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. Mentions in brackets are coreferential to each other. This cross-lingual coreference chain suggests that parallel multilingual data could be useful for training coreference models.

Parallel data has been applied to project coreference annotations in non-neural coreference models (de Souza and Oršašan, 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.
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. 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 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 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.

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

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 cross-lingual encoder generates a list of spans, varying the span width. Then the cross-lingual 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 $M_s$: $\{m_{s1}, m_{s2}, \ldots, m_{sm}\}$ and a target mention list $M_t$: $\{m_{t1}, m_{t2}, \ldots, m_{tn}\}$. 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 $\mathcal{Y}(i)$ as the possible antecedent set of the source mention $i$. The cross-lingual coreference loss is defined in Equation 2, where $j = \arg \max_{j \in \mathcal{Y}(i)} s_{ij}$ for a given $i$.

$$\mathcal{L}_x = \sum_{i=1}^{m} e^{-s_{ij}}$$

During training, the model learns to minimize both the coreference loss and the cross-lingual coreference loss $\mathcal{L}_x$ 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,
museums, 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 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).

### 3.2 Experimental Settings

Our experiments are based on the code released by Xu and Choi (2020). 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,
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 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 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 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.

<table>
<thead>
<tr>
<th>Source Mentions(English)</th>
<th>Target Mentions(English/German)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hong Kong</td>
<td>the city’s</td>
</tr>
<tr>
<td>It</td>
<td>the Supreme Court</td>
</tr>
<tr>
<td>he</td>
<td>28-jähriger Koch (28-Year-Old Chef)</td>
</tr>
<tr>
<td>The 19-year-old American gymnast</td>
<td>Simone Biles</td>
</tr>
</tbody>
</table>

Table 2: Examples of correct coreferential mention pairs predicted by the cross-lingual coreference model.
References


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


