Neural End-to-end Coreference Resolution for German in Different Domains

Fynn Schröder[*], **Hans Ole Hatzel**[*], and **Chris Biemann**
Language Technology Group
Universität Hamburg, Germany
{fschroeder, hatzel, biemann}@informatik.uni-hamburg.de

Abstract

We apply neural coreference resolution to German, surpassing the previous state-of-the-art performance by a wide margin of 10–30 points F1 across three established datasets for German. This is achieved by a neural end-to-end approach, training contextual word-embeddings jointly with mention and entity similarity scores. We explore the impact of various parameters such as language models, pre-training and computational limits with respect to German data. In an effort to support datasets representing the domains of both news and literature, we make use of two distinct model architectures: a mention linking-based and an incremental entity-based approach that should scale to very long documents such as literary works. Our code and ready-to-use models are publicly available.

1 Introduction

Coreference resolution is the task of resolving text spans in documents that refer to the same entities. These are grouped into mention-clusters with each cluster representing one entity. Figure 1 shows coreference annotations on a literary text with different entities being denoted by both subscripts and colors. Tasks such as question answering (Morton, 1999) or text summarization (Steinberger et al., 2007) can rely on coreference resolution as part of the language processing pipeline. Bamman et al. (2014) demonstrated that coreference resolution is also applicable to literary analysis. The task has recently seen large improvements as systems moved from rule-based (e.g. Roesiger and Kuhn, 2016; Lee et al., 2011) to neural approaches (e.g. Lee et al., 2017; Joshi et al., 2019). This advancement from a CoNLL-F1-score of 57.8, achieved by a rule-based system in the original CoNLL-2012 shared task (Pradhan et al., 2012), to 67.2 in the first end-to-end neural system (Lee et al., 2017) has shown that neural systems are key to state-of-the-art performance.

[Alice]$_1$ was not a bit hurt, and **[she]**$_1$ jumped up on to **[her]**$_1$ feet in a moment: **[she]**$_1$ looked up, but it was all dark overhead; before **[her]**$_1$ was **[another long passage]**$_2$, and **[the White Rabbit]**$_3$ was still in sight, hurrying down **[it]**$_2$.

Figure 1: Coreference gold annotations for "Alice's Adventures in Wonderland" (annotations from Bamman et al., 2020)

Coreference resolution on German using neural networks has received little attention. There has, to our knowledge, no work been reported on German news datasets using neural networks yet. This work is also the first to use cross-task learning to improve performance on German literary datasets.

We apply and adapt exiting approaches to coreference on German, making our code and models publicaly available.[1] There are two approaches to neural coreference resolution that we consider: A mention-linking-based and an entity-linking-based approach. Both have an initial mention proposal step, finding text spans that are likely to represent mentions. In mention-linking approaches, out of the cross-products of mentions, those mentions with the highest likelihood are considered. Each such mention is connected to its highest-scoring antecedent with transitively connected mentions forming entities.

The entity-representation-based approach also involves the initial mention proposal step. However, rather then creating links on a per-mention basis, initial mentions are considered to be entity representations, with each subsequent mention be-

[*] denotes equal contribution

[1] https://github.com/uhh-lt/neural-coref/tree/konvens

ing compared to existing entity representations and assigned to those that match them best. This way memory usage and computational effort can be reduced, as it is proportional to the number of entities, rather than the square of the number of mentions.

2 Related Work

Relevant prior work can be put into two distinct categories: (a) Neural, state-of-the-art coreference resolution developed primarily on English (b) Coreference resolution applied to German.

Most neural coreference resolution models perform a ranking of antecedents based on the pairwise scores of mention candidates (Wiseman et al., 2015; Clark and Manning, 2016a; Lee et al., 2017), at this only relying on local decisions that may not be globally optimal to form coherent entities (Lee et al., 2018). This general architecture has been improved on in multiple ways.

To address the issue of global optimization, Clark and Manning (2016b) and Wiseman et al. (2016) create entity representations during the ranking step. Lee et al. (2018); Kantor and Globerson (2019) iteratively refine mention representations with associated antecedent information, performing what they refer to as higher-order inference.

While the end-to-end coreference model of Lee et al. (2017) uses a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to produce span representations, Lee et al. (2018) see a 3.2 F1 score increase on the English CoNLL-2012 shared task by additionally using ELMo (Peters et al., 2018) embeddings. Lee et al. (2018) also modify the model to perform coarse-to-fine antecedent pruning enabling an efficient computation and potentially allowing the processing of longer documents. Joshi et al. (2019) and Kantor and Globerson (2019) improve upon this by using BERT (Devlin et al., 2019) embeddings instead of the LSTM-based representations and gain another 3.3 F1 points.

