Multilingual Coreference Resolution with Harmonized Annotations

Ondřej Pražák and **Miloslav Konopík** and **Jakub Sido**

{ondfa,sidoj}@ntis.zcu.cz
konopik@kiv.zcu.cz

Department of Computer Science and Engineering,
NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Technická 8, 306 14 Plzeň
Czech Republic

Abstract

In this paper, we present coreference resolution experiments with a newly created multilingual corpus CorefUD (Nedoluzhko et al., 2021). We focus on the following languages: Czech, Russian, Polish, German, Spanish, and Catalan. In addition to monolingual experiments, we combine the training data in multilingual experiments and train two joined models – for Slavic languages and for all the languages together. We rely on an end-to-end deep learning model that we slightly adapted for the CorefUD corpus. Our results show that we can profit from harmonized annotations, and using joined models helps significantly for the languages with smaller training data.

1 Introduction

Coreference resolution is the task of finding language expressions that refer to the same real-world entity (antecedent) of a given text. Sometimes the corefering expressions can come from a single sentence. However, the expressions can be one or more sentences apart as well. It is necessary to see the whole document in some hard cases to judge whether two expressions are corefering adequately. This task can be divided into two subtasks. Identifying entity mentions, and grouping the mentions together according to the real-world entity they refer to. The task of coreference resolution is closely related to anaphora resolution – see (Sukthanker et al., 2020) to compare these two tasks.

One of the challenging difficulties of coreference resolution lay in linguistically complicated annotations. Some examples of linguistic complications are *split antecedents* (a mention refer to more than one real-word entities), *near identity* relations, *anaphoric* and *cataphoric* relations, etc (Nedoluzhko et al., 2021).

In this paper, we rely on a CorefUD corpus (Nedoluzhko et al., 2021) of harmonized annotations. This corpus enables us to battle linguistic complications since it simply presents corefering mentions in clusters. Since the corpus is compiled from 11 different corpora in 8 different languages, we can conduct multilingual experiments in this work. Our research goal is to evaluate whether the harmonized annotations open the possibility to obtain some performance gain by joint learning on multiple languages. We aim to compare the harmonized annotations with the original corpora as well.

2 Related Work

In agreement with many other NLP tasks, deep learning models prevail in the coreference resolution task. Lee et al. (2017) were first to introduce the end-to-end approach that many following papers adopted (they obtained an average of 67.2 of F1 score). The task experienced a big leap in performance with the introduction of large pre-trained models. BERT based models deliver the best results; Kantor and Globerson (2019) F1 76.6, and Joshi et al. (2019) F1 76.9. Joshi et al. (2020) came up with a new pretraining task focused on better span representations. Their model called SpanBERT brings additional improvements in the coreference resolution task (F1 79.6). Xu and Choi (2020) question the importance of modeling higher-order inference (HOI). They show that with advanced encoders, HOI has only a minor effect on the performance of models.

Research is significantly less evolved for other languages than English. However, some notable experiments were published in recent years. Recasens et al. (2010) describe multilingual experiments (for English, Catalan and Spanish, Dutch, German and Italian) similarly to our paper. However, the annotations were not harmonized as in our case. Therefore, they provide no experiments with

joint training.

Other cross-lingual experiments include Portuguese by learning from Spanish (Cruz et al., 2018); Spanish and Chinese relying on an English corpus (Kundu et al., 2018); and Basque based on an English corpus as well (Urbizu et al., 2019). All these approaches employ neural networks, and they transfer the model via cross-lingual word embeddings.

Treex CR (Novák, 2017) is a coreference resolution module in the Treex NLP framework[1]. It produces an advanced syntactic analysis with semantic features that the tool uses to find coreference relations – offers models for Czech and English. Other non-English experiments include Polish (Nitoń et al., 2018), Russian (Sboev et al., 2020), and German (Srivastava et al., 2018).

3 Dataset

For our experiments, we use the harmonized multilingual coreference dataset *CorefUD* (Nedoluzhko et al., 2021). The dataset was created by converting 17 existing datasets for 11 different languages into a common format on the top of universal syntactic annotations – Universal Dependencies. For coreference representation, a cluster-based approach was selected instead of the link-based approach. It is simpler and moreover the most frequently used dataset for English – OntoNotes adopt this approach too. In a cluster-based approach, every mention belongs to one cluster, represented by a unique ID. In a link-based approach, coreferences are expressed by the links between corefering mentions. In the link-based approach, coreference structures form a chain, but there are more complex coreference structures in some cases (Nedoluzhko et al., 2021). Datasets that use the link-based approach were converted to cluster-based at the cost of some information loss.

There are some notable differences between the datasets. One of the most prominent ones is the presence of singletons. Singletons are clusters that contain only one mention. Singletons are not present in any coreference relation. However, they are annotated as mentions in all datasets. Discontinued mentions represent another notable difference. A discontinuous mention consists of a sequence of words that is interrupted at least once with some words that do not belong to the mention. Such mentions can cause problems to models that assume

mentions to be continuous (such as our model).

Table 1 shows the statistics of the datasets including the above-mentioned differences. More detailed analysis can be found in Nedoluzhko et al. (2021).

