Neural End-to-End Coreference Resolution using Morphological Information

Tuğba Pamay Arslan[*] and **Kutay Acar**[†] and **Gülşen Eryiğit**[*]
İTÜ NLP Research Group
Department of [AI&Data[*], Computer[†]] Engineering
Faculty of Computer&Informatics
Istanbul Technical University
[[*]pamay, [†]acarku18, [*]gulsen.cebiroglu]@itu.edu.tr

Abstract

In morphologically rich languages, words consist of morphemes containing deeper information in morphology, and thus such languages may necessitate the use of morpheme-level representations as well as word representations. This study introduces a neural multilingual end-to-end coreference resolution system by incorporating morphological information in transformer-based word embeddings on the baseline model. This proposed model participated in the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023). Including morphological information explicitly into the coreference resolution improves the performance, especially in morphologically rich languages (e.g., Catalan, Hungarian, and Turkish). The introduced model outperforms the baseline system by 2.57 percentage points on average by obtaining 59.53% CoNLL F-score.

1 Introduction

Coreference Resolution (CR) is the task of determining coreferential relations between mentions referring to the same real-world entity in a document. CR is one of the essential components of comprehending natural language and is investigated under the semantic level of natural language processing (NLP). An end-to-end CR system consists of two stages which are trained jointly: 1) Mention detection and 2) Coreference linking. In the first, all possibly referential mentions are extracted in a text. Then, the coreferential relations between the automatically predicted mentions are created during the linking stage. When the CR task crosses with the complex linguistic diversity of natural languages, it becomes even more difficult, and morphological richness is one of such diversity. Morphologically rich languages require considering sub-word units (or morphemes) which carry deeper information at the morphology level. Therefore, this study explores the impact of including morphology information explicitly in a neural multilingual end-to-end CR system. Moreover, even if CR is an actively studied topic for quite a long time, the multilingual studies are currently in the process of development. Most studies propose CR datasets in their own data format and report their performances in one language only. The lack of quality and standardized datasets makes building multilingual CR systems harder. CorefUD initiative fills this gap in the CR literature by proposing a universal coreference representation scheme which was built on top of the Universal Dependencies (Nivre et al., 2017, 2020; Grobol and Tyers, 2023) initiative.

In this paper, we propose a neural, multilingual, end-to-end CR model trained with the data convened in CorefUD v1.1 (Novák et al., 2022); we extend the baseline model (Pražák et al., 2021) by enhancing the transformer-based word embeddings with dense and sparse (i.e., one/multi-hot encoding) vector representations of morphological information (i.e., POS tags and morphological features). The CorefUD v1.1 contains 17 different datasets for twelve languages in a harmonized, universal scheme. The proposed CR model employing sparse vector representations of morphological information achieves 59.53% CoNLL score on the test set (average across all languages), which means a 2.57 percentage points improvement over the baseline. The results show that the impact of explicitly incorporated morphological information is particularly high in the CR performance of morphologically rich languages. The paper is structured as follows: Section 2 gives the related work, Section 3 introduces the proposed neural model in detail, Section 4 presents the experimental setup and results, and Section 5 gives the conclusion.

2 Related Work

Deep learning-based CR approaches conforming to the end-to-end fashion have begun to be studied extensively in the last few years. Lee et al. (2017) proposes the first end-to-end neural CR system,

Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, pages 34–40
December 7, 2023

which creates a base for later studies. This study is enhanced with transformer-based embeddings via BERT (Kantor and Globerson, 2019) and SpanBERT (Joshi et al., 2020) with the higher-order inference (HOI) mechanism on top which are featured by Lee et al. (2018). Moreover, Liu et al. (2020) proposes a neural CR system employing entity-based features which were obtained by graph neural networks. In parallel, Park et al. (2020) introduce BERT-SRU-based Pointer Networks with the integration of morpheme boundaries as features for Korean. There are many studies proposed in the previous multilingual CR shared task (CRAC 2022) (Žabokrtský et al., 2022). The winning team of the shared task was ÚFAL CorPipe with 2 of their 3 submissions being on the leaderboard. The best model, *straka*, is trained jointly on all training data in all languages, and provides 70.72% CoNLL F-score by primary metric. *ondfa* (Pražák et al., 2021) is a baseline-based model using pre-trained XLM-RobertaLarge (Conneau et al., 2019) and also containing mention-head prediction. With the power of the mention-head prediction component, the model ends up getting a higher head-match score. *K-Sap* (Saputa, 2022) is introduced for only Polish. In addition to neural CR systems, rule-based models (*berulasek*, *simple-rule-based*, *Moravec*) also exist in the CRAC 2022.