Recently, Joshi et al. (2020) presented a model optimized for span representations named SpanBERT and saw another 2.5 point increase in F1 score, which has been reproduced by Xu and Choi (2020). Wu et al. (2020) have taken a different approach to coreference resolution; they outperform previous state of the art by 3.5 F1 points in part due to the ability to recover missed mentions by framing the task as a question-answering problem.

Toshniwal et al. (2020); Xia et al. (2020) both introduce incremental approaches to coreference resolution. Instead of comparing mention pairs like Lee et al. (2017), they compare mentions with entity representations, with the entity representations being produced from a linear combination of their mentions. Both approaches work by iteratively processing all mentions and scoring each mention with regard to a set of entities; as a result, an evaluation of the full cross-product of mentions is not necessary. The two approaches differ slightly in how they handle the introduction of new entities.

For coreference resolution on German texts, published work predates the age of neural networks in natural language processing. The CorZu system (Klenner and Tuggener, 2011; Tuggener and Klenner, 2014) is a rule-based incremental entity-mention model that has been extended with Markov Logic Networks for the antecedent selection.

Roesiger and Kuhn (2016) adapted the English system of Björkelund and Kuhn (2014) to German. A directed tree where each node represents a mention is used to model the coreferences in a document. For determining antecedents, both local and non-local handcrafted features are employed. They created the current state-of-the-art approach for German news datasets, evaluating their system on the SemEval-2010 shared task and on version 10 of the TüBa-D/Z dataset.

The domain of literature has, for both German and English, received increased attention in recent years with regard to coreference resolution. Roesiger et al. (2018) considered the domain specific challenges and phenomena of literature. Bamman et al. (2020) released an English dataset and Krug et al. (2018) released a German dataset (see Section 3.2 for details). While Krug (2020) performed coreference resolution on German literary data, Toshniwal et al. (2020) used the English dataset. Krug (2020) compare various approaches to coreference resolution on German historic novels using the DROC dataset (Krug et al., 2018). Their best-performing system in a gold-mention scenario uses a rule-based Stanford Sieve approach (Lee et al., 2011), iteratively applying rules starting from the most precise rule, going to less precise rules. When mention spans are generated by the model, the end-to-end neural network, based on the approach by Lee et al. (2017), performs about on par with the rule-based systems in conjunction with preprocessing pipelines.

Evaluation of coreference data presents a challenge, different proposed metrics emphasise different aspects of a model's performance. An average of the three metrics MUC, B^3, and $CEAF_{\phi_4}$ has been used in the CoNLL-2012 task (Pradhan et al., 2012). As these metrics are widely used we focus on them for reporting our results, including an average of the three, the CoNLL-F1 score.

3 German Coreference Datasets

3.1 News

The standard corpus for coreference resolution in German is TüBa-D/Z (Telljohann et al., 2017; Naumann and Möller, 2006), a manually annotated collection of newspaper articles released in multiple versions that incrementally add more documents. It was also used as the data source for the German part of the SemEval-2010 shared task on coreference resolution (Recasens et al., 2010).

To be comparable with previous work, we chose to use SemEval-2010 and TüBa-D/Z release 10.0 instead of the marginally larger 11.0 for most of our experiments. As there is no official split for the TüBa-D/Z, we use the same splits as previous work (Roesiger and Kuhn, 2016).[2]

While TüBa-D/Z does not contain singletons (on average 3.65 mentions per entity, 10.89 entities per article), these mentions are annotated in SemEval-2010 (on average 1.34 mentions per entity, 73.07 entities per article). Across the dataset, 84.6% of all entities and 64.1% of all mentions are singletons.

Compared to the standard English coreference corpus, OntoNotes (Weischedel et al., 2013), used in the CoNLL-2012 shared task on coreference resolution (Pradhan et al., 2012), TüBa-D/Z neither contains different genres of texts nor additional metadata such as speaker information. Regarding statistics such as average mentions per entity, mentions/sentence length and tokens/sentences/entities per document, German TüBa-D/Z 10.0 and English OntoNotes 5.0 are remarkably similar.

