DramaCoref: A Hybrid Coreference Resolution System for German Theater Plays

Janis Pagel
Institute for Natural Language Processing
University of Stuttgart
pageljs@ims.uni-stuttgart.de

Nils Reiter
Department of Digital Humanities
University of Cologne
nils.reiter@uni-koeln.de

Abstract

We present a system for resolving coreference on theater plays, *DramaCoref*. The system uses neural network techniques to provide a list of potential mentions. These mentions are assigned to common entities using generic and domain-specific rules. We find that DramaCoref works well on the theater plays when compared to corpora from other domains and profits from the inclusion of information specific to theater plays. On the best-performing setup, it achieves a CoNLL score of 32% when using automatically detected mentions and 55% when using gold mentions. Single rules achieve high precision scores; however, rules designed on other domains are often not applicable or yield unsatisfactory results. Error analysis shows that the mention detection is the main weakness of the system, providing directions for future improvements.

1 Introduction

We present experiments on resolving coreferences in theater plays/dramas. *Coreference resolution* (CR) aims to assign all mentions in a text to their respective discourse entities. Doing so offers many benefits for working with theater plays, yet CR on this type of text is rarely performed.

Next to prose and poetry, drama is considered one of the three main literary genres since Ancient Greece (Aristoteles, 1982). Dramas are scripts to be performed in theatre. Furthermore, drama is a highly structured text type: To a large extent, it consists of character speech with clear indication of the speaker, as well as some stage directions and a segmentation into acts/scenes. Usually, dramas also contain a so called *dramatis personæ*, which is a list of cast members which will appear during the course of the play. This list typically contains further information about the characters, like family relations, which in turn can be exploited to better resolve certain coreferent mentions.

ACT I, Scene I. *The Prince's Cabinet*
The Prince, seated at a desk, which is covered with papers.
PRINCE. Complaints; nothing but complaints! [...] To be sure, if we could relieve every one, we might indeed be envied. Emilia? *opening a petition, and looking at the signature.* An Emilia? Yes—but an Emilia Bruneschi—not Galotti. Not Emilia Galotti. What does she want, this Emilia Bruneschi? *Reads* She asks much—too much. But her name is Emilia. It is granted *signs the paper, and rings.*

Figure 1: Opening scene of Gotthold Ephraim Lessing's *Emilia Galotti*. Translation from German by the authors.

Characters are a constitutive ingredient for theater plays, and scholarly analysis often revolves around characters, their depiction, and their relations. While dramas make it straightforward to detect character speech, references to characters within the speech of other characters can really only be detected using CR.[1] These character references provide insight into how characters are introduced and seen by other characters, sometimes well before appearing on stage themselves. This can already be seen in Figure 1, which is the opening scene of a play. The character Emilia Galotti only enters the stage in the second act, but from the very beginning, other characters talk about her, thus shaping her reception by readers and audience.

In this paper, we present a system to automatically resolve these coreferences. It makes use of both neural network and rule-based components. Both have been adapted to the dramatic text type, either through fine-tuning or the addition of specific rules.

2 Related work

CR has received a lot of attention, mostly focused on the English language and evaluated on news texts (Lee et al., 2013; Björkelund and Kuhn, 2014; Clark and Manning, 2016; Martschat, 2017; Lee et al., 2017, 2018; Joshi et al.,

[1] Andresen and Vauth (2018) point out similar things for prose texts.

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 36–46
Punta Cana, Dominican Republic, November 10–11, 2021.

2019). The two published CR systems for German are CorZu (Tuggener, 2016) and IMS HotCoref DE (Rösiger and Kuhn, 2016). CorZu is a rule-based system that iteratively eliminates possible mention pairs by checking a number of linguistic features. It achieves 64.79 MELA F-Score (Pradhan et al., 2012) on TüBa-D/Z, an annotated corpus of German news text (Naumann, 2007)[2]. IMS HotCoref DE is a machine learning system that is based on the multi-language IMS HotCoref system by Björkelund and Kuhn (2014). IMS HotCoref searches for possible antecedents based on latent search trees and uses global features to train a perceptron for classification. IMS HotCoref DE modifies IMS HotCoref by adding language-specific properties such as morphological information and making use of GermaNet (Hamp and Feldweg, 1997). Recently, Schröder et al. (2021) presented a system for neural end-to-end coreference resolution on German data, including literary data. Their system outperforms the previous state-of-the-art system IMS HotCoref DE by up to 30 percentage points.[3]

There are only few publications that focus on **CR for literary texts**. BookNLP (Bamman et al., 2014) is a full NLP pipeline optimised for (English) long texts. To resolve coreferential links, noun phrases are clustered, following Davis et al. (2003) (without evaluation). The resolution of anaphora is based on linguistically motivated features and a Bayesian classifier. It achieves an accuracy of 82%, evaluated on under 900 mention pairs from three literary novels. Krug et al. (2015) describe a CR system that only resolves references to literary characters. The system is an adaptation of Lee et al. (2011) and achieves a performance of about 56 B^3 F-Score (outperforming CorZu). Van Cranenburgh (2019) also build upon Lee et al. (2011) and present results for CR on Dutch novels. Like Krug et al. (2015), they limit their annotation of mentions to literary characters, but also include inanimate objects. Table 1 shows self-reported results from the two aforementioned works for CR on literary texts, contrasted with results from Lee et al. (2011), on which system both works are build upon.

