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. “ACL 2023”, “it” in the English sentence, and “ACL 2023”, “它” in the Chinese sentence are coreferential to each other. Compared to the two separate monolingual coreferential pairs: <ACL 2023, it>, <ACL 2023, it> in this parallel sentence pair. This cross-lingual coreference chain suggests that parallel multilingual data can provide extra coreferential knowledge compared to monolingual data which could be useful for training coreference models.

Parallel data has been applied to project coreference annotations in non-neural coreference models (de Souza and Orășan, 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 a relatively large amount of annotations are available?

We propose a cross-lingual model which exploits cross-lingual coreference knowledge from parallel data. Our model is based on the most popular neural coreference model (Lee et al., 2018), which consists of an encoder, a mention span scorer, and a coreference scorer. We extend these three modules, which are applied to the source-side data, with a target-side encoder and adapters for the mention span scorer and the coreference scorer, allowing these to resolve cross-lingual coreference. 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 popu-
lar 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 Related Work

Lee et al. (2017) first propose end-to-end neural coreference models (neural-coref) and achieve better performance on the OntoNotes English dataset compared to previous models. Most current neural coreference models are based on neural-coref and replace the statistic word embeddings used by Lee et al. (2017) with contextualized word embeddings from ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2020), etc. (Lee et al., 2018; Joshi et al., 2019, 2020).

Neural-coref only models the relation between pairs of mentions. Many studies propose to consider entity-level information while predicting clusters (Lee et al., 2018; Kantor and Globerson, 2019; Xu and Choi, 2020). However, Xu and Choi (2020) find that these models considering higher-order inference are not significantly better or even worse. Instead, the observed differences can be explained by the powerful performance of SpanBERT.

Because these models are expensive in terms of memory and time, especially when using higher-dimensional representations. Xia et al. (2020) and Toshniwal et al. (2020) propose models that only keep a limited number of entities in the memory, without much performance drop. Kirstain et al. (2021) introduce a start-to-end model where the model computes mention and antecedent scores only through bilinear functions of span boundary representations. To cope with the enormous number of spans, Dobrovolskii (2021) proposes a word-level coreference model, where the model first considers the coreference links between single words, and then reconstruct the word spans.

All these models are trained on monolingual coreference annotations. In this paper, we introduce a simple model building on the top of neural-coref, which exploits cross-lingual coreference from parallel data in an unsupervised way.

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.

3 Coreference Models

3.1 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. 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 \(\epsilon\) 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

\[s(i, j) = \text{score}(i, j)\]

\[P(y_i) = \frac{1}{\sum_{\epsilon \in \mathcal{Y}} s(i, \epsilon)}\]

\[\text{Loss} = -\sum_{i} \log P(y_{\text{true}})\]

\[^1\]The number of generated spans is decided by hyperparameters, i.e., the maximum width of a span, the ratio of entire span space, the maximum number of spans.
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.

\[
P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y_i' \in Y(i)} e^{s(i,y_i')}}
\]

### 3.2 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. This component serves to capture latent coreference information. During training, as source-side modules are shared across languages, source-side parameters are jointly updated when optimizing the cross-lingual coreference loss.

There is no specific range for antecedents in the cross-lingual setting. Thus, we introduce a restriction to target-side antecedents, where the cross-lingual antecedent’s position number in the target sentence should not surpass the source mention’s position number in the source sentence more than 50. This pruning can make the model more efficient and effective.

Say the model has predicted a source mention list \(M_s: \{m_{s1}, m_{s2}, \ldots, m_{sm}\}\) and a target mention list \(M_t: \{m_{t1}, m_{t2}, \ldots, m_{tm}\}\). The model has also generated a two-dimensional coreference score matrix, where \(s_{ij}\) represents the coreference score between \(m_{si}\) and \(m_{tj}\). We denote \(Y(i)\) as the possible antecedent set of the source mention \(i\). The cross-lingual coreference loss is defined in Equation 2, where \(j = \arg \max_{j \in Y(i)} s_{ij}\) for a given \(i\).²

\[
L_x = \sum_{i=1}^{m} e^{-s_{ij}}
\]

During training, the model learns to minimize both the coreference loss and the cross-lingual coreference loss \(L_x\) with a ratio 1 : 1. During inference, we only employ the source-side modules, which are trained with coreference supervision and latent cross-lingual coreference knowledge, to predict coreference clusters.

### 4 Experiments

Due to the page limit, we leave our experimental settings in Appendix A.