4 Model

We use the basic end-to-end model from Xu and Choi (2020) with no higher-order inference (HOI), so it is the same model as it was proposed by Lee et al. (2017).

In the model, we start by modeling the probability $P(y_i|D)$ of a mention i corefering with the antecedent y_i in a document D. Since the model adopts the end-to-end approach, the mentions are identified together with the coreference links. We consider every continuous sequence of words as a mention i. Therefore, we work with $N = \frac{T(T+1)}{2}$ possible mentions, where T is the number of words in a document D.

We model the score of a mention i corefering with an antecedent y_i as a combination of two types scores $s_m(i)$ and $s_a(i, y_i)$. The s_m is a score of a sequence of words (spans) i being a mention. The $s_a(i, y_i)$ score is the score of a span y_i being an antecedent of span i. The scores are combined as a sum of $s_m(i)$, $s_m(y_i)$ and $s_a(i, y_i)$ as follows:

$$s(i, y_i) = \begin{cases} 0 & y_i = \epsilon \\ s_m(i) + s_m(y_i) + s_a(i, y_i) & y_i \neq \epsilon \end{cases},$$

(1)

where ϵ is an empty antecedent. Both scores $s_m(i)$ and $s_a(i, y_i)$ are estimated with a feed-forward neural network over the BERT-based encoder. In our experiments we use two encoders – multilingual BERT (Devlin et al., 2018) and Slavic BERT (Arkhipov et al., 2019).

The probability of an antecedent y_i can be expressed as the $softmax$ normalization over all possible antecedents $y' \in Y(i)$ for a mention i:

$$P(y_i|D) = \frac{\exp(s(i, y_i)}{\sum_{y' \in Y(i)} \exp(s(i, y'))}$$

(2)

The formula for all antecedents uses a product of multinomials of all individual antecedents:

$$P(y_1, ..., y_N|D) = \prod_{i=1}^{N} P(y_i|D)$$

(3)

CorefUD dataset	Total size						Division [% of words]		
	docs	sents	words	empty	singletons	discont.	train	dev	test
Catalan-AnCora	1550	16,678	488,379	6,377	74.6%	0%	78.6	10.7	10.8
Czech-PDT	3165	49,428	834,721	33,086	35.3%	3.1%	78.3	10.6	11.1
German-PotsdamCC	176	2,238	33,222	0	76.5%	6.3%	80.3	10.2	9.5
Polish-PCC	1828	35,874	538,891	864	82.6%	1.0%	80.1	10.0	9.9
Russian-RuCor	181	9,035	156,636	0	2.5%	0.5%	78.9	13.5	7.6
Spanish-AnCora	1635	17,662	517,258	8,111	73.4%	0%	80.9	9.5	9.6

Table 1: Basic dataset statistics including train/dev/test split of CorefUD data sets. Column *discont.* shows the percentage of discontinuous mentions. Taken from Nedoluzhko et al. (2021).

In the training phase, we maximize the marginal log-likelihood of all correct antecedents:

$$J(D) = \log \prod_{i=1}^{N} \sum_{\hat{y} \in Y(i) \cap \text{GOLD}(i)} P(\hat{y}) \qquad (4)$$

where GOLD(i) is the set of spans in the training data that are antecedents.

5 Experiments

First, we perform monolingual experiments with the model described in Section 4 on several largest datasets from *CorefUD*. Namely Czech, Russian, Polish, Spanish, Catalan, and German-PotsdamCC. The employed datasets are summarized in Table 1 along with some basic statistics. The datasets are split to train, dev, and test, but the test datasets are not publicly available. Therefore, we use the original dev datasets as test datasets, and we create new dev datasets by taking 10% of the training parts. We tune the hyperparameters and perform early stopping on the development parts.

As the next step, we perform multilingual experiments, where we train one model for all the Slavic languages (Czech, Russian, and Polish) and another model for all the languages (Czech, Russian, Polish, German, Spanish, and Catalan). Multilingual results in comparison with the monolingual ones are shown in Table 2.

The results in Table 2 are influenced by the presence of singletons in the datasets. Particularly, Polish, German, Spanish, and Catalan contain a large portion of singletons, which negatively impacts the results. Since our employed model cannot model singletons, we have removed them from the test datasets. We show the results on filtered datasets in Table 4. Singletons are not important for coreference resolution since they form no coreference

relation. However, they can be important in the training phase, if the model can use them for mention recognition.

We report the average F1 measure from the official scoring script[2] for the coreference resolution task in CoNLL. The metric is computed as the average of MUC, B^3 and $CLEAF_4$. Definition of these metric can be found in Pradhan et al. (2014). The F1 scores are reported with 95% confidence intervals measured from 5 runs. We use the same set of hyperparameters for all the languages and for all models. We train the models for approximately for 100k steps. We employ the Adam optimizer with the learning rate of 0.00001 for BERT layers and 0.0002 for other layers.

6 Discussion

From the result (See Table 3), we can see that joined multilingual models helps all the languages except for Czech – which does make sense because the Czech dataset is the largest one in the CorefUD corpus.