To our knowledge, no system has been published for German data that is tailored to dramatic or dialogical texts. Only a few publications deal explicitly with domain adaptation for CR systems (Apostolova et al., 2012; Yang et al., 2012; Zhao and Ng, 2014). Do et al. (2015) describe a method for adapting the Berkeley CR system to narrative texts using linear programming.

In terms of **data sets**, there are only a few readily available. Bamman et al. (2019) released a corpus with annotated entity types (following the ACE standard) on English literary texts. Rösiger et al. (2018) describe challenges in the annotation of literary texts, including plays. In previous own work (Pagel and Reiter, 2020), we have presented a corpus of 31 German language plays annotated for coreference, which will be used as the main resource in this paper.

3 DramaCoref

We propose a hybrid system, using neural and rule-based components, for resolving coreference on theater plays, called *DramaCoref*. The system follows the classic distinction into mention detection and coreference resolution component. For mention detected, we apply a system that is based on a BERT-model, and fine-tune it to our task and domain. We generally see mention detected as similar to named entity recognition (NER), and follow best practices and settings for NER.

The coreference resolution component is based on Stanford's Multi-Pass Sieve Coreference Resolution System (Raghunathan et al., 2010; Lee et al., 2011, 2013). Rule-based systems have been shown to provide good results for CR on literary data (Krug et al., 2015; van Cranenburgh, 2019). For this type of data, rule-based systems also come with certain advantages, as (i) no training data is needed, which is usually sparse for the literary domain (cf. Krug et al., 2015), (ii) rule-based systems tend to generalize better on unseen domains (Lee et al., 2013, p. 886) and (iii) the rules can be easily crafted and changed to fit the needs of the specific domain (van Cranenburgh, 2019, p. 28).

We adopt all passes of Raghunathan et al. (2010) and Lee et al. (2011) to test how general purpose rules fair on dramas and add passes designed to yield high precision on this type of text. In particular, we add a pass that matches lemmas of heads of mentions, since the dramas are in German language and head words can be morphologically complex. Furthermore, we advance the pronoun related rule of Lee et al. (2011) and split it into first, second and third person rules. This allows us to specifically target these different cases and exploit the fact that

[2] https://uni-tuebingen.de/de/134290

[3] The paper got released during the final preparation of this paper.

Paper	Mentions	MUC	B^3	$CEAF_e$	CoNLL
Lee et al. (2011)	auto	0.61	0.69	0.45	0.58
	gold	0.65	0.71	0.48	0.61
Krug et al. (2015)	auto	0.86	0.56	NA	NA
	gold	NA	NA	NA	NA
van Cranenburgh (2019)	auto	0.71	0.62	0.67	0.67
	gold	0.78	0.72	0.78	0.76

Table 1: Comparison of results from different papers. All scores for the different evaluation measures are F1 scores.

dramas mark speaker turns.

Like in Lee et al. (2011), single passes are tested for their precision on a held-out dataset and then ordered by precision; the pass with the highest precision is applied first. Passes which are applied later receive the clusters of previous passes as input, so clusters are build up successively.

Table 2 gives an overview of all passes deployed within DramaCoref. Pass 1 groups mentions into clusters that have identical surface strings. Pass 2 matches structurally predictable constructions, such as appositions and relative clauses (cf. Raghunathan et al., 2010). Pass 3 matches for identical heads of two mentions, while disallowing mentions that are embedded within each other or that do not contain stop words or modifiers present in the cluster of the other mentions. Passes 4 and 5 are variations on this and drop the aforementioned constraints, respectively. Pass 6 drops the constraints of passes 4 and 5, but requires two mentions to share the same named entity category. Pass 7 and 9 resolve pronouns. Pass 10 matches mentions that are identical after dropping any words following the head words. Pass 11 matches mentions with proper nouns as head words. Pass 12 handles information from lexical resources and matches mentions which heads are in a synonymy or hyponymy relationship. For a detailed description of the functionality of passes 1–12, see Raghunathan et al. (2010) and Lee et al. (2011). Pass 2 from Raghunathan et al. (2010) and pass 12 and 14 (here pass 9) from Lee et al. (2011) have been subdivided into passes 2a-c, 9a-b and 12a-b to enable a more detailed evaluation, but still serve the same purpose as in the original papers. While pass 7 from Raghunathan et al. (2010) originally resolved general pronouns, it here only resolves third person pronouns. Pass 11 from Raghunathan et al. (2010) has been extended by a new variation 11a. All changes to passes from Lee et al. (2011) and Raghunathan et al. (2010) as well as passes newly introduced to the system are listed below:

Pass 11a Like pass 11, but adds the constraint that modifiers of either candidates cannot contain a first person pronoun if the speakers differ or a second person pronoun if the next or previous speaker differs.

