relaxation given an antecedent distribution, based on the following observation: in a document containing m mentions there are m potential entities $E_1, ..., E_m$ where E_i has mention i as the first mention. Let $Q(i \in E_j)$ be the probability that mention i corresponds to entity E_j (that is, to the entity that has j as its first mention). Le and Titov (2017) showed that this probability can be computed recursively based on the antecedent distribution $P(y_i)$ as follows:

$$Q(i \in E_j) = \begin{cases} \sum_{k=j}^{i-1} P(y_i = k) \cdot Q(k \in E_j) & \text{if } j < i \\ P(y_i = \epsilon) & \text{if } j = i \\ 0 & \text{if } j > i \end{cases}$$

Note that this is a fully differentiable procedure that calculates the clustering distribution for each entity i.

The distribution $Q(i \in E_j)$ above leads to a simple differentiable implementation of the equalization operation in (3), as described next. In order to use entity representations for equalizing mention representations, we need a representation for each entity E_i at each time step t, so we won't represent a mention using mentions not yet encountered. We denote it by:

$$\boldsymbol{e}_i^{(t)} = \sum_{j=1}^{t} Q(j \in E_i) \cdot \boldsymbol{g}_j$$

Finally, an entity representation for each mention i is calculated using the entity distribution of mention i and the global entity representations:

$$\boldsymbol{a}_i = \sum_{j=1}^{i} Q(i \in E_j) \cdot \boldsymbol{e}_j^{(i)}$$

It can be seen that the above $\boldsymbol{a}_i$ will indeed lead to (3) when the distributions $P(y)$ are deterministic.

4 Using BERT Embeddings

Our coreference model relies on input representations for each input token. Lee et al. (2018) used the ELMo context-dependent embeddings for this purpose. Here we propose to use the more recent BERT embeddings (Devlin et al., 2018) instead, which have achieved state of the art performance on many natural language processing tasks. BERT is a bidirectional contextual language model based on the Transformer architecture (Vaswani et al., 2017). Using BERT for coreference resolution is not trivial: BERT can only run on sequences of fixed length which is determined in the pre-training process. In the pre-trained model published by Devlin et al. (2018), this limitation is 512 tokens, which is shorter than many of the documents in the CoNLL-2012 coreference resolution task. Even without considering the pre-training limitation, because the attention mechanism grows as the square of the sequence length, and because of the large number of parameters of the BERT model, running it on very large sequences is not feasible on most machines due to memory constraints.

In order to obtain BERT embeddings for sequences of unlimited length, we propose to use BERT in a convolutional mode as follows. Let D be a fixed window length. We obtain a representation for token i by applying BERT to the sequence of tokens from D to the left of i to D to the right of i. We then take the four last layers of the BERT model for token i and apply a learnable weighted averaging to those, similar to the process used in ELMo. The output of the network is taken as the representation of token i, and replaced the ELMo representation in the model of Section 3.1. We use $D = 64$, since using the maximum size of $D = 256$ is computationally intensive, and good results are already obtained with 64.[2]

5 Related Work

Several works have addressed the issue of entity-level representation (Culotta et al., 2007; Wick et al., 2009; Singh et al., 2011). In Wiseman et al. (2016) an RNN is used to model each entity. While this allows complex entity representations, the assignment of a mention to an RNN is a

[2]We note that BERT uses a special tokenization called WordPiece (Wu et al., 2016) which can split words to sub-words. When a word was split to several sub-words, only the embedding of the first sub-word was taken.

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Lee et al. (2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
+ BERT	**83.51**	82.8	83.16	**74.51**	74.14	74.32	71.93	70.6	71.26	76.25
– Second-order	82.61	83.48	83.04	73.56	75.44	74.49	71.6	**71.55**	71.57	76.37
+ EE (Ours)	82.63	**84.14**	**83.38**	73.31	**76.17**	**74.71**	**72.37**	71.14	**71.75**	**76.61**

Table 1: Results on the test set of the English CoNLL-2012 shared task. The average F1 of MUC, B^3 and $CEAF_{\phi_4}$ is the main evaluation metric.

hard decision, and as such cannot be optimized in an end-to-end manner. Clark and Manning (2015) use whole-entity representations as obtained from agglomerative clustering. But again the clustering operation in non-differentiable, requiring the use of imitation learning. In Lee et al. (2018), entity refinement is more restricted, as it is only obtained from the attention vector at each step. Thus, we believe that our model is the first to use entity-level representations that correspond directly to the inferred clusters, and are end-to-end differentiable.

Mention-entity mappings have been used in the context of optimizing coreference performance measures (Le and Titov, 2017; Clark and Manning, 2016). Here we show that these mappings can also be used for the resolution model itself. We note that we did not try to optimize for coreference measures as in Le and Titov (2017), and this is likely to further improve results.

6 Experiments

Data for all our experiments is taken from the English portion of the CoNLL-2012 coreference resolution tasks (Pradhan et al., 2012). Our experimental setup is very similar to Lee et al. (2018), and our code is built on theirs. We did not change the optimizer or any of the training hyperparameters. The following changes were made to the model:

- We used BERT word embeddings instead of ELMo as input to the LSTM (see Section 4).
- We replaced the span representation refinement mechanism with our Entity Equalization approach (see Section 3).

7 Results

Following Pradhan et al. (2012), we report precision, recall and F1 of the MUC, B^3 and $CEAF_{\phi_4}$ metrics, and average the F1 score of all three metrics to get the main evaluation metric used in the CoNLL-2012 coreference resolution task. We calculated the metrics using the official evaluation scripts of CoNLL-2012.

Results on the test set are shown in Table 1. Our baseline is the span-ranking model from Lee et al. (2018) with ELMo input features and second-order span representations, which achieves 73.0% Avg. F1. Replacing the ELMo features with BERT features achieves 76.25% average F1. Removing the second-order span-representations while using BERT features achieves 76.37% F1, achieving higher recall and lower precision on all evaluation metrics, while somewhat surprisingly being better overall. Replacing second-order span representations with Entity Equalization achieves 76.64% average F1, while also consistently achieving the highest F1 score on all three evaluation metrics. Our results set a new state of the art for coreference resolution, improving the previous state of the art by 3.6% average F1.

8 Conclusion

In this work we presented a new state-of-the-art coreference resolution system. Key to our approach is the idea that each mention should contain information about all its coreferring mentions. Here we implemented this idea by summing all mention representations within a cluster. In the future we plan to further enrich these representations by considering information from across the document. Furthermore, we can consider more structured representations of entities that reflect entity attributes and inter-entity relations.

Acknowledgments

This work was supported by a grant from the Israel Science Foundation.

References

Kai-Wei Chang, Rajhans Samdani, and Dan Roth. 2013. A constrained latent variable model for coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 601–612.

Kevin Clark and Christopher D Manning. 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, pages 1405–1415.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262.

Aron Culotta, Michael Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88.

Pascal Denis and Jason Baldridge. 2008. Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 660–669. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Greg Durrett, David Hall, and Dan Klein. 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, pages 114–124.

Phong Le and Ivan Titov. 2017. Optimizing differentiable relaxations of coreference evaluation metrics. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 390–399.

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.

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)*, volume 2, pages 687–692.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

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. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 968–977. Association for Computational Linguistics.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 793–803. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Michael Wick, Aron Culotta, Khashayar Rohanimanesh, and Andrew McCallum. 2009. An entity based model for coreference resolution. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 365–376. SIAM.

Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. Learning global features for coreference resolution. *arXiv preprint arXiv:1604.03035*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.