For the smallest datasets (Russian and German), multilingual models outperform the monolingual ones by a large margin (cca $2-6$ % F1). We can see that at least in small training datasets, using joined models definitely helps, and we can profit from the harmonized coreference annotations. It is also worth noticing that the confidence intervals for these datasets are significantly wider than for other datasets.

After the singleton filtering the employed model achieves good results for all languages tested.

For German, there are 6.3% of discontinuous entity mentions. The model iterates over all possible continuous spans. Therefore, it is not able to identify discontinuous mentions at all. For Geraman, the effect of singletons and discontinuous mentions

---

[2]https://github.com/conll/reference-coreference-scorers

	czech	russian	polish	german	spanish	catalan
Mono-mBERT	58.883 ± 0.204	62.665 ± 1.028	42.411 ± 0.303	39.958 ± 0.775	49.654 ± 0.118	47.962 ± 0.302
Mono-SlavicBert	**60.283 ± 0.013**	62.097 ± 1.153	43.234 ± 0.114	-	-	-
Slavic-mBERT	58.734 ± 0.198	**66.762 ± 0.495**	44.091 ± 0.413	-	-	-
Slavic-SlavicBERT	60.096 ± 0.103	64.414 ± 0.750	**44.943 ± 0.110**	-	-	-
Joined-mBERT	58.990 ± 0.304	65.243 ± 0.942	44.346 ± 0.342	**46.098 ± 0.641**	**51.192 ± 0.221**	**49.881 ± 0.126**

Table 2: Overall results of F1 averages obtained from the official scoring script.

	czech	russian	polish	german	spanish	catalan
Joined-mBERT	+0.107	+1.926	**+1.935**	**+6.140**	**+1.538**	**+1.919**
Slavic-mBERT	-0.149	**+3.445**	**+1.680**	-	-	-
Slavic-SlavicBERT	-0.187	**+2.317**	**+1.709**	-	-	-

Table 3: F1 gains of multilingual models over the same monolingual ones. Bold numbers indicate that the difference is bigger than the width of confidence interval. Table depicts absolute differences.

	czech	russian	polish	german	spanish	catalan
Mono-mBERT	64.383 ± 0.153	63.135 ± 0.521	60.247 ± 0.242	52.541 ± 1.183	67.88 ± 0.543	64.394 ± 0.685
Mono-SlavicBert	**65.835 ± 0.141**	63.453 ± 0.615	61.726 ± 0.395	-	-	-
Slavic-mBERT	63.980 ± 0.211	**66.794 ± 1.105**	61.584 ± 0.396	-	-	-
Slavic-SlavicBERT	65.443 ± 0.231	64.192 ± 0.475	**62.883 ± 0.068**	-	-	-
Joined-mBERT	64.176 ± 0.120	65.618 ± 0.314	61.959 ± 0.431	**61.439 ± 1.216**	**68.9825 ± 0.209**	**66.456 ± 0.092**

Table 4: Overall results of F1 averages obtained from the official scoring script after singleton removal.

combines.

7 Future Work

Currently, we experimented only on a subset of languages available in *CorefUD*. This was caused mainly by the resource-exhaustive training. We need 32GB graphic cards to capture long-enough contexts. We plan to experiment with the rest of the languages in the future.

Additionally, it would be interesting to explore the possibilities of zero-shot cross-lingual transfer in *CorefUD*, where we will not use the training data for the target language at all.

8 Conclusion

We performed pilot experiments to evaluate inter-language transferability of the models based on the *CorefUD* dataset. To do so, we used an end-to-end coreference resolution model based on BERT-like models. Our experiments show that learning from multiple languages in CorefUD annotation scheme helps significantly especially for languages with smaller training data (like Russian and German data in *CorefUD*).

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Andre Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2018. Exploring spanish corpora for portuguese coreference resolution. pages 290–295.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. 2018. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 395–400, Melbourne, Australia. 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.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. 2021. *Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages*. ÚFAL MFF UK, Praha, Czechia.

Bartłomiej Nitoń, Paweł Morawiecki, and Maciej Ogrodniczuk. 2018. Deep neural networks for coreference resolution for Polish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Michal Novák. 2017. Coreference resolution system not only for czech. In *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017), Martinské hole, Slovakia, September 22-26, 2017*, volume 1885 of *CEUR Workshop Proceedings*, pages 193–200. CEUR-WS.org.

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.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden. Association for Computational Linguistics.

A. Sboev, R. Rybka, and A. Gryaznov. 2020. Deep neural networks ensemble with word vector representation models to resolve coreference resolution in russian. In *Advanced Technologies in Robotics and Intelligent Systems*, pages 35–44, Cham. Springer International Publishing.

Ankit Srivastava, Sabine Weber, Peter Bourgonje, and Georg Rehm. 2018. Different german and english coreference resolution models for multi-domain content curation scenarios. In *Language Technologies for the Challenges of the Digital Age*, pages 48–61, Cham. Springer International Publishing.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Gorka Urbizu, Ander Soraluze, and Olatz Arregi. 2019. Deep cross-lingual coreference resolution for less-resourced languages: The case of Basque. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 35–41, Minneapolis, USA. Association for Computational Linguistics.

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.