3.2 Literature

The DROC dataset (Krug et al., 2018) contains 90 coreference annotated literary documents where each document comprises one chapter with an average length of 4369.49 tokens. We use the splits established by Krug (2020), i.e. 58 training, 14 development and 18 test documents. There is a total of 51 797 mentions in 5365 clusters, 2409 of these are singleton clusters. As a result, while 45% of clusters are singleton clusters, only 4.7% of mentions are singletons. Our calculations for the performance of human annotators on the subset of DROC are listed in Table 1, providing an upper bound for our performance expectations. In contrast to other datasets (e.g. Bamman et al., 2020), only mention heads are annotated, rather than whole nominal phrases. This means that in the sentence, "and [the driver] was none other than [that cursed Englishman]" (from the dataset by Bamman et al. (2020) "The Scarlet Pimpernel"), only the spans "Englishman" and "driver" would be annotated as coreferring instead. Thus, only spans up to a short length need to be considered in the mention proposal step. DROC also differentiates itself from other datasets in that it only annotates references to characters.

Mention-F1	MUC-F1	B^3-F1	CEAF$_{\phi_4}$-F1	CoNLL-F1
97.05	93.67	84.69	69.25	82.54

Table 1: Inter-annotator F1 scores for DROC as calculated using the scorer by Pradhan et al. (2012) based on the individual annotator's data by Krug et al. (2018).

More generally, literary data, when compared to news texts, comes with the added challenge of document length. Longer documents tend to come with more mentions, DROC, for example, contains an average of 575.52 mentions per document whereas SemEval only has an average of 97.79. In general, increased document length lead to longer processing time, larger computational effort and higher memory requirements.

4 Model

In this section, we describe our German coreference resolution models in detail. We build on the widely adapted neural end-to-end architecture developed by Lee et al. (2017, 2018), improved by Joshi et al. (2019) and re-implemented in PyTorch (Paszke et al., 2019) by Xu and Choi (2020). Although the CorefQA system (Wu et al., 2020) is currently the top-performing system for English, we chose to not build upon it because it is more complex and requires vastly more computational resources than our chosen approach.

The general idea of our models is to first detect mentions and then to link them. Each document is processed individually during both training and inference; Figure 2 visualizes a single document being processed by both model variants. First, contextual ELECTRA (Clark et al., 2020) embeddings are obtained for each token and all possible

[2] for corpus statistics, see Table 9 in the appendix

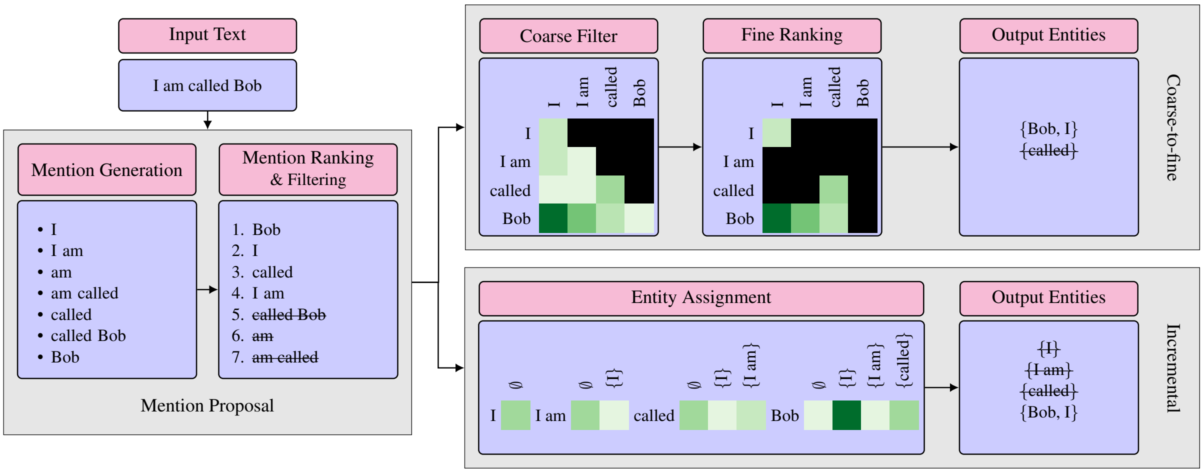

Figure 2: Conceptual visualization of our two end-to-end model variants processing an example document. Both models are based on the same mention proposal step. While the incremental model operates on an ever-growing set of entities, the coarse-to-fine model performs one comparison on the cross product of all mentions. Dark green color indicates a good match between mention and its assignment candidate, whereas black squares indicate that, due to filtering, no scoring was performed. All values are manually chosen for illustration purposes.

mention spans up to a configurable length are enumerated. Mention embeddings are created, containing start and end token embeddings and the attention-weighted average of all span tokens. In contrast to the English models, our models contain neither genre nor speaker embeddings as the German datasets do not supply this information.