Available annotated CR datasets in the literature are in a lack of standardization, which makes the development of multilingual CR systems complicated. By means of the CorefUD scheme (Novák et al., 2022), a multilingual coreference dataset collection is established and the task is shaped into a more generalized form. In parallel, CRAC organizations encourage researchers to develop and submit their own systems utilizing CorefUD dataset under the shared representation. CRAC 2022 was organized with the CorefUD v1.0 (Nedoluzhko et al., 2022) containing 13 datasets for 10 languages. CRAC 23 is organized with CorefUD v1.1 release. This version consists of 17 different datasets for 12 languages. Recent contributions involve Hungarian with one dataset, Turkish with one dataset, and Norwegian with two datasets. These made Turkish and Norwegian to appear in the CorefUD collection for the first time.

3 The Proposed Model

The introduced model is a modified version of the baseline model provided in the CRAC 2023 Shared Task (Žabokrtský et al., 2023), with the span representations updated. The baseline model (Pražák et al., 2021) provides a multilingual, end-to-end neural CR system which is a re-implementation of an available study (Xu and Choi, 2020). Basically, the model learns the probability distribution of coreferential links in the training data by maximizing the marginalized log-likelihood of gold antecedents for each possible span. To rank automatically detected referential mentions and link them with their possible antecedents, the model estimates the combination of two types of scores: 1) individual mention score, and 2) paired antecedent score. Individual mention score represents the likelihood of a span being a referential mention. Antecedent score entails a span pair and ranks their possibility of being coreferent. Since spans are considered as a sequence of words, they are represented by their words' embeddings obtained from a transformer, i.e., BERT.

This study introduces an enhanced span representation by incorporating morphological information explicitly in addition to contextual embeddings obtained by BERT. Each span embedding consists of three main sub-parts[1]: the embeddings of its first and last tokens, and the head attended embeddings of all tokens, as formularized in Equation (1) in the baseline model. In the equation, s_i represents the i$^{\text{th}}$ span, and $e(s_i)$ indicates the embedding of the related span.

$$e(s_i) = e(s_{i_{first}}) \oplus e(s_{i_{last}}) \oplus e(s_{i_{head}}) \quad (1)$$

This study extends the first and last tokens' embeddings by incorporating one/multi-hot encoded morphological information explicitly. There are two types of morphological information utilized: universal part-of-speech (UPOS) and universal morphological features (Feats). The output sample of this procedure is shown for the first token's embedding in the Equation (2). The operation annotated by $\oplus$ is concatenation. Therefore the size of $e(s_i)$ is extended by the total unique number of universal POS tags and morphological features in the CorefUD collection. The same procedure is also applied to the last token's embedding.

$$\begin{aligned} e(s_{i_{first}}) = & e(s_{i_{first}}[form]) \oplus \\ & enc(s_{i_{first}}[upos]) \oplus \\ & enc(s_{i_{first}}[feats]) \end{aligned} \quad (2)$$

[1]The span representation also contains various metadata (speaker, genre, span width) embeddings, and also embedded distance between a span and its antecedent. These secondary information are not formularized in equations to make them more readable.

4 Experimental Setup & Results

This section introduces the performance of the proposed model and also intermediate results.

4.1 Experimental Setup

The model utilizes a transformer-based neural language model, BERT[2] (Devlin et al., 2019), which is multilingual, base, and case sensitive. The model is trained using the default hyper-parameters, except maximum segment length being 256 instead of 512[3]. The hardware used in training is Tesla v100 graphic card. We trained our model for 24 epochs and the number of documents in the joint training data is 9595. The gradient update frequency is 1 so the total gradient update count is 230280 accordingly. The total time for training is 25-30 hours on average across the experiments.

