Parallel Data Helps Neural Entity Coreference Resolution

Anonymous ACL submission

Abstract

Coreference resolution is the task of finding expressions that refer to the same entity in a text. Coreference models are generally trained on monolingual annotated data but annotating coreference is expensive and challenging. Hardmeier et al. (2013) have shown that parallel data contains latent anaphoric knowledge, but it has not been explored in end-to-end neural models yet. In this paper, we propose a simple yet effective model to exploit coreference knowledge from parallel data. In addition to the conventional modules learning coreference from annotations, we introduce an unsupervised module to capture cross-lingual coreference knowledge. Our proposed cross-lingual model achieves consistent improvements, up to 1.74 percentage points, on the OntoNotes 5.0 English dataset using 9 different synthetic parallel datasets. These experimental results confirm that parallel data can provide additional coreference knowledge which is beneficial to coreference resolution tasks.

1 Introduction

Coreference resolution is the task of finding expressions, called mentions, that refer to the same entity in a text. Current neural coreference models are trained on monolingual annotated data, and their performance heavily relies on the amount of annotations (Lee et al., 2017, 2018; Joshi et al., 2019, 2020). Annotating such coreference information is challenging and expensive. Thus, annotation data is a bottleneck in neural coreference resolution.

Hardmeier et al. (2013) have explored parallel data in an unsupervised way and shown that parallel data has latent cross-lingual anaphoric knowledge. Figure 1 shows a coreference chain in an English–Chinese parallel sentence pair. “it”, “EMNLP 2022” in the English sentence, and “EMNLP 2022”, “它”(it) in the Chinese sentence are coreferential to each other. This cross-lingual coreference chain suggests that parallel multilingual data could be useful for training coreference models.

[*EMNLP 2022*] is coming , [*it*] is a top-tier NLP conference .
[*EMNLP 2022*] 即将召开 , [它] 是一个 NLP 领域的顶会 .

Figure 1: A coreference chain in an English–Chinese parallel sentence pair. Mentions in brackets are coreferential to each other.

Parallel data has been applied to project coreference annotations in non-neural coreference models (de Souza and Orăsan, 2011; Rahman and Ng, 2012; Martins, 2015; Grishina and Stede, 2015; Novák et al., 2017; Grishina and Stede, 2017). Instead, we focus on neural coreference models and ask the following main research question: *Can parallel data advance the performance of coreference resolution on English, where large amount of annotations are available?*

We propose a cross-lingual model which exploits cross-lingual coreference knowledge from parallel data. As there is no annotated cross-lingual coreference data, the model computes the coreference scores between target spans and source spans without any supervision. We conduct experiments on the most popular OntoNotes 5.0 English dataset (Pradhan et al., 2012). Given the English data, we generate 9 different synthetic parallel datasets with the help of pretrained neural machine translation (NMT) models. The target languages consist of Arabic, Catalan, Chinese, Dutch, French, German, Italian, Russian, and Spanish. The experimental results show that our cross-lingual models achieve consistent improvements, which confirms that parallel data helps neural entity coreference resolution.

2 Coreference Models

2.1 *neural-coref*

Most neural coreference models are variants of *neural-coref* (Lee et al., 2017), whose structure is illustrated in Figure 2 (a). It consists of a text encoder, a mention scorer, and a coreference scorer.

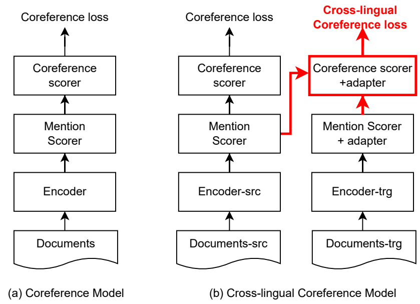

Figure 2: Overview of (a) the conventional monolingual coreference model and (b) our cross-lingual coreference model using synthetic parallel data. The main differences are marked in red. The red block is a cross-lingual coreference scorer which is expected to capture cross-lingual coreference knowledge.

The final coreference clusters are predicted based on the scores of these modules.

Given a document, the encoder first generates representations for each token. Then the model creates a list of spans, varying the span width.[1] Each span representation is the concatenation of 1) the first token representation, 2) the last token representation, 3) the span head representation, and 4) the feature vector, where the span head representation is learned by an attention mechanism (Bahdanau et al., 2015) and the feature vector encodes the size of the span. Then the mention scorer, a feed-forward neural network, assigns a score to each span. Afterwards, the coreference scorer computes how likely it is that a mention refers to each of the preceding mentions.