A naïve approach of comparing each mention candidate with every other to find links between them raises computational issues, quickly becoming infeasible to compute as it requires $O(M^2)$ comparisons for $M = \text{max\_mention\_length} \cdot |D|$ mention candidates, for a document D where $|D|$ is the document length in word-piece tokens. To reduce computational effort over a naïve approach to find the best antecedent for each mention, we employ two established strategies: A coarse-to-fine and an incremental approach, with the incremental approach being able to handle documents of arbitrary length with limited memory.

4.1 Coarse-to-fine

Our model is based on the implementation by Xu and Choi (2020). For each mention span, the model learns a distribution over its antecedents based on how likely both individual spans are to be valid mentions and how likely they to refer to the same entity. Two pruning steps are used to make this mention linking computationally feasible.

To reduce the number of mentions, all mention embeddings are scored individually with a feed-forward neural network (FFNN). For each document D only the top $n = \min(4096, 0.4 \cdot |D|)$ mentions are kept after pruning. Instead of performing a pairwise comparison of all N mentions, only a fraction is used. Thus, removing obvious non-mentions and limiting the complexity to $O(n^2 \ll N^2)$, a step that we refer to as mention filtering.

In the coarse antecedent pruning step, the pairwise similarity scores of the remaining mention embeddings are summed with the individual mention scores. A subsequent fine-grained ranking is performed with the top $a = 64$ antecedents per mention; to this effect, pairwise mention-antecedent embeddings consisting of mention, antecedent and similarity embedding are created. These embeddings are scored with a FFNN and combined with scores from the coarse step resulting in scores for the top antecedents per mention. We do not use so-called higher-order inference as this effectively doubles the computational cost of the fine-grained antecedent scoring without improving the quality according to Xu and Choi (2020).

During training, the model learns to optimize the marginal log-likelihood of possibly correct antecedents for each mention, i.e. for each antecedent the score should be 1 if mention and antecedent belong to the same gold entity, 0 otherwise.

During inference, an undirected graph of mentions is created by connecting each mention with its highest-scoring antecedent. In this graph, each connected component of mentions forms an entity.

4.2 Incremental

The general approach of the incremental model follows Xia et al. (2020) and Toshniwal et al. (2020). Mention filtering is performed as in the course-to-fine model. We process the document iteratively, splitting the document into multiple windows for transformer language model inference. Unlike Toshniwal et al. (2020) but following Xia et al. (2020) we reuse all model weights, including both the transformer weights and all task-specific layers.

In a step we call entity assignment, every mention candidate chooses its entity in an iterative fashion. In our standard setup, this is modeled as a classification task with a dynamic number of classes and the initial set of classes, each class representing an entity $C_0 = \{\emptyset\}$. If, for any mention embedding m being processed, $\emptyset$ is selected as the class, the mention is added as a new class. Entity representations are tracked with $R(E_n)$ being set to m when the n-th entity is added. As a result, after the first mention is processed the set of classes is always extended: $C_1 = \{\emptyset, E_0\}$. Subsequently, new mentions E_x are added iteratively. Whenever any existing E_x is selected as the best fitting entity, its representation is updated using an update gate: $R(E_x) := (1-\alpha)m + \alpha R(E_x)$.

Training is done by means of cross entropy loss across all existing entities and the new entity class, with the gold class for each entity being its most recently assigned mention gold class. As a result, early in training many entity representations likely contain mentions that, from a gold label perspective, should not belong together. Toshniwal et al. (2020) use teacher forcing to address this issue and thereby reach earlier convergence; we test this approach in our setup, assigning each mention to its gold class for further computations, rather than relying on predicted classes.

The only way mention candidates can be discarded (either because they are not a mention or because they are singleton mentions) is by means of creating a new entity and never assigning any additional mentions to it, in postprocessing any such singleton entity would be removed, yielding the final output entities. To support detection of singleton mentions, we follow Xia et al. (2020) in adding an additional class representing the discarding of any given entity. In this "discard" scenario, singleton mentions are not removed in postprocessing since non-mentions are modeled explicitly.

Language Model	CoNLL-F1
BERT-Base, multilingual uncased	74.50
BERT-Base, multilingual cased	74.60
GBERT base, cased	75.35
GELECTRA base, cased	77.01
GNG_ELECTRA base, uncased	**77.86**
GBERT large, cased	79.05
GELECTRA large, cased	**79.24**

Table 2: TüBa-D/Z 10 development score of coarse-to-fine models with different language models (using 512 as segment size)

5 Experiments: News Domain

We perform preliminary experiments to select the best pre-trained German language model, its best context size and to optimize other hyperparameters. For the main experiments on the news datasets TüBa-D/Z 10 and SemEval-2010, we train and evaluate our coarse-to-fine model as it is easily capable of processing the typically rather short documents. We use the training, development and test splits as described in Section 3.1. The SemEval dataset contains singletons, but our coarse-to-fine model predicts only clusters of at least two entities. Following Roesiger and Kuhn (2016), we ignore singletons when scoring our systems' predictions.