Pass 12a and 12b Like pass 12 in Lee et al. (2011), but instead of WordNet (Fellbaum, 1998), GermaNet (Hamp and Feldweg, 1997) is used.

Pass 14 This pass works the same as pass 1, except it does not operate on tokens but lemmas only. This change targets the morphological complexity of German, so that inflected word forms are not counted as a mismatch.

Post-processing In a post-processing step, clusters are merged when they contain at least one mention that is referring to the same cast member of the play. In the cases of first-person pronouns, the speaker can be identified as the referred cast member. In the cases of proper names, DramaCoref gets a list of all cast members from the *dramatis personæ* as input and tries to assign the correct cast member via string comparison between all mentions in the dramatis personæ and all mentions in the cluster.

In a very last step, all singletons, i.e. clusters that only contain one single mention, are removed, since all the datasets used in the experiments do not contain singletons.

4 Experiments

4.1 Data

The experiments are performed on the GerDraCor-Coref data (Pagel and Reiter, 2020). GerDraCor-Coref is a corpus of coreference annotations on dramas in German language, including 31 texts

Pass ID	Short Name	Source
1	ExactMatch	Raghunathan et al. (2010)
2a	Acronyms	Raghunathan et al. (2010)
2b	Appositions	Raghunathan et al. (2010)
2c	RelPron	Raghunathan et al. (2010)
3	StrictHeadMatch	Raghunathan et al. (2010)
4	StrictHeadMatchVar1	Raghunathan et al. (2010)
5	StrictHeadMatchVar2	Raghunathan et al. (2010)
6	HeadEntail	Raghunathan et al. (2010)
7	Pron3rdPers	Raghunathan et al. (2010)
9a	SpeakerPron1stPers	Lee et al. (2011), modified
9b	SpeakerPron2ndPers	Lee et al. (2011), modified
10	RelaxedStringMatch	Lee et al. (2011)
11	ProperHeadWordMatch	Lee et al. (2011)
11a	PoperHeadWordMatchVar1	New
12a	LexicalSynonym	Lee et al. (2011), modified
12b	LexialHyponym	Lee et al. (2011), modified
14	ExactLemmaMatch	New

Table 2: All passes of DramaCoref.

from 1732 to 1921. All texts come from GerDraCor (Fischer et al., 2019), a corpus of German dramas, which is encoded in TEI/XML to mark act and scene segmentations and speaker changes, among others. In total, the dataset is comprised of 298,352 tokens, 61,126 mentions and 5,473 entities. The most frequent part-of-speech for mentions are pronouns with 40% of the overall number of mentions, followed by noun phrases and named entities.

Pagel and Reiter (2020) point out several observations about the data and regarding the use of coreference in the texts: Since the texts are considerably longer than in common domains, such as newspaper texts, coreference chains can become quite long as well. Especially chains involving main characters of the plays entail many mentions and span the entirety of the text. On average, dramas contain more mentions, but less entities when compared to news and interview corpora. Additionally, the number of pronouns in coreference chains is much higher compared to other domains (40% for GerDraCor-Coref, compared to 20% in TüBa-D/Z).

For the experiments, we use all mentions and coreference clusters from GerDraCor-Coref except for clusters containing non-nominal antecedents (cf. Kolhatkar et al., 2018).

4.2 Neural mention detection

We extract all possible mentions using a transformer-based approach via fine-tuning a pre-trained German BERT model (Devlin et al., 2019) provided by HuggingFace[4]. While the pre-trained BERT model is not specifically tuned to work on Named Entity Recognition (NER) or similar tasks, it nevertheless performs well on German benchmark NER evaluation datasets, like CoNLL 2003 (0.80 F1) or GermEval14 (0.84 F1)[5].

We also compare the performance to HuggingFace's DistilBERT model[6] (Sanh et al., 2019), a smaller version of BERT with comparable performance, and a BERT model fine-tuned on German literary texts and NER, also hosted by HuggingFace[7].

While mention detection is not identical with NER, we assume that the tasks are similar enough that a model for mention detection will benefit from knowledge about named entities.

Experimental setup The pre-trained BERT models all have embedding and hidden state dimension size of 768. Dropout was set to 0.1 and 12 attention

[4] https://huggingface.co

[5] https://huggingface.co/bert-base-german-cased

[6] https://huggingface.co/distilbert-base-german-cased

[7] https://huggingface.co/severinsimmler/literary-german-bert

heads were used while training. In all cases, we run fine-tuning over 4 epochs with an early stopping criterion of 3 and use a train-test split of 80%-20%. AdamW was used as optimizer. Learning rate was set to $4e{-}5$ and the batch size to 8. The maximum length a sequence could have was set to 128.

The BERT-based models are compared to a baseline that uses the Berkeley Parser (Petrov et al., 2006). From the parser's output, we treat all predicted noun phrases as mentions[8].

Results The results of our neural mention detection experiments are shown in Table 3. We find that our setup outperforms the baseline based on parsing in terms of accuracy and precision, however, the baseline yields slightly better results in the macro F1 score. Using a BERT-based model fine-tuned on German literary text is not able to outperform the generic German BERT model, but rather lowers the performance quite notably. We do not find a difference in the performance of the base BERT model and DistilBERT.