Universal POS tags and morphological features are employed as morphological information in this study. One should note that, in the dataset, multiple morphological features are collected under the same information unit, separated by pipe symbols. Therefore, while one-hot encoding is suitable for POS tags, morphological features require multi-hot encoding, e.g., UPOS="NOUN" and FEATS="Case=Nom|Number=Plur" have upos_one_hot = [00100000...] (supposing NOUN is the third UPOS) and feats_multihot = [0100100...] (supposing Number=Plur is the second and Case=Nom the fifth feature). The total numbers of unique POS tags and morphological features are 20 and 210, respectively. Since morphological information is inserted to both the first and last tokens' embeddings, the dimensionality of the span embedding has increased by 460 for the one/multi-hot encoding technique.

In the case of dense vector representation, embedding layers with the dimensionality of 5 are deployed for POS tags and morphological features separately. To preserve the dimensionality, multiple morphological features are averaged out. All experiments are operated on a joined multilingual training set containing training data from all CorefUD languages. The only official evaluation criterion for the shared task is CoNLL, calculated as the macro-average F1 values of MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998) and CEAFe (Pradhan et al., 2014) scores of the predictions. The primary score is calculated using the head-match. That is, if the heads of a gold-standard and predicted mentions correspond to the same token, they are considered as a match. For that reason, the predicted mentions are reduced to their head tokens during the evaluation, with the help of `MoveHead` utility of Udapi[4] (Popel et al., 2017).

4.2 Results & Discussion

Several experiments were conducted to maximize the performance while enhancing span representations with morphological information. The results are given in Table 1. All contributions are made to the baseline model.

System	CoNLL
BASELINE	58.99
+{U,F}$_{emb}$	60.75
+{U}$_{enc}$	61.27
+{U,F}$_{enc}$ (morphbase)	**61.35**

Table 1: The performances of the intermediate and the proposed models evaluated on the development sets (CoNLL score in %).

While the first two rows below the *BASELINE* indicate the intermediate systems, the final system, named *morphbase* hereinafter, is the proposed model which participates in the CRAC 2023 shared task by our team, `TrCR`. As intermediate investigations, in Table 1, the models are named with the employed linguistic information; U indicates the use of universal POS tags and F indicates the use of morphological features. The results show all models exploiting morphological information surpass the performance of the baseline model by varying amounts.

We try to use both dense embeddings and one hot encodings for our morphological information representations; The first attempt is to utilize dense representations of both universal POS tags and morphological features, named {U,F}$_{emb}$. This model provides 60.75% CoNLL score which is 1.76 percentage points higher than the baseline. The second model, {U}$_{enc}$ uses a one-hot encoded version of only universal POS tags, and surpasses our first intermediate model by 0.52 percentage points. The model submitted to the shared task, {U,F}$_{enc}$ (named morphbase) employs encoded versions of both universal POS tags and morphological features. The *morphbase* model gives the best per-

[2] https://huggingface.co/bert-base-multilingual-cased

[3] https://github.com/ondfa/coref-multiling/blob/master/experiments.conf

[4] https://github.com/udapi/udapi-python

System	**AVG**	ca_ancora	cs_pcedt	cs_pdt	de_parcor	de_potsdam	en_gum	en_parcor	es_ancora	fr_democrat	hu_szeged	lt_lcc	pl_pcc	ru_rer	hu_korkor	no_bokmaal	no_nynorsk	tr_itcc
BASELINE	58.99	65.60	65.72	65.66	**57.25**	56.07	**66.87**	**56.56**	67.00	57.22	58.96	66.96	64.17	**63.04**	48.38	58.44	68.78	16.15
morphbase	**61.35**	**68.85**	**67.97**	**66.05**	50.10	**63.51**	65.42	44.85	**69.98**	**59.77**	**59.19**	**72.74**	**65.61**	62.93	**53.25**	**71.02**	**69.15**	**32.63**
Diff	↑2.36	↑3.25	↑2.25	↑0.39	↓7.15	↑7.44	↓1.45	↓11.71	↑2.98	↑2.55	↑0.23	↑5.78	↑1.44	↓0.11	↑4.87	↑12.58	↑0.37	↑16.48

Table 2: Dev set results for individual languages in the primary metric (CoNLL).