5.1 Pre-trained Language Models

We evaluated multiple pre-trained language models for our coreference resolution model. As a baseline, we include the multilingual BERT-Base model (in both the cased and uncased variants) by Devlin et al. (2019). Chan et al. (2020) recently published German BERT and ELECTRA (cased, both base and large) denoted as GBERT / GELECTRA in Table 2. In addition, we included another ELECTRA model (uncased, base) by German-NLP-Group denoted as GNG_ELECTRA[3].

We find that all of the recent German language models perform better than the multilingual BERT. For the base models, ELECTRA outperforms BERT by a substantial margin. Using large models, ELECTRA performs marginally better. Based on the results shown in Table 2, we selected GNG_ELECTRA as the base and GELECTRA as the large model for our remaining experiments.

[3] Model description at https://huggingface.co/german-nlp-group/electra-base-german-uncased

Segment Length	F1 (base)	F1 (large)
128	75.69	76.28
256	76.56	77.29
384	77.01	78.51
512	**77.50**	**79.27**

Table 3: TüBa-D/Z 10 development score of coarse-to-fine models GNG_ELECTRA (base) and GELECTRA (large) with different segment lengths.

5.2 ELECTRA Context Size

Following Joshi et al. (2019), we split documents into non-overlapping ELECTRA contexts, evaluating different splits for contexts as shown in Table 3. While Joshi et al. (2019) show that for English BERT-base/large a segment length of 128/384 is optimal, this does not hold true for our German models and dataset where larger segment lengths perform better. Our results are in line with the intuition that larger context sizes provide more contextual information for any given mention. Thus, we use a segment length of 512 in our models.

5.3 Hyperparameters

In general, parameters affecting computational limits have a large impact, all other parameters that we tested had only limited effect. Parameters controlling the pruning (top_span_ratio, max_top_spans and max_top_antecedents) have a strong negative effect when set too low, resulting in too aggressive pruning. Higher values increase evaluation scores with quickly diminishing returns; yet strongly increase computation time and memory.

To reduce GPU memory usage and computation time, we reduced the size of all feed-forward neural networks from 3000 used in previous work to 2048 without seeing distinct score changes on the TüBa-D/Z 10 development set. We also increased the size to 4096, resulting in more memory usage and slower computation, but negligible changes in evaluation performance.

6 Experiments: Literature Domain

For the literary dataset (DROC), we explore the use of both model variants. We initialize the incremental model with weights from the coarse-to-fine variant.

CoNLL-F1		News-Pretrain	
		✓	✗
Singletons	✓	61.66 ± 0.52	59.93 ± 0.33
	✗	**65.58 ± 0.46**	**64.26 ± 0.51**

Table 4: The effect of using pre-training on the DROC coarse-to-fine model on data with and without singletons. All results were averaged over 5 runs and the standard deviation is given.

6.1 Coarse-to-fine Model

Given the relatively small size of the DROC dataset, we explore the impact of pretrained weights from the news tasks. We expected that while the different approaches to mention annotation (heads or entire noun phrases) would somewhat limit applicability of existing weights they would still lead to an improvement.

Table 4 shows the development set results for the DROC dataset, with the same set of initial weights that was pretrained on TüBa-D/Z 10 being used for all of our runs. Standard deviation for the ConLL-F1 scores are given, based on five runs with different random initializations. All layer weights, including task specific ones as well as language model ones were reused. The experiment was repeated for a variant of the DROC dataset with all singleton mentions removed.

Using Welch's t-test we can infer that the pretrained version does, on average, perform better for the no singleton variant ($p < 0.005$). As a result we will use the news-pretrained model variant in all our further experiments. This finding is also supported by the recent publication by (Xia and Durme, 2021) which establishes that, especially for short datasets, using pretrained weights is beneficial. We are unsure if further significant improvements could be gained by pre-training on additional datasets, for example GerDraCor (Pagel and Reiter, 2020), given that TüBa-D/Z is already a large dataset.

Table 5 shows how two configuration parameters affect the coarse-to-fine model's performance. The two options enable different features, where "segment info" describes how many BERT segments lie between the current and candidate mention while "token info" describes the token distance from the candidate mention to the document start. Further, "token info" encodes the length of the candidate mention span. This experiment was performed as

Distance Features			
Segment	Token	Coarse-To-Fine	Incremental
✗	✗	61.11 ± 0.57	**65.79**
✓	✗	**62.31 ± 0.27**	64.20
✗	✓	61.70 ± 0.22	62.57
✓	✓	59.93 ± 0.33	65.42

Table 5: Performance of the coarse-to-fine and incremental models with respect to two configuration parameters relevant to recency bias.