During training, given a span i, the model predicts a set of possible antecedents $\mathcal{Y} = \{\epsilon, 1, \ldots, i-1\}$, a dummy antecedent ϵ and preceding spans. The model generates a probability distribution $P(y_i)$ over antecedents for the span i, as shown in Equation 1 below. $s(i, j)$ denotes the coreference score between span pair i and j. The coreference loss is the marginal log-likelihood of the correct antecedents. During inference, the model first recognizes potential antecedents for each mention, then it predicts the final coreference clusters. More specifically, given a mention, the model considers the preceding mention with the highest coreference score as the antecedent.

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y' \in \mathcal{Y}(i)} e^{s(i,y')}} \tag{1}$$

[1]The number of generated spans is decided by hyper-parameters, i.e., the maximum width of a span, the ratio of entire span space, the maximum number of spans.

2.2 Cross-Lingual Model

We hypothesize that parallel data can provide additional coreference information which benefits learning coreference. As there is no supervision to the target-side and cross-lingual modelling, we attempt to transfer the source-side learned parameters to the target-side unsupervised modules by adding additional adapters, which has been shown efficient and effective (Houlsby et al., 2019). Therefore, we extend *neural-coref* by introducing a target-side encoder, adapters for target-side mention scorer, and cross-lingual coreference scorer, where each adapter is a one-layer feed-forward neural network with 500 hidden nodes. The overview of our cross-lingual model is shown in Figure 2 (b).

For the target-side, we can use a shared cross-lingual encoder or a target-side monolingual encoder. The coreference scorer computes coreference scores between target-side spans and source-side spans. This is the key component to learn cross-lingual coreference knowledge. The strategy we follow is the same as that in *neural-coref* during inference: Given a source mention, the target mention with the highest coreference score is considered as the corresponding cross-lingual antecedent.

Say the model has predicted a source mention list M_s: $\{m_{s_1}, m_{s_2}, \ldots, m_{s_m}\}$ and a target mention list M_t: $\{m_{t_1}, m_{t_2}, \ldots, m_{t_n}\}$. The model has also generated a two-dimensional coreference score matrix, where s_{ij} represents the coreference score between m_{s_i} and m_{t_j}. We denote $\mathcal{Y}(i)$ as the possible antecedent set of the source mention i. The cross-lingual coreference loss is defined in Equation 2, where $\hat{j} = \arg\max_{j \in \mathcal{Y}(i)} s_{ij}$ for a given i.

$$\mathcal{L}_x = \sum_{i=1}^{m} e^{-s_{i\hat{j}}} \tag{2}$$

During training, the model learns to minimize both the coreference loss and the cross-lingual coreference loss $\mathcal{L}_x$ with a ratio $1:1$. During inference, we only employ the source-side modules, which are trained with coreference supervision, to predict coreference clusters.

3 Experiments

3.1 Data

We experiment with the OntoNotes 5.0 English dataset. The number of documents for training, development, and test is 2802, 343, and 348, respectively. The data is originally from newswire,

Data	$F1_{mention}$	MUC			B^3			$CEAF_e$			$F1_{avg}$	Δ F1
		R	P	F1	R	P	F1	R	P	F1		
English	85.42	80.31	81.40	80.85	71.31	70.92	71.10	65.81	70.97	68.30	73.42	0
English–Arabic	86.13	81.73	81.80	81.77	72.91	71.77	72.34	67.85	71.53	69.64	74.58	1.16
English–Catalan	86.17	81.38	82.36	81.87	72.55	72.75	72.65	67.77	72.19	69.91	74.81	1.39
English–Chinese	86.02	81.16	82.43	81.78	71.91	72.74	72.32	66.96	72.17	69.47	74.53	1.11
English–Dutch	**86.29**	81.53	**82.84**	**82.18**	72.67	**73.31**	**72.99**	**68.36**	**72.41**	**70.33**	**75.16**	**1.74**
English–French	85.93	81.12	82.15	81.63	72.06	72.36	72.20	67.36	71.31	69.28	74.37	0.95
English–German	86.02	81.86	81.28	81.56	73.06	70.82	71.92	67.42	70.93	69.14	74.20	0.78
English–Italian	86.13	81.71	82.09	81.90	72.82	72.09	72.45	67.73	71.60	69.61	74.65	1.23
English–Russian	86.17	**82.38**	81.31	81.84	**73.75**	70.62	72.15	67.94	71.12	69.49	74.50	1.08
English–Spanish	86.21	81.72	81.88	81.80	72.62	71.88	72.25	67.88	71.11	69.45	74.50	1.08