System	**AVG**	ca_ancora	cs_pcedt	cs_pdt	de_parcor	de_potsdam	en_gum	en_parcor	es_ancora	fr_democrat	hu_szeged	lt_lcc	pl_pcc	ru_rer	hu_korkor	no_bokmaal	no_nynorsk	tr_itcc
BASELINE	56.96	65.26	**67.72**	**65.22**	**44.11**	57.13	**63.08**	35.19	66.93	**55.31**	55.32	63.57	66.08	**69.03**	40.71	65.10	65.78	22.75
morphbase	**59.53**	**68.23**	64.89	64.74	39.96	**64.87**	62.80	**40.81**	**69.01**	53.18	**56.41**	**64.08**	**67.88**	68.53	**52.91**	**68.17**	**66.35**	**39.22**
Diff	↑2.57	↑2.97	↓2.83	↓0.48	↓4.15	↑7.74	↓0.28	↑5.62	↑2.08	↓2.13	↑1.09	↑0.51	↑1.8	↓0.5	↑12.2	↑3.07	↑0.57	↑16.47

Table 3: Test set results for individual languages in the primary metric (CoNLL).

formance among all investigated models with a 61.35% CoNLL score which is 2.36 percentage points higher than the baseline. The one/multi-hot encoding technique performs better in capturing sparse tag combinations, which may be one reason why models using this technique are more successful than the model using dense representations.

There were 10 submissions in this year's shared task, CRAC 2023. The winner model of this year, *CorPipe* (Straka and Straková, 2022) preserved their positions on the leaderboard in the previous shared task and provides 74.90% CoNLL scores on average. We are ranked at 7th place (Table 5) on the macro-averaged score, indicated as *morphbase* in Table 5. On individual dataset scores, our highest rank is on Catalan (ca_ancora), which is the 5th place. Then it is followed by 6th place on Turkish (tr_itcc), Hungarian (hu_korkor), German (de_potsdamcc), and English (en_parcorfull) datasets. Tables 2 and 3 present the performance of the *morphbase* model, in all languages with the primary metric. The top row lists the name of datasets for each language. The row 'Diff' indicates the improvement of the *morphbase* over the baseline model. Enhanced span representation achieves 61.35% and 59.53% CoNLL performance on average, which are higher than 2.36 and 2.57 percentage points on development and test sets, respectively. Including morphological information explicitly into the baseline model improves the performance of the following morphologically rich languages: Catalan, Czech, Hungarian, Spanish, French, Lithuanian, Polish, Norwegian, and Turkish, however, in Czech and French, improvements are only observed on the development sets.

The highest performance increment is on Turkish (tr_itcc) by 16.48 percentage points on the development set and 16.47 percentage points on the test set. Since Turkish possesses prominently rich morphology, such enhancement is not utterly surprising. For Hungarian, a significant increase is obtained on hu_korkor dataset by 4.87 percentage points on the development set and 12.2 percentage points on the test set. It is followed by Norwegian which exhibits agglutinative characteristics on verbal suffixes and the baseline model is surpassed by 12.58% percentage points on the development set. The performance of Spanish is improved by 2.98 and 2.08 percentage points compared to baseline by obtaining 69.98% and 69.01% CoNLL scores on development and test sets, orderly. While there is an undeniable drop in performance for German (de_parcorfull) and English (en_parcorfull) datasets, there is no such drop in the remaining datasets of these languages. The small sizes of these datasets (for details, check Table 4) might be the reason for such results. Moreover, it can be observed that in the languages having large datasets such as Czech, Spanish, and Polish, the effect of morphological information integration seems not as prominent as in medium-sized datasets.

5 Conclusion

This study proposed a neural, end-to-end, multilingual CR model which is an improved version of the baseline model incorporating morphological information into transformer-based span embeddings. The results show that extending word representations with morphological information helps CR

systems on average but especially for languages with high morphological complexity and agglutinative characteristics (e.g., Catalan, Hungarian, Norwegian, and Turkish). The proposed model completed the CRAC 2023 shared task at 7th place on average. Besides, on individual dataset scores, our highest rank is on Catalan (ca_ancora), which is the 5th place. Then it is followed by 6th place on Turkish (tr_itcc), Hungarian (hu_korkor), German (de_potsdamcc), and English (en_parcorfull).