CoNLL-F1		Teacher Forcing	
		✓	✗
Discard	✓	63.92	**65.42**
	✗	58.52	57.27

Table 6: DROC incremental model configurations

we saw a recency bias in terms of connecting mentions in our early result explorations (see Section 7), an effect that could be caused by these distance based features. On average, the variant without token distance representation performs significantly better than the the one with both features enabled ($p < 0.001$). We attribute this to a greater mention recency bias that is encouraged by the additional features.

6.2 Incremental Model

The memory usage of the coarse-to-fine approach, while not prohibitive for the DROC dataset, will prevent its application to full length literary documents.

Table 5 illustrates the impact of the same configuration parameters that were used for the coarse-to-fine model. The impact of the parameters appears to be lessened in the incremental case.

Unsurprisingly, due to the possibility of handling singleton mentions, Table 6 clearly shows that the discard functionality is critical to model performance. Teacher forcing appears to have a negative impact on performance; this does come as a surprise but while convergence early in training was faster the final results were slightly worse.

6.3 Impact of Document Length

We seek to analyze how well incremental models fare as document length increases. To this end, we split DROC at the nearest sentence boundary into sub documents that are no longer than 512, 1024 and 2048 tokens. Previous work (Krug, 2020; Joshi et al., 2019) has established that, with longer documents, the performance of coreference systems drops. This can be interpreted as the inherent difficulty of the coreference task growing with document length. Figure 3 shows that for longer documents the gap in performance between the model variants increases slightly.

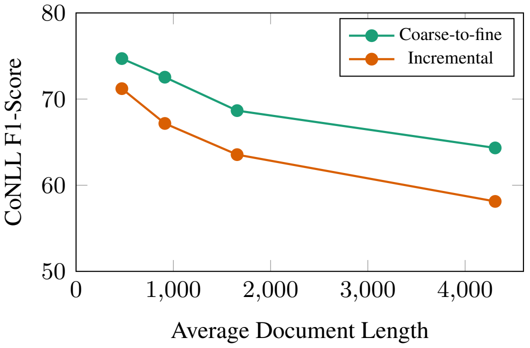

Figure 3: The performance of incremental systems compared to coarse-to-fine model as document lengths increases.

System	CoNLL-F1	
	TüBa	SE'10
German coarse-to-fine base	77.21	72.54
German coarse-to-fine large	**78.79**	**74.46**
IMS HotCoref	48.54	48.61
+ gold mentions	65.76	63.61
CorZu		45.82
+ gold mentions		58.11

Table 7: Results of our coarse-to-fine models and previous systems on the test set of TüBa-D/Z 10 and SemEval-2010 (without singletons). IMS HotCoref and CorZu scores as reported by Roesiger and Kuhn (2016). Full metrics in Table 10 in the appendix.

7 Results & Error Analysis

Our neural coarse-to-fine models outperform the previous state of the art by a large margin on both SemEval-2010 (+25.85 F1) and TüBa-D/Z (+30.25 F1) as shown in Table 7. In fact, even if the other systems are allowed to use gold mentions, our models still outperform them by more than 10 F1 points. Using ELECTRA large for contextual embeddings yields a small improvement over the base model (+1.58 F1 / +1.92 F1). Figure 4 shows an example of our systems prediction on an unseen document.

We manually analyze the predictions of our

[Bahn-Chef]$_{1}$ legt Statistik vor. Bisher keine Erklärung für [das Unglück von [Eschede]$_{3}$]$_{2}$ [Frankfurt]$_{4}$ (taz) - Der Eindruck, daß sich die Unfälle bei [der Bahn]$_{5}$ häuften, sei nur durch die " Berichterstattung der Medien " provoziert, erklärte [der Vorstandsvorsitzende [der Deutschen Bahn AG]$_{5}$, Johannes Ludewig (CDU)]$_{1}$), gestern in [Frankfurt]$_{4}$. Zum bevorstehenden ersten Jahrestag [der ICE-Katastrophe von [Eschede]$_{3}$]$_{2}$ (3. Juni) verwies [Ludewig]$_{1}$ auf [die - bahneigene - Statistik]$_{6}$. [...]

