Efficient and Effective Coreference Resolution for German

Fynn Petersen-Frey and Hans Ole Hatzel and Chris Biemann
Universität Hamburg
Language Technology Group & Hub of Computing and Data Science
{fynn.petersen-frey,hans.ole.hatzel,chris.biemann}@uni-hamburg.de

Abstract

Coreference resolution is the task of identifying and clustering text spans referring to the same entities. While many current models provide good results for short texts, they are computationally inefficient for long documents, and their quality degrades with longer documents. In this paper, we propose a new German coreference resolution model improving on the state of the art while being computationally efficient even for long documents. We perform experiments on multiple German datasets and ablations across different model variations. Our source code and trained model weights are available on Github and Huggingface. Further, we provide a ready-to-use Python library to perform coreference resolution on German texts with minimal setup.

1 Introduction

Coreference occurs when two or more phrases, called mentions, within a text refer to the same entity, such as a person, object, or event. A mention refers back to a previous text span, called its antecedent, which references the same entity. Coreference resolution deals with resolving these references by connecting each mention with its antecedent. In more technical terms, it is the task of identifying and clustering mentions in the form of text spans in a document referring to an entity. All mentions referring to the same entity need to be assigned to the same cluster to build mention clusters where each cluster represents exactly one entity. Figure 1 shows a short example with resolved coreference annotations.

The task of coreference resolution is a widely applicable foundational part of a language processing pipeline to enable or improve other tasks such as question answering, text summarization, quotation attribution, or aid in literary analysis, discourse analysis, etc.

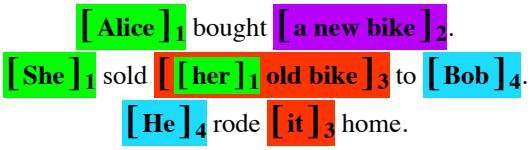

Figure 1: Visualization of coreference mentions and cluster information using indexed brackets.

While early models relied on rule-based approaches, current state-of-the-art systems are built as end-to-end neural models leveraging large pretrained language models. As a result, the quality of the system's outputs has drastically improved over the last years. However, the computational demands have also sharply increased, rendering many approaches impractical for use with large document collections, especially with longer documents, as runtimes typically increase superlinearly. The bulk of the research has focused on English as a target language, with comparatively few approaches targeting German.

In this work, we adapt an existing approach for English coreference resolution to German. We make the following contributions:

- New state-of-the-art coreference resolution models for German news and literary texts
- that are also computationally efficient on long texts.[1]
- Ablation experiments across various model parameters and design choices.
- A thorough evaluation of our models by comparison with existing models and an error analysis.
- A ready-to-use Python library to perform coreference resolution on pre-tokenized or plain text.[2]

2 Related Work

Rule-based approaches have long been used to address the task of coreference resolution (Raghu-

[1] https://github.com/uhh-lt/maverick-coref-de
[2] https://pypi.org/project/maverick-coref-de

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 128–137, 2025

nathan et al., 2010; Lee et al., 2011). Advancements in neural architectures, specifically encoder models, have enabled end-to-end neural systems which generally follow mention linking approaches but differ in architectural details and in how they represent mentions (Lee et al., 2017, 2018; Dobrovolskii, 2021). In recent years, text-to-text approaches have increasingly been used to essentially perform inline coreference annotations (Wu et al., 2020; Bohnet et al., 2023; Zhang et al., 2023). While such text-to-text approaches yield promising results, recent advances in encoder-based approaches have caught up in terms of resolution quality while offering vastly quicker runtime performance (Martinelli et al., 2024). Recent work cast doubts on whether coreference resolution architectures have improved over time, attributing the gradual improvement mainly to the underlying model's sizes instead (Porada et al., 2024).

In terms of datasets, there are three German ones that are relevant to this work: (1) TüBa-D/Z (Telljohann et al., 2004), (2) SemEval 2010 (Recasens et al., 2010), and (3) DROC (Krug et al., 2018). While the former two both contain news articles, DROC contains annotated excerpts from literary works. The excerpts are, on average, around 4,000 tokens in length, which presents a challenge for traditional encoder models that often handle only around 500 sub-word input tokens.

Rule-based approaches have also been applied to coreference resolution in German (Krug et al., 2015; Tuggener, 2016). Roesiger and Kuhn (2016) presented a tree-based machine learning approach using handcrafted features. More recently, we (Schröder et al., 2021) contributed a mention-linking-based encoder model for coreference resolution, the current state-of-the-art model for TüBa-D/Z. Bohnet et al. (2023) test a transition-based coreference system on, among other languages, German. Gupta et al. (2024) focus specifically on literary texts, contributing a novel technique for handling texts with lengths that exceed the underlying encoder model's context.