Limitations

The main limitation of this study is that the training is operated on the joined data including all languages and there are no language-specific adjustments to the model. Therefore, the model treats all data equally even if the language has specific characteristics which might be useful to detect referential mentions and/or make coreferential relations between them. It is considered that it might increase the performance of particular languages having distinctive linguistic characteristics.

The proposed model is trained by only default hyper-parameters with the baseline model, that is, no hyper-parameter tuning could be done due to the time and resource constraints. The introduced model may need another set of parameters to perform better. For example, due to the hardware constraints, the transformer's segment size is used as 256, which is smaller than the usual, 512. The effect of such a constraint is most likely to be negative since it is a limiting factor when it comes to capturing the longer context.

Beyond the listed limitations, this study showed the positive impact of the interaction between transformer-based word representations and morphological information on the CR, despite the increasing popularity of deep learning, and the power of transformers. In future work, firstly, we plan to conduct error-analysis on languages which our model, morphbase, provided lower performances than the baseline model. We also plan to apply the proposed idea to other SOTA end-to-end neural multilingual CR systems. Moreover, we will work on increasing the performance of other languages by representing language-specific features in the model.

Acknowledgements

This work is funded by the Scientific and Technological Research Council of Turkey (TUBITAK) with a TUBITAK 2515 (European Cooperation in Science and Technology - COST) project Grant No. 123E079. Computing resources used in this work were provided by the National Center for High Performance Computing of Turkey (UHeM) under grant number 4015042023 and also by İTÜ Artificial Intelligence and Data Science Application and Research Center.

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. Citeseer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Loïc Grobol and Francis Tyers, editors. 2023. *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*. Association for Computational Linguistics, Washington, D.C.

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.

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.

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.

Lu Liu, Zhenqiao Song, and Xiaoqing Zheng. 2020. Improving coreference resolution by leveraging entity-centric features with graph neural networks and second-order inference. *arXiv preprint arXiv:2009.04639*.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M.Antònia Martí, Marie Mikulová, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2022. Coreference in universal dependencies 1.1 (CorefUD 1.1). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cheoneum Park, Jamin Shin, Sungjoon Park, Joonho Lim, and Changki Lee. 2020. Fast end-to-end coreference resolution for Korean. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2610–2624, Online. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. 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.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123.

Karol Saputa. 2022. Coreference resolution for Polish: Improvements within the CRAC 2022 shared task. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 18–22, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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*.

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.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, and Daniel Zeman. 2023. Findings of the Second Shared Task on Multilingual Coreference Resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Appendix

Stat	ca_ancora	cs_pcedt	cs_pdt	de_parcorfull	de_potsdamcc	en_gum	en_parcorfull	es_ancora	fr_democrat	hu_szegedkoref	lt_lcc	pl_pcc	ru_rucor	hu_korkor	no_bokmaalnarc	no_nynorsknarc	tr_itcc
docs	1,298	2,312	3,165	19	176	195	19	1,356	126	400	100	1,828	181	94	346	394	24
sents	13,613	49,208	49,428	543	2,238	10,761	543	14,159	13,057	8,820	1,714	35,784	9,035	1,351	15,742	12,481	4,733
words	435,690	1,191,599	857,109	10,602	33,222	187,515	10,798	466,530	284,883	128,825	37,014	539,355	156,636	26,556	245,515	206,660	55,341
empty	6,377	35,844	22,389	0	0	99	0	8,112	0	4,857	0	470	0	1,988	0	0	0
train [%]	77.6	81.0	78.3	81.6	80.3	78.9	81.2	80.0	80.1	81.1	81.3	80.1	78.9	79.3	82.8	83.6	81.5
dev [%]	11.4	14.2	10.6	10.4	10.2	10.5	10.7	10.0	10.0	9.6	9.2	10.0	13.5	10.2	8.8	8.7	8.8
test [%]	11.0	4.9	11.2	8.1	9.5	10.6	8.1	10.0	10.0	9.4	9.6	9.9	7.6	10.5	8.4	7.7	9.7

Table 4: CorefUD v1.1 statistics. Last 4 datasets are newly introduced, the rest is presented in previous versions.