### 4.1 Data

We experiment with the OntoNotes 5.0 English dataset. The number of documents for training, development, and test is 2,802, 343, and 348, respectively. The data is originally from newswire, 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 clusters³ are not annotated in OntoNotes.

Given the English data, we use open access pre-trained 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).

The input to monolingual models are the English data and the inputs to cross-lingual models are these parallel data. They have the same amount of data entries. These parallel data have the same coreference annotations as the data fed into monolingual models, the only difference is that the English data is paired with its target translations, and there are no annotations in the translations at all.

²We assume that there should be at least one antecedent on the other side for each mention, either the translation of the mention or a translation of its antecedent. In practice, the quality of synthetic parallel data is not guaranteed which introduces noise. On the other hand, synthetic data may actually be more parallel than natural translations.

³An entity cluster that only contains a single mention.
4.2 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 is 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. Given the reported BLEU scores of the pre-trained NMT models, we find that the improvements do not correlate with the quality of generated translations.

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. We can tell that models with a higher mention detection F1 score do not always achieve higher coreference F1 score. There is no consistency across different language pairs, so the improvements are not merely from better mention detection performance, namely, memorizing mentions.

As Table 1 shows, our cross-lingual model, 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.

5 Analysis

5.1 Unsupervised Cross-Lingual Coreference

To explore whether the unsupervised module can capture cross-lingual coreference information, we check the cross-lingual mention pairs predicted by the cross-lingual coreference scorer.

ParCorFull (Lapshinova-Koltunski et al., 2018) 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 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

\*\*We also conduct preliminary experiments with parallel data from multiple language pairs, concatenating the parallel data of EN-DE, EN-ES, EN-IT, EN-NL, and EN-RU five language pairs. Our proposed cross-lingual model achieves better performance compared to using data from one single language pair, showing the capability of our model to work with multiple parallel data.\*\*
a cross-lingual model with English-English synthetic data, and we then feed the OntoNotes 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.

### 5.2 Separate Monolingual Encoders

Multilingual pretrained models suffer from the curse of multilinguality which makes them less competitive as monolingual models. Thus, we test the robustness of our model with separate encoders, i.e., 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 SpanBERT, which makes it harder to learn the cross-lingual coreference.

### 6 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.

### Limitations

We expect that our cross-lingual models have learnt some coreference knowledge on the target languages and we conduct experiments on some languages in zero-shot settings. However, we do not get consistent and significant improvements compared to monolingual models. This should be further investigated which potentially helps languages with few or no coreference annotations. Compared to monolingual models, our cross-lingual model improves the source-side coreference resolution but it requires almost two times GPU memory during training. Thus, this model architecture imposes restrictions on using larger pretrained models given limited resources.

### Acknowledgments

We thank all reviewers for their valuable and insightful comments. This project is mainly funded by the Swedish Research Council (grant 2017-930), under the project Neural Pronoun Models for Machine Translation. GT is also supported by Science Foundation of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (22YBB36). We also acknowledge the CSC – IT Center for Science Ltd., for computational resources, with the help of Jörg Tiedemann.
References


A Experimental Settings

Our experiments are based on the code released by Xu and Choi (2020).5 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)6 to compute the F1 scores of three metrics (MUC (Vilain et al., 1995), B3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005)) and report the average F1 score.

Regarding the pretrained NMT models, the English-German/French/Russian models are transformer.wmt19* and transformer.wmt14.en-fr from https://github.com/pytorch/fairseq/blob/main/examples/.

5https://github.com/lxucs/coref-hoi
6https://github.com/conll/reference-coreference-scorers
The NMT models for other translation directions are `opus-mt-en-*` or `opus-mt-*en` from [https://huggingface.co/Helsinki-NLP](https://huggingface.co/Helsinki-NLP).

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 randomly initialize the parameters of adapters. As we train a unified 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 in Section 5.2.

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.
ACL 2023 Responsible NLP Checklist

A  For every submission:

✓ A1. Did you describe the limitations of your work?
   in the Limitation section

□ A2. Did you discuss any potential risks of your work?
   Not applicable. Left blank.

✓ A3. Do the abstract and introduction summarize the paper’s main claims?
   Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
   Left blank.

B  ☒ Did you use or create scientific artifacts?
   Left blank.

□ B1. Did you cite the creators of artifacts you used?
   No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
   No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
   No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
   No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
   No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
   No response.

C  ✓ Did you run computational experiments?
   Section 3

✓ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
   Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.
C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Appendix A

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

No response.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

No response.

D. Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants’ demographic (e.g., country of residence)?

No response.

D3. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.