Considering the higher score in accuracy with an almost identical F1 score, we opt to use the output of DistilBERT over the parser output as input for the coreference resolution component.

4.3 Sieve-based coreference resolution

We use the best performing model of the experiment in Section 4.2 as input for *DramaCoref*, which is DistilBERT. This component works as described in Section 3. All evaluation scores are measured using the coreference scorer from Pradhan et al. (2014). For the development set, on which the ordering of the passes will be determined, we randomly assign 20% of the documents from GerDraCor-Coref, while the rest is used for testing.

Baselines Before discussing the results of our resolution component, we show different baselines on the GerDraCor-Coref data in order to get a better understanding of the general performance of DramaCoref, since there are no other results to compare with on this dataset.

We showcase three different baselines[9]: (B1) N mentions in N clusters, meaning that every mention is assigned to its own cluster, so that we end up with as many clusters as mentions, each containing exactly one mention, (B2) N mentions in 1 cluster, meaning all mentions are grouped together in a single cluster and (B3) Closest agreeing candidate, which assigns each mention to the closest previous mention it agrees with syntactically and semantically, i.e. in number, gender and NE class.

We can see from Table 4 that B2, which groups all mentions in a single cluster, performs best, closely followed by B3. Looking at the structure of coreference clusters in the data, this makes sense, since the clusters with the most mentions are clusters for dramatic characters, making up a large amount of all mentions, and resemble the setup in the second baseline very closely. Still, performance scores are rather low.

Results Table 6 shows the main results of our coreference resolution component. We show both results on predicted and gold mentions and for different setups. For *act* and *scene*, DramaCoref is applied on whole acts or scenes contained within these acts, respectively. For all following setups, the *scene* setup has been used as a basis. For the setup of *castmembers-only*, we remove all mentions that are not listed in the dramatis personæ, so only literary characters are considered as mentions and resolved. This resembles more closely the setups by Krug et al. (2015) and van Cranenburgh (2019). The setup *dramatispersonae* applies the post-processing step described before, i.e. it merges clusters where it can find common strings found in the dramatis personæ.

All setups perform relatively equal, with *dramatispersonae* boosting the results slightly compared to *act*. The best results on automatic mentions is achieved by the *scene* setup, but for gold mentions, *dramatispersonae* is best. Applying DramaCoref only on cast members yields worst results, mainly because of the CEAF_e score dropping. This shows a need to find better ways of resolving clusters of literary characters, as approximately 61% of all mentions in GerDraCor-Coref refer to literary characters.

We also show the results of DramaCoref on texts from other domains, namely TüBa-D/Z (Naumann, 2007) as newspaper domain, DIRNDL (Björkelund et al., 2014) as radio news domain and CRETA (Rösiger et al., 2018) as a corpus on other literary texts which are not dramas, in Table 7. For the other corpora, the automatic mentions were provided with them, while for GerDraCor-Coref, we again

[8]The parser uses the PennTree format (Marcus et al., 1993), for which in the used implementation, pronouns are not marked as noun phrases and noun phrases inside prepositional phrases are completely neglected. We therefore additionally add these to our set of output mentions.

[9]B1 and B2 are identical to baselines also applied in Krug et al. (2015)

Model	MacroPrec	MacroRec	MacroF1	Acc
Parsing Baseline	0.5	**0.57**	**0.53**	0.7
DistilBERT	**0.54**	0.51	0.52	**0.84**
Base BERT	**0.54**	0.51	0.52	**0.84**
Literary BERT	0.24	0.33	0.28	0.73

Table 3: Results of the mention detection experiments.

	auto			gold		
Baseline	Precision	Recall	F1	Precision	Recall	F1
B1: N mentions in N clusters	0.23	0.13	NA	0.35	0.21	NA
B2: N mentions in 1 cluster	**0.29**	**0.33**	**0.22**	**0.45**	**0.67**	**0.38**
B3: Closest agreeing candidate	0.24	0.23	**0.22**	0.39	0.41	0.37

Table 4: CoNLL scores for running the different baselines on GerDraCor-Coref.

use the mentions coming from the BERT model in Section 4.2.

Looking at the CoNLL scores, DramaCoref works very well on CRETA and DIRNDL when using automatically detected mentions. The results on GerDraCor-Coref are slightly higher than on TüBa-D/Z. For the gold mentions, the CoNLL scores for TüBa-D/Z are highest and lowest for GerDraCor-Coref. On the one hand, this shows that DramaCoref is working in general, as it is able to predict coreference on texts of a different type. On the other hand, it shows that dramas might have properties that require more domain-specific rules and are not covered by systems build for newspaper texts. What can furthermore be seen is that with gold mentions, DramaCoref is able to predict much better than with the automatically detected mentions, showing the need for a more sophisticated mention detection for dramas.

System	MUC	B^3	$CEAF_e$	CoNLL
HotCoref	**56.55**	14.98	14.84	**28.79**
DramaCoref	42.54	**19.87**	**18.97**	27.12

Table 5: Results for comparing DramaCoref and IMS HotCoref DE.

