Adapting Coreference Algorithms to German Fairy Tales

Schmidt, David
david.b.schmidt@uni-wuerzburg.de
Universität Würzburg, Germany

Krug, Markus
markus.krug@uni-wuerzburg.de
Universität Würzburg, Germany

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Germany

Introduction

Coreference Resolution has been posing an ongoing challenge to researchers for more than 50 years. It is the task of grouping mentions (concrete or abstract references, represented as textual spans) into clusters representing entities. The approaches for solving this problem have been manifold and range from rule-based approaches (Lee et al., 2013) over classical machine learning approaches (Rahman and Ng, 2009) to modern approaches based on Deep Learning (Lee et al., 2017; Joshi et al., 2020). Coreference Resolution can act as a ”glue” between information that is extracted on a local level (usually sentences) in order to obtain representations for an entire document or a collection of documents. This enables many interesting downstream applications such as the creation of character networks (Elson et al., 2010; Krug, 2020) or tracking of events involving central objects in textual media (such as the dagger in Emilia Galotti) (Hatzel and Biemann, 2021). The transfer of existing approaches to new domains or types of text usually comes with a drop in performance. In this work, we examine the performance of a rule-based and an end-to-end Deep Learning algorithm and their adaptability to the domain of German fairy tales. These experiments should provide insight into: a) the drop experienced from one kind of texts to another b) the reliability of state-of-the-art Deep Learning approaches compared to rule-based approaches for a change of texts and c) Capabilities for the adaptation to mitigate this natural drop in performance. This helps to estimate the required amount of manual work that is to be expected when transferring to a new kind of text, especially when the new type features a low number of annotated documents.

For this we use fragments of German novels provided from the DROC corpus (Krug et al., 2018) and as target domain we make use of annotated fairy tales by the Brothers Grimm. In the next section we present our data, followed by the coreference algorithms as well as the methods for the domain adaptation in more detail. We conclude the paper by presenting and discussing the results of our experiments and potential follow up work.

Related Work

There are several recent works that evaluate the performance of coreference resolution models when applied to a different domain than they have been trained on. Srivastava et al. (2018) examine the performance of several coreference resolution systems (rule-based, statistical and projection-based) on English and German out-of-domain data and find that the rule-based system is the best choice for their use cases. Han et al. (2021) train a coreference resolution model based on c2f (Lee et al., 2018) and SpanBERT (Joshi et al., 2020) on two different corpora, Ontonotes (Hovy et al., 2006) and their new corpus FantasyCoref. They then evaluate both models on FantasyCoref and find that the model trained on the same domain outperforms the other one. (Toshniwal et al., 2021) examine the generalization capabilities of coreference models by evaluating the performance of longdoc (Toshniwal et al., 2020) on out-of-domain data using several English datasets and find that models which have been trained on several datasets jointly perform better than those trained on a single dataset.

Data

The data sets for our experiments were the DROC corpus (Krug et al., 2018), comprising 90 fragments of German novels, and 46 tales from the seventh edition of the Children’s and Household Tales by the Brothers Grimm[1]. 40 of these fairy tales have been released together with networks of the important characters and their relations (Schmidt et al., 2021). The mentions and their coreference ids have been annotated by human annotators in both data sets. A notable difference between the two data sets is that DROC has gold information about direct speeches, speakers and addressees while for the fairy tales this information was generated automatically. All other information that the algorithms might use, e.g. POS tags or dependency parse trees, has been annotated automatically in both data sets.

Tab. 1: Average number of mentions per document in DROC and the fairy tales, and the ratios of names, noun phrases and pronouns

	Number of Mentions	Names	Noun Phrases	Pronouns
DROC	579	11.4%	20.1%	68.5%
Fairy Tales	296	5.2%	31.0%	63.8%

Table 1 shows some statistics about the mentions in the documents of DROC and the fairy tales. One can see that a document in DROC is on average about twice as long as a document of the fairy tales. Names are used a lot less often in fairy tales, while the usage of noun phrases increases and that of pronouns is comparable.

There is also an important difference regarding the entities that are referred to by the annotated mentions: In DROC, only human characters are annotated. In the fairy tales, animals and legendary beings (like giants) are also annotated because they are important (and sometimes the only) characters (e.g. the Wolf in *Little Red Riding Hood/Rotkäppchen* or the main characters in *Town Musicians of Bremen/Die Bremer Stadtmusikanten*).

A notable difference of both corpora to a lot of other corpora like OntoNotes (Hovy et al., 2006) and LitBank (Bamman et al., 2020) is how the mentions are annotated: OntoNotes annotates the maximal extent of a span (e.g. ’[eine kleine süße Dirne]’) while DROC and the fairy tales only annotate the heads (’eine kleine süße [Dirne]’).

For the experiments, DROC was split (a fix split) into a training set and a test set in a ratio of 80% to 20%: 72 documents for training and 18 for evaluation. The fairy tales were evaluated via five-fold cross validation.