Table 1: F1 scores on mention detection ($F1_{mention}$) and coreference resolution ($F1_{avg}$) of the monolingual model trained on English and cross-lingual models trained on 9 different synthetic parallel datasets. Δ F1 is the improvement over the monolingual model. Bold numbers are the best scores in each column. $F1_{avg}$ scores of all the cross-lingual models are statistically significant (t-test, $p < 0.05$).

magazines, broadcast news, broadcast conversations, web, conversational speech, and the Bible. It has been the benchmark dataset for coreference resolution since it is released. The annotation in OntoNotes covers both entities and events, but with a very restricted definition of events. Noun phrases, pronouns, and head of verb phrases are considered as potential mentions. Singleton clusters[2] are not annotated in OntoNotes.

Given the English data, we use open access pretrained NMT models released by Facebook and the Helsinki NLP group to generate synthetic parallel data (Wu et al., 2019; Ng et al., 2019; Tiedemann and Thottingal, 2020).

3.2 Experimental Settings

Our experiments are based on the code released by Xu and Choi (2020).[3] We keep the original settings and do not do hyper-parameter tuning. As Xu and Choi (2020) have shown that higher-order, cluster-level inference does not further boost the performance on coreference resolution given the powerful text encoders, we do not consider higher-order inference in our experiments. Even though the mention boundaries are provided in the data, we still let the model learn to detect mentions by itself. For evaluation, we follow previous studies and employ the CONLL-2012 official scorer (Pradhan et al., 2014, v8.01) to compute the F1 scores of three metrics (MUC(Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$(Luo, 2005)) and report the average F1 score.

The baseline model is trained on monolingual data while the cross-lingual models are trained on synthetic parallel data. Note that we use the trained monolingual model to initialize the source-side modules of the cross-lingual model. We mainly employ cross-lingual pretrained models, the XLM-R base model, as our encoders, but we also explore using two separate monolingual encoders. All the models are trained for 24 epochs with 2 different seeds, and the checkpoint that performs best on the development set is chosen for evaluation. We only report the average scores. Each model is trained on a single Nvidia V100 GPU with 32GB memory.

3.3 Experimental Results

Table 1 shows the detailed scores of each model on the OntoNotes 5.0 English test set. Compared to the baseline model, which is trained only on English data, our cross-lingual model trained on different synthetic parallel datasets achieves consistent and statistically significant (t-test, $p < 0.05$) improvements, varying from 0.78 to 1.74 percentage points. The model trained on English–Dutch achieves the best F1 performance on coreference resolution. The model trained on English–Russian achieves the best recall score on MUC and B^3.

It is interesting to see that the model trained on English–German achieves the least improvement, although German together with Dutch are closer to English compared to other languages. Meanwhile, the models trained on English–Arabic, English–Chinese, English–Russian obtain moderate improvements, even though Arabic, Chinese, and Russian are more different from English.

In addition to the results on coreference resolution, we also report the mention detection results, which are based on mention scores, i.e., the outputs of mention scorers. Models trained on parallel data are consistently superior to the monolingual model, and the model trained on English–Dutch gets the best F1 score of 86.29.

As Table 1 shows, our cross-lingual model,

[2] An entity cluster that only contains a single mention.
[3] https://github.com/lxucs/coref-hoi

which exploits parallel data, is superior to the model trained only on monolingual data. This confirms that parallel data can provide additional coreference knowledge to coreference models, which is beneficial to coreference modelling, even if the parallel data is synthetic and noisy.

4 Analysis

4.1 Unsupervised Cross-Lingual Coreference

To further explore what the unsupervised coreference resolution module can learn, we check the cross-lingual mention pairs predicted by the cross-lingual coreference scorer.