We also compare the results of DramaCoref to IMS HotCoref DE on the plays. To allow for a comparison, we split the data in training and test set of 70% and 30%, respectively. IMS HotCoref DE is then trained on the training set, while for DramaCoref, the ordering of rules is determined on this set. Both systems are tested on the same test set. The results can be seen in Table 5. We can see that DramaCoref achieves comparable results to IMS HotCoref DE, outperforming it in the B^3 and $CEAF_e$ scores. However, IMS HotCoref DE performs better for the MUC score, resulting in a slightly higher CoNLL score.

5 Error analysis

By having a closer look at the performance of the single passes in Figure 2, we can see that, as expected, passes 9a and 9b, which handle first and second person pronouns, perform very well. Also pass 1 for exact string matches is relatively reliable, the remaining passes perform relatively on par. Only passes 2a-c have almost no hits, as they handle constructions normally not present in dramas. We also include the three baselines B1-3 for comparison, which manage to outperform several passes. Pass 9a performing well is not surprising, as first person pronouns can often be assigned very reliably to the cluster of the current speaker. Assigning second person pronouns with pass 9b works surprisingly well, showing that a simple heuristic of assuming that the current character addresses the previous or following speaker works in many cases. Pass 1, which performed highest for Raghunathan et al. (2010), is also strong for GerDraCor-Coref, but does not achieve a precision of 90+% as in Raghunathan et al. (2010). This might be due to the length of the texts, where mentions with identical surface forms can appear many times in different contexts. Using lemmas instead of tokens for pass 14 does not outperform pass 1.

	Mentions	MUC	B^3	$CEAF_m$	$CEAF_e$	CoNLL
act	auto	0.45	0.23	0.32	0.19	0.29
	gold	0.79	0.42	0.43	0.41	0.54
castmembers-only	auto	0.44	0.26	0.35	0.15	0.28
dramatispersonae	auto	0.46	0.23	0.33	0.19	0.29
	gold	0.79	0.44	0.45	0.42	0.55
scene	auto	0.43	0.27	0.37	0.24	0.32
	gold	0.72	0.47	0.49	0.42	0.54

Table 6: Results of DramaCoref on GerDraCor-Coref with different setups. All scores are F1 scores.

Mentions	Corpus	Mention Fscore	MUC	B^3	$CEAF_m$	$CEAF_e$	CoNLL
auto	CRETA	0.70	0.51	0.30	0.34	0.26	0.36
	DIRNDL	0.43	0.36	0.30	0.38	0.37	0.38
	TüBa-D/Z	0.46	0.26	0.25	0.33	0.31	0.28
	GerDraCor-Coref	0.60	0.45	0.23	0.32	0.19	0.29
gold	CRETA	0.96	0.72	0.46	0.46	0.47	0.55
	DIRNDL	0.70	0.61	0.58	0.63	0.61	0.61
	TüBa-D/Z	0.86	0.66	0.62	0.65	0.64	0.64
	GerDraCor-Coref	0.96	0.79	0.42	0.43	0.40	0.54

Table 7: Scores for running DramaCoref on different corpora. All scores are F1 scores.