Method

In order to assess the capabilities of domain adaptation from German novels to German fairy tales, we made use of a rule-based coreference resolution system and a model based on neural networks. We briefly present both methods followed by the way of adaptation.

The rule-based approach we use is an adaptation of the sieves algorithm by (Lee et al., 2013) to German (Krug et al., 2015). It partitions its rules into so-called sieves, which are ordered by the precision of their rules and applied one after the other to a document. This enables the rules to make use of the decisions of previously applied rules. Most rules use string matching to resolve names and noun phrases. Among the first sieves is also one that uses information about direct speeches to resolve all first person pronouns to the speaker and all second person pronouns to the addressee, and another that resolves relative and reflexive pronouns based on dependency parse trees. All other pronouns are resolved at the end since they do not possess much helpful information and can only be resolved unreliably (compared to a lot of names and noun phrases).

As Deep Learning architecture, we decided to use c2f (Lee et al., 2018)[2]. It is based on e2e (Lee et al., 2017), which was the first end-to-end neural network-based architecture for coreference resolution. e2e begins by building span representations for all spans up to a pre-defined length. All span representations are scored by a feed-forward neural network and only the top-scoring spans are kept (usually about 40% of all spans), all others are discarded. Each remaining span representation is then paired with a pre-defined number of potential antecedents and the pairs are scored by another feed-forward neural network (aside from the span representations the feed forward neural network (FFNN) also receives additional information like the domain of the document and whether both spans have the same speaker). Since not all span representations actually have an antecedent they are also paired with a dummy antecedent. For each span, the highest-scoring partner is picked as antecedent (unless the dummy antecedent was the highest-scoring partner, then the span does not have an antecedent). The architecture of c2f extends e2e in several ways, the two most important are the following: First, it uses a coarse bilinear scoring function, which is easier to compute, to prune the span representations before they are scored by the FFNN. Secondly, it scores the span representation pairs more than once and refines the span representations based on the scoring results between the iterations.

The adaptation of both approaches was done as follows:

Rule-based approach: Most rules in the sieves algorithm previously skipped family relation words and did not try to resolve them to an antecedent. In fairy tales, family relation words are most often unique (e.g. there is only one character called mother and one called father), so family relation words now are resolved to an antecedent if they are preceded by a definite article. Reflexive pronouns are resolved with the help of a dependency parse tree, which was not possible for several reflexive pronouns in the fairy tales. These are now resolved together with most other pronouns (lacking information about gender and number, hardly any antecedent can be ruled out, so they are usually resolved to the first that is checked). In addition to that, there were a few small changes that were done as the result of an error analysis on the fairy tales but are not motivated by the domain (they would probably also slightly improve the results on DROC).

Deep Learning approach: We trained and evaluated three variants of the c2f algorithm: c2f trained on DROC (c2f D) for 75000 steps, c2f trained on the fairy tales (c2f FT) for about 50000 steps and c2f pre-trained on DROC for 75000 steps and fine-tuned on the fairy tales for an additional 20000 steps (c2f D+FT)[3]. The maximum number of words a span may contain was reduced from 30 to 4 since the mentions annotated in DROC and the fairy tales are significantly shorter than the mentions in most English corpora. As language model we used a German ELMo model trained on Wikipedia (May, 2019)[4].

Results and Discussion

Table 2 displays the results of the sieves algorithm (old and adapted version) and c2f (trained on DROC, the fairy tales or both) on DROC (first two rows) and the fairy tales. As metrics we use MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_E$ (Luo, 2005) and as well as LEA (Moosavi and Strube, 2016).

	MUC			B^3			$CEAF_E$			LEA		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Sieves old	85.2	77.7	81.2	68.8	43.5	52.3	32.3	56.0	39.6	57.5	37.1	44.3
c2f	89.8	91.2	90.5	48.5	59.8	52.7	31.0	46.2	36.6	45.8	57.7	50.1
Sieves old	86.1	77.5	81.5	**67.5**	44.7	52.9	28.2	49.9	34.9	55.6	38.0	44.2
Sieves	87.3	80.9	84.0	67.1	47.6	55.0	33.9	**54.5**	**40.9**	57.8	41.3	47.4
c2f D	88.6	90.2	89.4	56.2	46.6	49.5	47.8	26.2	32.4	53.9	42.7	46.3
c2f FT	90.4	90.7	90.5	59.0	56.4	57.0	52.2	30.4	37.1	57.1	53.3	54.4
c2f D+FT	**91.2**	**91.6**	**91.4**	64.2	**61.6**	**62.5**	**56.5**	34.1	**40.9**	**62.4**	**58.7**	**60.1**

Tab. 2: Results (Precision, Recall and F1 score) of the rule-based sieves algorithm and the coarse-to-fine algorithm on DROC (first two rows) and on the fairy tales (all other rows). c2f was either trained on DROC (D), the fairy tales (FT) or both. Note that that the sieves algorithm uses gold mentions while c2f does not. The best results on the fairy tales are marked in bold

