Multilingual Coreference Resolution: Adapt and Generate

System Description – CRAC 2023 Shared Task

Tatiana Anikina, Natalia Skachkova, Anna Mokhova

DFKI / Saarland Informatics Campus, Saarbrücken, Germany







- 1. Introduction
- 2. DFKI-Adapt
- 3. DFKI-MPrompt
- 4. Conclusion

Introduction

The CRAC-2023 shared task [Žabokrtskỳ et al., 2023] focused on **multilingual coreference resolution**, which includes (a) mention prediction and (b) mention clustering.

Goal: One model that can be applied to different languages. BUT:

- Languages may differ a lot in grammar, morphology, writing systems, etc.
- Annotated corpora are often not parallel and differ in size.
- Datasets may differ in how markables are defined.

We investigate:

- how to combine the existing data, features and fine-tuning approaches to improve the baseline results without larger models or additional data;
- if knowledge accumulated in large multilingual language models can be extracted using prompt fine-tuning to perform mention detection, and if this method can compete with the state-of-the-art one.

DFKI-Adapt

DFKI-Adapt

- \cdot is based on the official baseline
- combines joint pre-training, combined datasets for related languages, loss-based re-training, character embeddings and adapters

Note: Configurations are evaluated using the official development data.



The available datasets are quite different in terms of size and annotations. However, the task of identifying and clustering coreferent mentions is the same.

We pre-train the baseline model on **all datasets combined together** and then continue fine-tuning this model on each dataset. We restrict the number of the pre-training steps to 100,000.

Joined pre-training is **beneficial for all languages** and it brings an average improvement of **+4.8 F1** points compared to the CRAC baseline.



Combining the training sets of the **related languages**. E.g., for Spanish we combine it with other Romance languages that include Catalan and French.

Combined datasets are beneficial, although the benefits differ. The average improvement is **+2.29 F1** compared to the CRAC baseline.

Combining datasets is especially **helpful when we have a small number of annotated documents**.

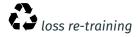


273 characters which include alphabet letters plus some additional symbols such as currency or copyright signs.

We run **bi-LSTM** to encode every token in the data.

In the coreference model we **concatenate the character embeddings** of the start and the end of each span with the corresponding BERT embeddings.

Character embeddings give a boost in performance compared to the CRAC baseline (+0.77 F1 points on average). The only two languages which show a decrease in performance are German and English.

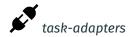


We store the loss associated with each document per epoch. At the end of each epoch we **sort the documents by their losses and take the 10%** of the most difficult ones (with the highest loss) for **additional training**.

This brings an average improvement of **+0.63% F1** points across all datasets. However, some datasets (e.g., Spanish and English-GUM) show worse performance.

The approach works better when there are less training data available.

DFKI-Adapt



We add **adapters** to the baseline model and then **pre-train them** for each dataset. Then we load the pre-trained adapters and train a new model for each dataset with the pre-trained adapter weights.

- task-adapters-frozen: we do not further train the adapters
- task-adapters-tuned: we continue training the adapters together with the rest of the model

With **task-adapters-tuned** the model underperforms by **-4.39 F1** points on average.

With **task-adapters-frozen** the results differ between the datasets. E.g., the model trained on German-Potsdam gains +8.21 F1 points compared to the baseline. However, English-GUM has a drop of -15.38 F1 points. The average improvement is **+0.67 F1**.

DFKI-Adapt: Evaluation

| | mbert- | mbert- | char- | joined- | combined- | loss- | task- | task- | DFKI- | CRAC- |
|----------------|--------|----------|-------|----------|-----------|----------|----------|----------|-------|----------|
| Dataset | joined | separate | embed | pre- | datasets | re- | adapters | adapters | Adapt | baseline |
| | | | | training | | training | frozen | tuned | | |
| ca_ancora | 68.97 | 65.06 | 66.56 | 68.72 | 66.29 | 65.59 | 66.19 | 61.99 | 68.34 | 65.60 |
| cs_pdt | 66.35 | 65.30 | 67.45 | 68.32 | 66.62 | 65.36 | 66.35 | 61.18 | 68.60 | 65.66 |
| en_gum | 65.80 | 52.01 | 54.05 | 62.41 | 35.25 | 51.38 | 51.49 | 47.54 | 69.63 | 66.87 |
| fr_democrat | 59.74 | 58.85 | 58.88 | 60.97 | 61.09 | 57.81 | 57.88 | 52.50 | 62.34 | 57.22 |
| de_potsdamcc | 65.77 | 58.92 | 55.16 | 62.03 | 67.12 | 59.77 | 64.28 | 60.27 | 69.29 | 56.07 |
| hu_szegedkoref | 59.78 | 59.98 | 59.53 | 62.29 | 60.42 | 60.13 | 57.39 | 53.70 | 62.60 | 58.96 |
| lt_lcc | 71.22 | 69.09 | 69.55 | 73.18 | 75.76 | 69.47 | 68.05 | 64.95 | 73.08 | 66.96 |
| no_bokmaal | 69.81 | 68.47 | 69.11 | 72.26 | 69.09 | 67.65 | 68.83 | 64.53 | 72.45 | 58.44 |
| pl_pcc | 65.41 | 63.64 | 65.32 | 66.38 | 66.21 | 63.74 | 64.30 | 59.44 | 65.89 | 64.17 |
| ru_rucor | 62.08 | 62.11 | 63.84 | 66.54 | 64.58 | 63.26 | 61.73 | 57.97 | 67.50 | 63.04 |
| es_ancora | 67.00 | 66.37 | 67.99 | 69.82 | 66.64 | 66.29 | 66.99 | 62.53 | 70.07 | 67.00 |
| tr_itcc | 31.66 | 31.35 | 17.98 | 30.80 | 33.88 | 23.28 | 20.68 | 6.91 | 37.80 | 16.15 |

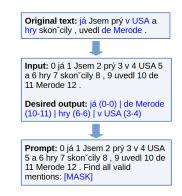
 Table 1: CoNLL F1 scores on the development data. The best performing setting is in bold

DFKI-Adapt takes the **4th** and the **6th places** among 8 (dev) and 10 (test) submissions, and shows an **improvement** over the CRAC baseline by **+9.07 F1** points on the **development data** and by **+4.9 F1** points on the **test data**.