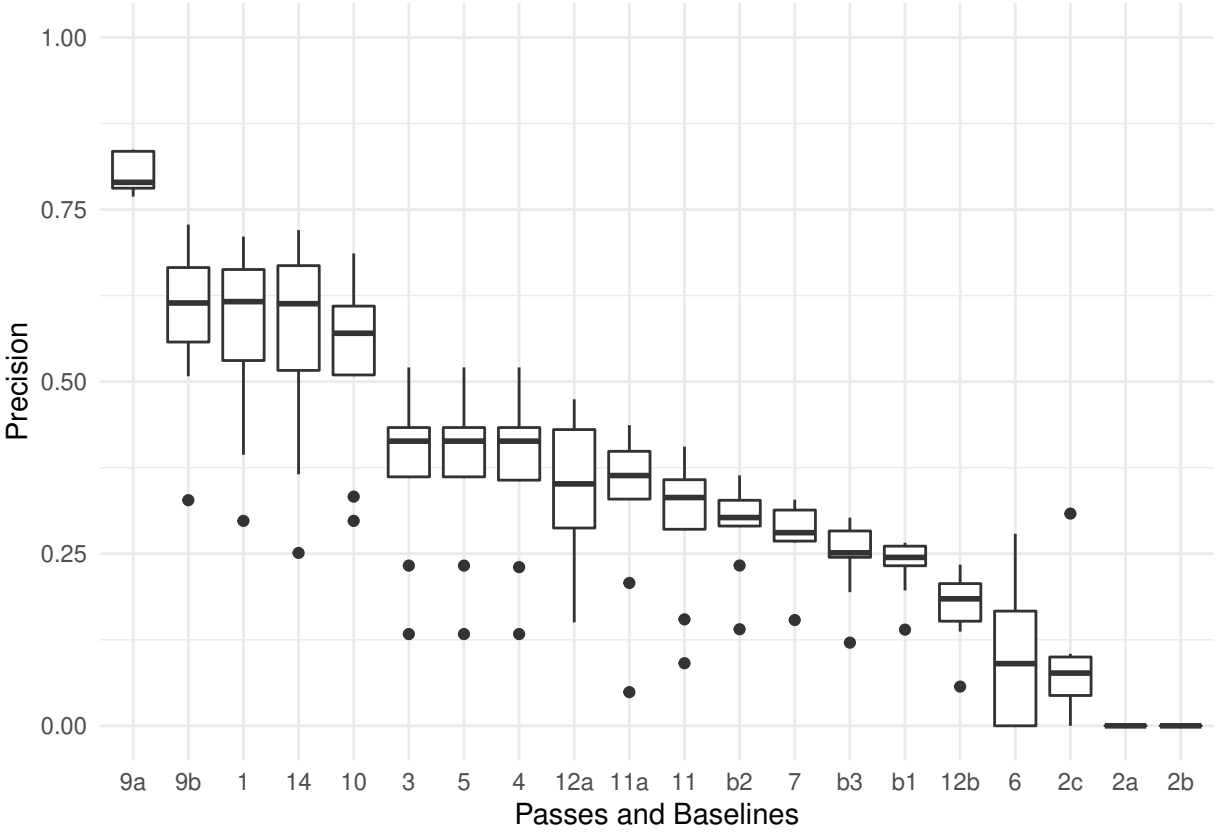

Figure 2: Performance of passes on the development set, ordered by precision. The precision is measured using the CoNLL score.

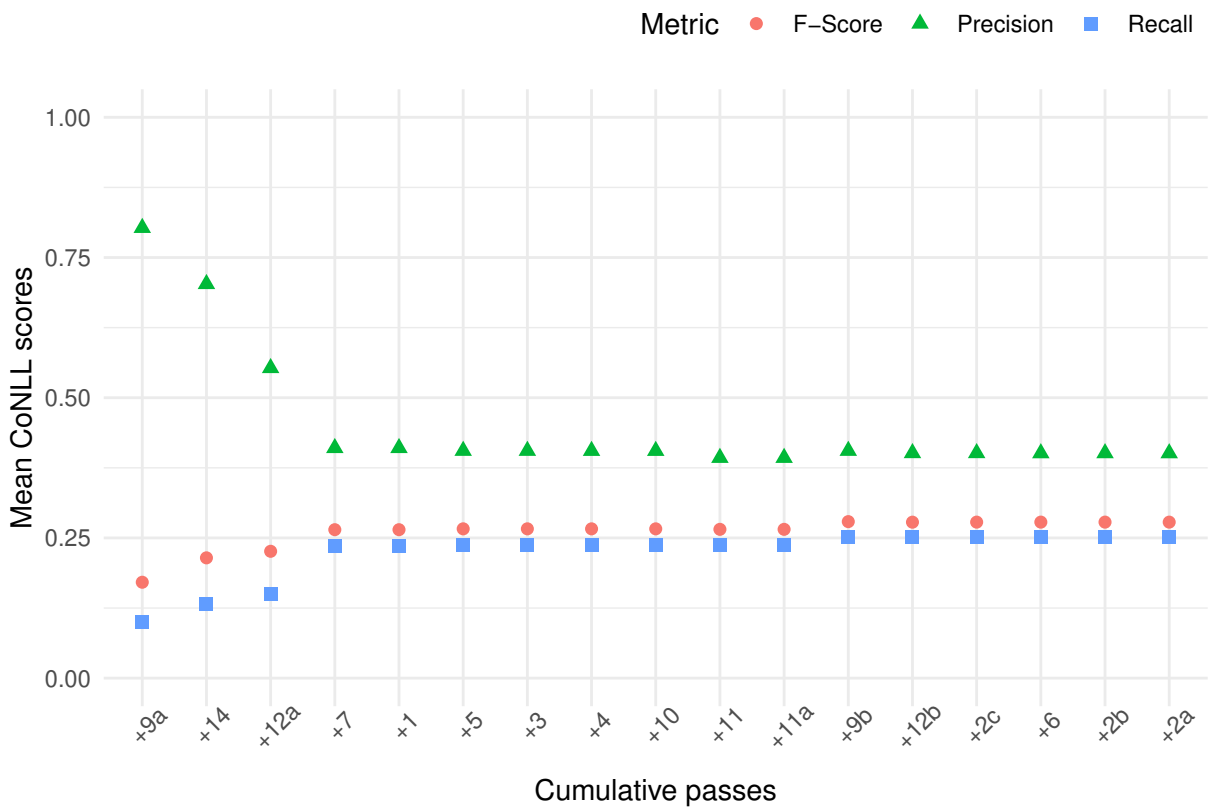

Figure 3: Passes ordered by their precision on the test set and applied cumulatively, showing mean precision, recall and F1 for the CoNLL score. Cumulatively means that from left to right, the given pass is added to all previously applied passes.