System	**AVG**	ca_ancora	cs_pcedt	cs_pdt	de_parcorfull	de_potsdamcc	en_gum	en_parcorfull	es_ancora	fr_democrat	hu_szegedkoref	lt_lcc	pl_pcc	ru_rucor	hu_korkor	no_bokmaalnaarc	no_nynorsknarc	tr_itcc
CorPipe	**74.90 (1)**	**82.59 (1)**	**79.33 (1)**	**79.20 (1)**	**72.12 (1)**	**71.09 (1)**	**76.57 (1)**	**69.86 (1)**	**83.39 (1)**	**69.82 (1)**	**69.47 (1)**	**75.87 (1)**	**79.54 (1)**	**82.46 (1)**	**68.92 (1)**	**78.74 (1)**	**78.77 (1)**	**55.63 (1)**
Anonymous	70.41 (2)	79.51 (2)	75.88 (2)	76.39 (2)	64.37 (3)	68.24 (5)	72.29 (2)	59.02 (3)	80.52 (2)	66.13 (2)	66.25 (2)	70.09 (2)	77.58 (2)	80.19 (2)	64.65 (3)	75.32 (2)	73.33 (2)	47.22 (2)
Ondfa	69.19 (3)	76.02 (3)	74.82 (3)	74.67 (3)	71.86 (2)	69.37 (3)	71.56 (3)	61.62 (2)	77.18 (3)	60.32 (4)	65.75 (4)	68.52 (3)	76.90 (3)	76.50 (4)	66.38 (2)	72.39 (4)	70.91 (4)	41.52 (4)
McGill	65.43 (4)	71.75 (4)	67.67 (7)	70.88 (4)	41.58 (7)	70.20 (2)	66.72 (4)	47.27 (4)	73.78 (4)	65.17 (3)	65.93 (3)	65.77 (6)	76.14 (4)	77.28 (3)	60.74 (4)	73.73 (3)	72.43 (3)	45.28 (3)
DeepBlueAI	62.29 (5)	67.55 (7)	70.38 (4)	69.93 (5)	48.81 (5)	63.90 (7)	63.58 (6)	43.33 (5)	69.52 (5)	55.69 (6)	63.14 (5)	66.75 (4)	73.11 (5)	74.41 (5)	54.38 (5)	69.86 (6)	68.53 (5)	36.14 (8)
DFKI-Adapt	61.86 (6)	68.21 (6)	68.72 (5)	67.34 (6)	52.52 (4)	69.28 (4)	65.11 (5)	36.87 (7)	69.19 (6)	58.96 (5)	58.56 (6)	66.01 (5)	67.98 (6)	72.48 (6)	51.53 (7)	70.05 (5)	68.21 (6)	40.67 (5)
Morphbase	59.53 (7)	68.23 (5)	64.89 (8)	64.74 (8)	39.96 (9)	64.87 (6)	62.80 (8)	40.81 (6)	69.01 (7)	53.18 (8)	56.41 (7)	64.08 (7)	67.88 (7)	68.53 (8)	52.91 (6)	68.17 (7)	66.35 (7)	39.22 (6)
BASELINE	56.96 (8)	65.26 (8)	67.72 (6)	65.22 (7)	44.11 (6)	57.13 (9)	63.08 (7)	35.19 (8)	66.93 (8)	55.31 (7)	55.32 (8)	63.57 (8)	66.08 (8)	69.03 (7)	40.71 (9)	65.10 (9)	65.78 (8)	22.75 (9)
DFKI-MPrompt	53.76 (9)	55.45 (9)	60.39 (9)	56.13 (9)	40.34 (8)	59.75 (8)	57.83 (9)	34.32 (9)	58.31 (9)	52.96 (9)	48.79 (9)	56.52 (9)	61.15 (9)	61.96 (9)	44.53 (8)	65.12 (8)	62.99 (9)	37.44 (7)

Table 5: This table presents the performances of the participated models in the CRAC 2023 Shared Task. These scores are the average CoNLL F-scores of the all languages. The numbers existing in parentheses indicate the rank of the team for each related language and dataset.