DFKI-MPrompt

DFKI-MPrompt: Mention generation task

- mT5-base (580M) & mT5-large (1.2B) for mention generation
- Models' weights are frozen
- No demonstrations
- A prefix of 5 tunable embeddings (randomly initialized)
- A generated mention is correct, if both the mention string and its indices are correct
- The OpenPrompt library [Ding et al., 2021]



DFKI-MPrompt: Mention generation results

- The CRAC baseline was trained on all the training data in one go.
- The scores are not directly comparable, as the baseline omits singletons.
- In comparison to the CRAC baseline, the approach underperforms by **-7.82 F1** points.

| Data | # men | mT5-base | mT5-large | baseline |
|----------------|---------|----------|-----------|----------|
| avg | 108,006 | 6.09 | 66.83 | 74.65 |
| ca_ancora | 7,280 | 54.79 | 61.77 | 81.55 |
| cs_pcedt | 23,784 | 61.61 | 66.95 | 80.90 |
| cs_pdt | 20,955 | 57.24 | 62.46 | 78.76 |
| en_gum | 5,508 | 69.97 | 76.15 | 80.24 |
| en_parcorfull | 79 | 39.29 | 37.33 | 58.13 |
| fr_democrat | 7,032 | 68.87 | 75.88 | 78.63 |
| de_parcorfull | 93 | 52.81 | 55.14 | 53.89 |
| de_potsdamcc | 558 | 62.91 | 72.92 | 73.47 |
| hu_korkor | 448 | 55.32 | 61.04 | 70.85 |
| hu_szegedkoref | 1,458 | 58.10 | 63.36 | 68.23 |
| lt_lcc | 366 | 53.39 | 59.01 | 77.06 |
| no_bokmaal | 6,446 | 72.38 | 80.79 | 84.07 |
| no_nynorsk | 5,193 | 72.97 | 80.75 | 85.16 |
| pl_pcc | 18,857 | 64.95 | 72.09 | 77.49 |
| ru_rucor | 2,297 | 73.16 | 77.97 | 83.43 |
| es_ancora | 7,161 | 54.97 | 61.72 | 82.56 |
| tr_itcc | 491 | 65.75 | 70.70 | 54.65 |

Table 2: F1 scores for mentionidentification on development data

DFKI-MPrompt: Mention Generation error analysis

- Shorter mentions in shorter sentences are more likely to be generated correctly
- Among 21,133 wrong outputs, given development data,
 - 379 (1.79%) do not have brackets with indices
 - **752 (3.56%)** cannot be split, as they have a wrong delimiter, and the majority contain correct indices
 - **20,002 (94.65%)** consist of one mention and one index pair, and about a third of them have correct indices

Example: "Rodolfo Bay Wright, fundador de la aerolínea Spantax (1-9) |, fundador de la aerolínea Spantax (4-9)" We modify the CRAC baseline's architecture so that it performs only coreference resolution. Next, it is **re-trained on gold mentions** (including singletons) using all training data in one go, and **evaluated on the generated ones**.

DFKI-MPrompt: Coreference resolution results

| Model | dev | test |
|--|-------|-------|
| Official CRAC baseline | 58.99 | 56.96 |
| CRAC baseline trained on all the data (pred) | 61.08 | N/A |
| CRAC baseline trained on all the data (gold) | 77.81 | N/A |
| CRAC baseline trained on all the data (gen) | 57.21 | 53.76 |

Table 3: Average coreference resolution F1 scores for 17 datasets

The approach takes the **last place** out of 8 (dev) and 10 (test) submissions. Compared to the official CRAC baseline, it shows an average **decrease** in performance by **-1.78** on the **development data** and by **-3.20 F1** points on the **test data**.

Only on **4 out 17 test** sets the model performs better than the baseline, e.g., on Hungarian-KorKor with **+3.82 F1** and on Turkish with **+14.69 F1** points.

Conclusion

Conclusion

DFKI-Adapt

- Joined pre-training with further fine-tuning on the respective dataset is the most beneficial setting.
- The largest gains can be achieved with the combination of different settings.
- Pre-trained and frozen adapter weights can be helpful for many languages

DFKI-MPrompt

- $\cdot\,$ Demonstrated worse results than the baseline
- Could be improved applying a better template, more optimal hyperparameters and a larger model
- Could be tried out to deal with split antecedents and discontinuous mentions

DFKI-Adapt: tatiana.anikina@dfki.de DFKI-MPrompt: natalia.skachkova@dfki.de

Thank you for your attention!

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.-T., and Sun, M. (2021).

OpenPrompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998.*

Žabokrtský, Z., Konopík, M., Nedoluzhko, A., Novák, M., Ogrodniczuk, M., Popel, M., Pražák, O., Sido, J., and Zeman, D. (2023).

Findings of the second shared task on multilingual coreference resolution.

In Proceedings of the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023), pages 1–15.