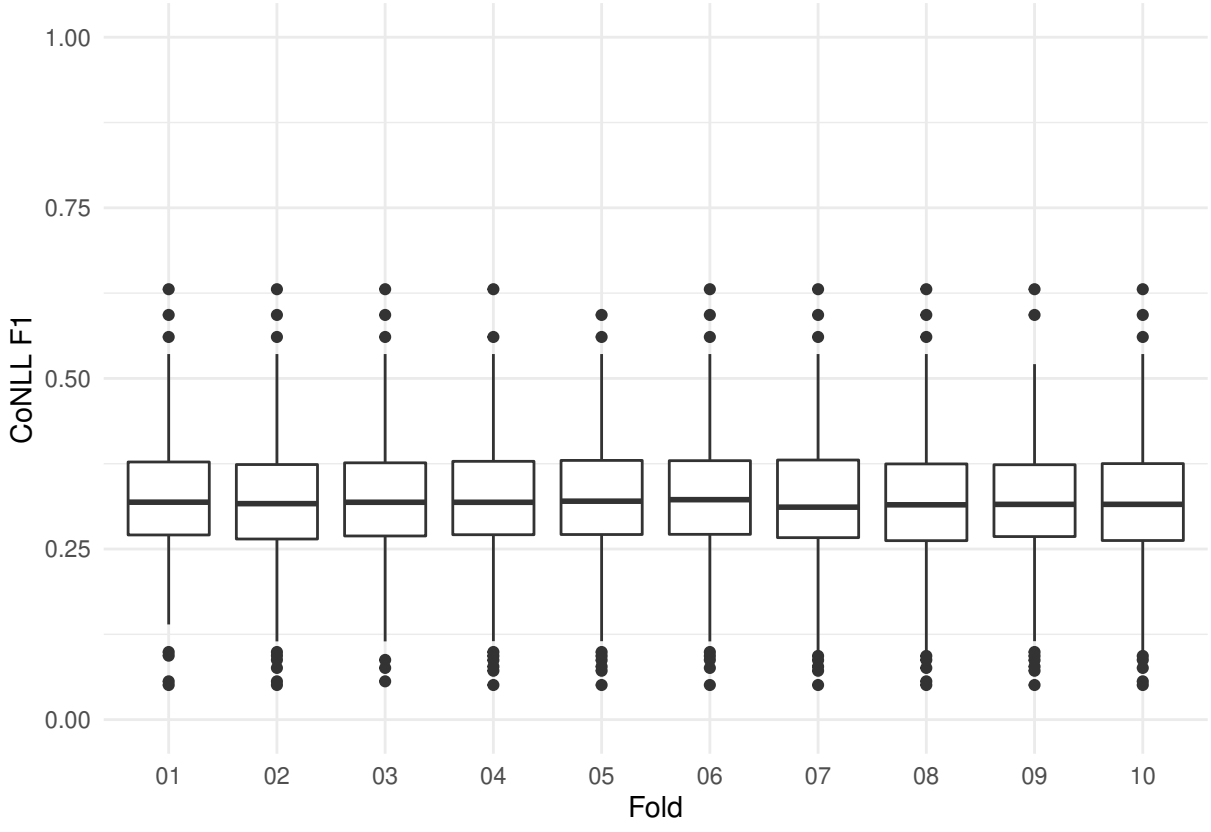

Figure 4: Results in a 10-fold cross validation setting.

Figure 3 shows all passes applied cumulatively. This means that the pass with the highest precision on the development set is evaluated on the test, then the pass with the highest and second-highest precision, and so on. This enables us to see how the evaluation scores change when more and more passes are added. Other than in Figure 2, the passes are evaluated on the test set instead of the development set, in order to get more meaningful numbers on a larger dataset. Thus, the ordering of the passes is different to Figure 2. Figure 3 emphasizes that, while the overall precision naturally falls when more passes with lower individual precision are added, the recall does not rise significantly. This calls for passes that are able to increase recall while not dropping precision too much. While the used passes did exactly that in the experiments of Raghunathan et al. (2010) and Lee et al. (2011), and also for van Cranenburgh (2019) on Dutch novels, this seemingly does not apply to the dramas in GerDraCor-Coref.

In general, passes from Lee et al. (2011), which are meant for general purpose domains, do not perform well enough on dramas, while passes specifically engineered for this domains, do not perform well enough in order to achieve results comparable to other domains like newspaper texts.

Lastly, we make sure that the results are not caused by splitting the data into a certain train-test split by applying 10-fold cross validation. The results are shown in Figure 4. As can be seen, the results are relatively constant across folds, meaning that the plays form a homogeneous whole.

6 Conclusion and future work

We present a system, DramaCoref, for coreference resolution on German theater plays. DramaCoref is hybrid in the sense that it uses a transformer architecture to retrieve potential mentions and in a second step clusters the retrieved mentions using ordered sieves. It performs comparable to other attempts of CR on literary data, however, theater plays appear to be a more difficult type of data for CR than texts of other types, e.g. newspaper texts. While some passes that work well on newspaper texts do not apply or underperform on the theater plays, specific passes and newly added passes perform reasonably well. Our goal for future research is the exploration of new passes, utilizing information coming from the plays like speaker turns and informations from the dramatis personæ like family relations, and improving the mention detection component. Developing an end-to-end neural network based system might also be worthwhile, as it makes the need for two distinct components obsolete; however, at this point, the available training data does not seem to be sufficient for such an approach.