Figure 4: Excerpt from a TüBa-D/Z 10 test set document (in total 438 tokens), where the shown output of our coarse-to-fine large model is identical to the human annotation (document score: 89.08 CoNLL-F1)

System	Model	F1 Score
Krug (2020) (with singletons)	Sieve CR E2E-NN	51.53 51.34 53.17
Ours (with singletons)	Incremental C2F	**64.72** 61.66
Ours (no singletons)	C2F	65.50

Table 8: Final results for the DROC dataset on the test set, with and without singleton mentions included.

coarse-to-fine model and find that it generally produces accurate coreference links both locally and document-wide. While entity assignment of mentions, identified in both prediction and gold data, is typically correct, missed and added mentions are more frequent errors. We assume that one reason is a contradicting training signal, i.e. while some mentions are annotated as such in the gold data, others are not because they are singletons or were missed in the annotation process.

Our incremental model on data including singletons outperforms the existing state of the art for DROC by 11.6 F1 points (see Table 8). Said results were achieved in a setup comparable to ours, with no gold information such as speakers or entity spans being used, except in the case of their end-to-end neural network (E2E-NN), where direct speech and speaker information were used.

We manually evaluate our model on entire literary texts. While we find local coreference relationships to be surprisingly accurate, when taking a more global view, some of our model's weaknesses are exposed. When searching the token "Holmes" in the German translation of "The Hound of the Baskervilles" [5] which should always refer to the same character we find the 212 tokens to occur in 31 different clusters with 4 mentions being assigned to no cluster. Our observation is that this often occurs after a long section of text without explicit mentions of the name, in fact the average distance from one mention of Holmes to the previous is 320.6 tokens whereas it is 655.3 for those cases where a new class is erroneously introduced. We suspect, that this could be attributed to the name taking less prominence in the entity representation after a while.

Es war einmal ein gar allerliebstes, niedliches Ding von einen [Mädchen]$_{1}$, [das]$_{1}$ hatte eine [Mutter]$_{2}$ und eine [Großmutter]$_{2}$, die waren gar gut und hatten das kleine [Ding]$_{1}$ so lieb. Die [Großmutter]$_{2}$ absonderlich, [die]$_{2}$ wußte gar nicht, wie gut sie ' s mit dem [Enkelchen]$_{1}$ meinen sollte [...]

(a) Model with token distance feature

Es war einmal ein gar allerliebstes, niedliches Ding von einen [Mädchen]$_{1}$, [das]$_{1}$ hatte eine [Mutter]$_{2}$ und eine [Großmutter]$_{3}$, die waren gar gut und hatten das kleine Ding so lieb. Die [Großmutter]$_{3}$ absonderlich, [die]$_{3}$ wußte gar nicht, wie gut sie ' s mit dem [Enkelchen]$_{1}$ meinen sollte, [...]

(b) Model without token distance feature

Figure 5: We observe a recency bias that appears to, in this case, be fixed by not including an explicit token distance feature. The term "Großmutter" (grandmother) is linked to the term "Mutter" (mother).[4]

Figure 5a illustrates a recency bias in our model, "grandmother" and "mother" were erroneously combined into one entity, presumably because the distance between the "mother" and "grandmother" mentions were very small. On a larger scale this effect can be observable as long sequences of the same cluster forming, an effect that is especially prominent in our incremental models. This observation motivated our experiments with removing distance features (see Table 5), resulting in an improved model and, in this case (as seen in Figure 5b), an improved result. However, this particular model no longer detects "thing" (Ding) as a

[4] Text from: https://www.projekt-gutenberg.org/bechstei/maerchen/chap053.html

[5] https://www.projekt-gutenberg.org/doyle/basker-1/

valid mention which could both be a side effect of removing the distance features or an effect of the random initialization and training.

8 Conclusion

We apply recent developments in neural architectures for coreference resolution on German data and achieve a substantial improvement over the previous state of the art on all three established German datasets. We conducted experiments with two variants: a coarse-to-fine model suitable for rather short documents, and an incremental model that should scale to long documents. In our analysis we found that while the task of coreference resolution itself becomes more difficult as document sizes increase, the incremental approach scales worse than the course-to-fine approach in terms of accuracy. While we found local decisions to be accurate, shortcomings of the incremental model in global consistency and recency bias were explored.

In future work, we would especially like to address remaining challenges for the processing of long-form literary documents. In spite of the large improvements we achieved, there is still a considerable headroom for coreference resolution, as reflected by a large performance gap between the human baseline of 82.54 F1 and our best model with 64.7 F1 on the DROC dataset. On a more theoretic note, another extension worth pursuing in the future especially for the literary domain is the notion of subjective coreference. As an example, in the fairy tale “Little Red Riding Hood” (see Figure 5), the girl temporarily perceives a highly plot-relevant coreference between the grandmother and the big bad wolf, which is not reflected in objectivized models.