ParCorFull is an English–German parallel corpus annotated with coreference chains. We first feed the data to the model and let the model predict English–German mention pairs. We go through the these pairs quickly and find that some of these pairs are coreferential, some of these pairs are translation pairs, but most of them are irrelevant. As the coreference chains in English and German are not aligned, we cannot conduct quantitative evaluation.

Alternatively, we evaluate the ability of the model to capture cross-lingual coreference knowledge using a synthetic mention pair set: an English–English mention pair set. Now we have "aligned" coreference chains, and we can evaluate the mention pairs automatically. Specifically, we first train a cross-lingual model with English–English synthetic data, and we then feed the OntoNotes 5.0 English validation set to the model, both the source and target sides, to predict English–English mention pairs.

The model predicts 18,154 pairs in total, including 131 mention pairs that are the same mention, 1,257 mention pairs that are coreferential, and 758 mention pairs with the same surface. This indicates that the model is able to resolve some cross-lingual coreference. However, since the cross-lingual module is trained without any supervision, most of predicted mention pairs are not coreferential.

Table 2 shows some correctly predicted coreferential mention pairs, in English–English and English–German settings. We can tell that our cross-lingual models are not simply generating a pair of two identical mentions, but coreferential mentions as well, which is different from word alignment. These mention pairs support our hypothesis that the cross-lingual model can capture cross-lingual coreference knowledge.

Source Mentions(English)	Target Mentions(English/German)
Hong Kong	the city 's
It	the Supreme Court
he	28-jähriger Koch (28-Year-Old Chef)
The 19-year-old American gymnast	Simone Biles

Table 2: Examples of correct coreferential mention pairs predicted by the cross-lingual coreference model, in English–English, English–German settings.

4.2 Separate Monolingual Encoders

Multilingual pretrained models suffer from the curse of multilinguality which makes them less competitive as monolingual models. Thus, we replace the unified cross-lingual encoder (XLM-R) with two separate monolingual encoders. The baseline is a monolingual model trained with SpanBERT, and the cross-lingual model is trained with SpanBERT and BERT on source- and target-side text, on the English–German synthetic dataset.

Our experimental results show that models employing SpanBERT perform much better, which is consistent with previous findings by Joshi et al. (2020). The monolingual model achieves 77.26 F1 score on the OntoNotes 5.0 English test set. Our cross-lingual model obtains an even higher F1 score, 77.79, which is statistically significant (t-test, p=0.044). Thus, our proposed model is applicable to settings with separate monolingual encoders.

The improvement on SpanBERT is smaller than that on XLM-R. One explanation is that SpanBERT is already very powerful and parallel data provides less additional knowledge. Another explanation is that the target-side encoder, a BERT model, is much weaker than the SpanBERT, which makes it more difficult to learn the cross-lingual coreference.

5 Conclusions and Future Work

In this paper, we introduce a simple yet effective cross-lingual coreference resolution model to learn coreference from synthetic parallel data. Compared to models trained on monolingual data, our cross-lingual model achieves consistent improvements, varying from 0.78 to 1.74 percentage points, on the OntoNotes 5.0 English dataset, which confirms that parallel data benefits neural coreference resolution.

We have shown that the unsupervised cross-lingual coreference module can learn limited coreference knowledge. In future work, it would be interesting if we can provide the model some aligned cross-lingual coreference knowledge for supervision, to leverage parallel data better.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA.

José Guilherme Camargo de Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Anaphora Processing and Applications*, pages 59–69, Berlin, Heidelberg. Springer.

Yulia Grishina and Manfred Stede. 2015. Knowledge-lean projection of coreference chains across languages. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 14–22, Beijing, China. Association for Computational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multi-source annotation projection of coreference chains: assessing strategies and testing opportunities. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 41–50, Valencia, Spain. Association for Computational Linguistics.

Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, Washington, USA. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. 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.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

André F. T. Martins. 2015. Transferring coreference resolvers with posterior regularization. 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 1427–1437, Beijing, China. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Michal Novák, Anna Nedoluzhko, and Zdeněk Žabokrtský. 2017. Projection-based coreference resolution using deep syntax. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 56–64, Valencia, Spain. Association for Computational Linguistics.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. 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.

Altaf Rahman and Vincent Ng. 2012. Translation-based projection for multilingual coreference resolution. In

Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 720–730, Montréal, Canada. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, New Orleans, USA.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.