The results show multiple interesting aspects:

1. The results of the non-adapted rule-based system appear to be rather stable between domains (with the exception of $CEAF_E$, which drops surprisingly), while the Deep Learning model has a significant drop when it is evaluated on a domain it has not been trained on (e.g. 46.3% LEA F1 vs 50.1% on DROC and 54.4% on the fairy tales). One reason for this drop is that it did not recognize a lot of references to animals as mentions since animals are not annotated in DROC. The sieves algorithm did not suffer from this because it uses gold mentions.
2. Training and evaluating c2f on fairy tales yields a performance better than doing both on DROC.
3. Domain Adaptation of a Deep Learning model is pretty easy by just training on different data first and then fine-tuning on in-domain data.
4. While requiring more effort, the adaptation of the rule-based system also yields significant improvements on the domain of the fairy tales. And so far only very rudimentary changes have been made and it is to be expected to further improve the results with more in-depth analysis.

5. On DROC, c2f shows overall better performance than the sieves algorithm. On the fairy tales, the performance gap (when c2f is trained only on fairy tales) is even larger. This is even though in both cases c2f is at a disadvantage since the sieves algorithm uses gold mentions and c2f does not.
6. The version of c2f that is pre-trained on DROC and fine-tuned on the fairy tales (c2f D+FT) outperforms all other systems (by over 5% when measured with LEA or B^3). This (unsurprisingly) shows that the neural network profits from larger data sets.

We have shown that domain adaptation of both, a rule-based system and a Deep Learning based system, yields substantial improvements to coreference resolution on a target domain (in our case fairy tales). The evaluation also opens possibilities for further combination of the results of the rule-based system and the Deep Learning based system, which we leave for further work.

Fußnoten

1. We use only 46 tales because the other documents have not been annotated yet. The data used can be found at https://gitlab.informatik.uni-wuerzburg.de/kallimachos/coref-adaptation
2. Most other NN architectures require even more memory and time to train. We also spent some effort experimenting with a more memory-efficient architecture (Kirstain et al., 2021) but could not get any substantial results.
3. Since we do not have any GPUs with sufficient memory capacity (more than 24 GB) c2f D was trained on CPUs, which took about one week. Training on the fairy tales was done on an RTX3090 in a few hours.
4. https://github.com/t-systems-on-site-services-gmbh/german-elmo-model

Bibliography

Bagga, A. and Baldwin, B. (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. Granada.

Bamman, D., Lewke, O., and Mansoor, A. (2020). An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference* , pages 44–54, Marseille, France. European Language Resources Association.

Elson, D. K., Dames, N., and McKeown, K. R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics* , pages 138–147. Association for Computational Linguistics.

Han, S., Seo, S., Kang, M., Kim, J., Choi, N., Song, M., and Choi, J. D. (2021). FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference* , pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hatzel, H. O. and Biemann, C. (2021). Towards layered events and schema representations in long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop* , pages 32–39.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers* , pages 57–60, New York City, USA. Association for Computational Linguistics.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* , 8:64–77.

Kirstain, Y., Ram, O., and Levy, O. (2021). Coreference resolution without span representations. *arXiv preprint arXiv:2101.00434* .

Krug, M. (2020). *Techniques for the Automatic Extraction of Character Networks in German Historic Novels* . Doctoral thesis, Universität Würzburg.

Krug, M., Puppe, F., Jannidis, F., Reger, I., Weimer, L., and Macharowsky, L. (2015). Rule based coreference resolution in german historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* , pages 98–104.

Krug, M., Puppe, F., Reger, I., Weimer, L., Macharowsky, L., Feldhaus, S., and Jannidis, F. (2018). Description of a corpus of character references in german novels - DROC [Deutsches ROman Corpus]. In *DARIAH-DE Working Papers* . DARIAH-DE.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., and Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational linguistics* , 39(4):885–916.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* , pages 188–197.

Lee, K., He, L., and Zettlemoyer, L. (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.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* , pages 25–32. Association for Computational Linguistics.

May, P. (2019). German ELMo Model.

Moosavi, N. S. and Strube, M. (2016). Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* , volume 1, pages 632–642.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* , pages 968–977. Association for Computational Linguistics.

Schmidt, D., Zehe, A., Lorenzen, J., Sergel, L., Düker, S., Krug, M., and Puppe, F. (2021). The FairyNet corpus - character networks for German fairy tales. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* , pages 49–56, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Srivastava, A., Weber, S., Bourgonje, P., and Rehm, G. (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.

Toshniwal, S., Wiseman, S., Ettinger, A., Livescu, K., and Gimpel, K. (2020). Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP) , pages 8519–8526, Online. Association for Computational Linguistics.

Toshniwal, S., Xia, P., Wiseman, S., Livescu, K., and Gimpel, K. (2021). On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference* , pages 111–120.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding* , pages 45–52. Association for Computational Linguistics.