Acknowledgments

This work was, in part, supported by the DFG through the project “Evaluating Events in Narrative Theory (EvENT)” (grants BI 1544/11-1 and GI 1105/3-1) as part of the priority program “Computational Literary Studies (CLS)” (SPP 2207). This work was partly supported by the Cluster of Excellence CLICCS (EXC 2037), Universität Hamburg, funded through the German Research Foundation (DFG).

References

David Bamman, Olivia Lewke, and Anya Mansoor. 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.

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 (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 47–57, Baltimore, Maryland. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German’s next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

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

Manfred Klenner and Don Tuggener. 2011. An Incremental Entity-Mention Model for Coreference Resolution with Restrictive Antecedent Accessibility. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 178–185, Hissar, Bulgaria. Association for Computational Linguistics.

Markus Krug. 2020. *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Ph.D. thesis, Universität Würzburg.

Markus Krug, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe, and Fotis Jannidis. 2018. Description of a corpus of character references in German novels-DROC [Deutsches ROman Corpus]. *DARIAH-DE Working Papers*, 27.

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*, pages 28–34, Portland, Oregon. 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.

Thomas S. Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*, pages 85–89, College Park, Maryland. Association for Computational Linguistics.

Karin Naumann and Vera Möller. 2006. Manual for the annotation of in-document referential relations. Technical report, Universität Tübingen.

Janis Pagel and Nils Reiter. 2020. GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 55–64, Marseille, France. European Language Resources Association.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

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

Ina Roesiger and Jonas Kuhn. 2016. IMS HotCoref DE: A Data-driven Co-reference Resolver for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 155–160, Portorož, Slovenia. European Language Resources Association (ELRA).

Ina Roesiger, 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, Santa Fe, New Mexico. Association for Computational Linguistics.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6):1663–1680.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2017. Stylebook for the tübingen treebank of written german (TüBa-D/Z). In *Seminar für Sprachwissenschaft, Universität Tübingen, Germany*.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 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.

Don Tuggener and Manfred Klenner. 2014. A Hybrid Entity-Mention Pronoun Resolution Model for German Using Markov Logic Networks. In *Proceedings of the 12th Edition of the Konvens Conference*, pages 21–29, Hildesheim, Germany. Universitätsbibliothek Hildesheim.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. OntoNotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, Pennsylvania*, 23.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China. Association for Computational Linguistics.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. *CoRR*, abs/2104.08457.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online. Association for Computational Linguistics.

A Appendix

dataset	articles	sentences	tokens
SemEval-2010	1,235	26,098	455,046
- training	900	19,233	331,614
- develop.	199	4,129	73,145
- test	136	2,736	50,287
TüBa-D/Z 10.0	3,644	95,595	1,787,801
- training	2190	65,416	1,258,514
- develop.	727	15,593	276,635
- test	727	14,586	252,652
TüBa-D/Z 11.0	3,816	104,787	1,959,474
OntoNotes 5.0	3,493	94,269	1,631,995
DROC	90	18,161	393,164
- training	58	11,368	249,817
- develop.	14	3,570	72,258
- test	18	3,223	70,999

Table 9: Overview of the dataset releases referred to in this work.

	MUC			B^3			$CEAF_{\phi_4}$			CoNLL	LEA		
	R	P	F1	R	P	F1	R	P	F1	F1	R	P	F1
TüBa-D/Z 10.0													
German c2f base	81.92	79.90	80.90	77.41	73.52	75.41	75.16	75.50	75.33	77.21	74.98	70.82	72.84
German c2f large	82.85	81.61	82.23	78.41	75.73	77.05	76.75	77.44	77.09	**78.79**	73.25	73.25	74.67
IMS HotCoref										48.54			
+ gold mentions										65.76			
SemEval-2010													
German c2f base	76.64	76.08	76.36	71.18	69.12	70.14	71.83	70.45	71.13	72.54	67.88	65.7	66.77
German c2f large	79.07	76.51	77.77	73.88	70.48	72.14	74.79	72.21	73.47	**74.46**	70.69	67.18	68.89
IMS HotCoref										48.61			
+ gold mentions										63.61			
CorZu										45.82			
+ gold mentions										58.11			

Table 10: Recall, precision and F1 score on the test set of TüBa-D/Z 10 and SemEval-2010 (without singletons). Our coarse-to-fine (c2f) models use either ELECTRA base or large. IMS HotCoref and CorZu system scores as reported by Roesiger and Kuhn (2016).