Acknowledgements

The work described has been conducted in the QuaDramA project, funded by the Volkswagen foundation and in the Q:TRACK project, funded by the German Research Foundation (DFG) in the context of SPP 2207 *Computational Literary Studies*. We thank both for making this possible.

References

Melanie Andresen and Michael Vauth. 2018. Added value of coreference annotation for character analysis in narratives. In *Proceedings of the Workshop on Annotation in Digital Humanities (annDH)*, pages 1–6.

Emilia Apostolova, Noriko Tomuro, Pattanasak Mongkolwat, and Dina Demner-Fushman. 2012. Domain adaptation of coreference resolution for radiology reports. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 118–121.

Aristoteles. 1982. *Poetik*. Reclam, Stuttgart, Germany.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2138–2144.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 370–379.

Anders Björkelund, Kerstin Eckart, Arndt Riester, Nadja Schauffler, and Katrin Schweitzer. 2014. The extended DIRNDL corpus as a resource for automatic coreference and bridging resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 3222–3228.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 47–57.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 643–653.

Andreas van Cranenburgh. 2019. A Dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

Peter T. Davis, David K. Elson, and Judith L. Klavans. 2003. Methods for precise named entity matching in digital collections. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, pages 125–127.

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 (NAACL-HLT)*, pages 4171–4186.

Quynh Ngoc Thi Do, Steven Bethard, and Marie-Francine Moens. 2015. Adapting coreference resolution for narrative processing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2262–2267.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT, Cambridge, MA.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable corpora: Introducing DraCor, an infrastructure for the research on European drama. In *Proceedings of DH2019: "Complexities"*.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

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.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: A survey. *Computational Linguistics*, 44(3):547–612.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimer. 2015. Rule-based coreference resolution in German historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 28–34.

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 (EMNLP)*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 687–692.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sebastian Martschat. 2017. *Structured Representations for Coreference Resolution*. Ph.D. thesis, Heidelberg University.

Karin Naumann. 2007. *Manual for the Annotation of in-document Referential Relations*. Abt. Computerlinguistik Universität Tübingen.

Janis Pagel and Nils Reiter. 2020. GerDraCor-Coref: A coreference corpus for dramatic texts in German. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 55–64.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Stroudsburg, PA, USA. 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 (ACL)*, pages 30–35.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL 2012,

pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 492–501.

Ina Rösiger and Jonas Kuhn. 2016. IMS HotCoref DE: A data-driven co-reference resolver for German. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 155–160.

Ina Rösiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Fynn Schröder, Hans Ole Hatzel, and Chris Biemann. 2021. Neural end-to-end coreference resolution for German in different domains. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS)*.

Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. thesis, University of Zürich.

Jian Bo Yang, Qi Mao, Qiao Liang Xiang, Ivor W. Tsang, Kian Ming A. Chai, and Hai Leong Chieu. 2012. Domain adaptation for coreference resolution: An adaptive ensemble approach. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Comutational Natural Language Learning (EMNLP-CoNLL)*, pages 744–753.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Mining and Information Analysis (Louhi)*, pages 21–29.

A Overview table with detailed results

Table 8 shows detailed results for applying DramaCoref on different corpora, for several evaluation metrics broken down by precision, recall and F1-score.

Mentions	Corpus	Mention Fscore	MUC			B^3			$CEAF_m$			$CEAF_e$			CoNLL		
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
auto	CRETA	0.70	0.51	0.51	0.51	0.39	0.25	0.30	0.33	0.36	0.34	0.21	0.35	0.26	0.37	0.37	0.36
	DIRNDL	0.43	0.19	0.21	0.36	0.22	0.24	0.30	0.28	0.30	0.38	0.27	0.28	0.37	0.23	0.24	0.38
	TüBa-D/Z	0.46	0.18	0.25	0.26	0.21	0.28	0.25	0.26	0.38	0.33	0.23	0.38	0.31	0.21	0.30	0.28
	GerDraCor-Coref	0.60	0.58	0.38	0.45	0.49	0.16	0.23	0.38	0.28	0.32	0.15	0.26	0.19	0.41	0.26	0.29
gold	CRETA	0.96	0.76	0.69	0.72	0.59	0.38	0.46	0.48	0.45	0.46	0.45	0.49	0.47	0.60	0.52	0.55
	DIRNDL	0.70	0.57	0.31	0.61	0.60	0.30	0.58	0.59	0.34	0.63	0.55	0.35	0.61	0.57	0.32	0.61
	TüBa-D/Z	0.86	0.71	0.51	0.66	0.73	0.47	0.62	0.68	0.51	0.65	0.63	0.54	0.64	0.69	0.51	0.64
	GerDraCor-Coref	0.96	0.83	0.74	0.79	0.66	0.31	0.42	0.44	0.41	0.43	0.37	0.46	0.40	0.62	0.51	0.54

Table 8: Overview of all results of the baselines and DramaCoref on the different corpora.