3 Model

Our model is based on Maverick (Martinelli et al., 2024), a highly-efficient, state-of-the-art coreference resolution model for English. It creates token representations using a transformer encoder, extracts mentions using two classifiers, and finally clusters the mentions by computing the probability of any two mentions belonging to the same cluster by computing the sum of multiple bilinear functions using the mentions' start and end representations. For a more thorough explanation of the Maverick architecture, additionally refer to Martinelli et al. (2024).

3.1 Mention Extraction

While the coarse-to-fine model (Lee et al., 2018) and its derivatives enumerate all spans up to a certain length as possible mentions and select the top k for the mention clustering step, the Maverick system uses a more efficient approach. Mentions are extracted by first identifying all possible mention starts using a feed-forward neural network classifier on each token representation $\mathbf{x_i}$ produced by a transformer encoder for all tokens $t_i, \ldots, t_n$ in a document

$$F_{start}(x) = \mathbf{W}'_{start} \cdot \mathrm{LN}(\mathrm{GeLU}(\mathrm{DO}(\mathbf{W}_{start}x)))$$
$$p_{start}(t_i) = \sigma(F_{start}(x_i))$$

with $\mathbf{W}'_{start}, \mathbf{W}_{start}$ being the learnable parameters, and dropout function (DO), GeLU activation, layer norm (LN) and σ the logistic sigmoid function.

For each plausible mention start t_s, i.e. tokens with $p_{start}(t_s) > 0.5$, span representations are created for all subsequent tokens t_j until the end of the sentence as the concatenation of x_s, and end, x_j. Another feed-forward neural network classifier is used to identify the likely mentions among all candidates

$$F_{end}(x) = \mathbf{W}'_{end} \cdot \mathrm{LN}(\mathrm{GeLU}(\mathrm{DO}(\mathbf{W}_{end}x)))$$
$$p_{end}(t_j|t_s) = \sigma(F_{end}([x_s, x_j]))$$

with $\mathbf{W}'_{end}$, $\mathbf{W}_{end}$ being learnable parameters. All candidate spans having $p_{end}(t_j|t_s) > 0.5$ are selected as the final mentions for the mention clustering step.

3.2 Mention Clustering

The Maverick model (Martinelli et al., 2024) uses the mention clustering of the LingMess model (Otmazgin et al., 2023), which in turn is an extension of the s2e model (Kirstain et al., 2021). We first briefly describe the mention clustering of the s2e model as implemented in Maverick and our model adoption. A mention $m = (\mathbf{x}_s, \mathbf{x}_e)$, consisting of the start $\mathbf{x}_s$ and end $\mathbf{x}_e$ token hidden states, is transformed into into a start and end representation using two fully-connected layers with

$\mathbf{W}_s, \mathbf{W}'_s, \mathbf{W}_e, \mathbf{W}'_e$ as the learnable parameters:

$$F_s(x) = \mathbf{W}'_s \cdot \mathrm{LN}(\mathrm{GeLU}(\mathrm{DO}(\mathbf{W}_s\mathbf{x})))$$
$$F_e(x) = \mathbf{W}'_e \cdot \mathrm{LN}(\mathrm{GeLU}(\mathrm{DO}(\mathbf{W}_e\mathbf{x})))$$

The probability p_a of the mention m_i to be an antecedent of mention m_j (in other words: to belong to the same cluster) is computed with a logistic sigmoid function σ over the antecedent logits

$$\begin{aligned} p_a(m_i, m_j) = \sigma(&F_s(x_s) \cdot \mathbf{W}_{ss} \cdot F_s(x_{s'}) + \\ &F_e(x_e) \cdot \mathbf{W}_{ee} \cdot F_e(x_{e'}) + \\ &F_s(x_s) \cdot \mathbf{W}_{se} \cdot F_e(x_{e'}) + \\ &F_e(x_e) \cdot \mathbf{W}_{es} \cdot F_s(x_{s'})) \end{aligned}$$

with $\mathbf{W}_{ss}, \mathbf{W}_{ee}, \mathbf{W}_{se}, \mathbf{W}_{es}$ being four learnable parameter matrices.

The LingMess model (Otmazgin et al., 2023) uses multiple mention-antecedent scorers instead of only the single generic scorer found in the s2e model. The scorers handle different linguistically motivated categories that any pair of mentions m_i and m_j belongs to: PRON-PRON-C, PRON-PRON-NC, ENT-PRON, MATCH, CONTAINS, OTHER. For example, in case m_i is a pronoun and m_j is a proper noun, the ENT-PRON scorer is used to compute p_a. The final score of two mentions is the sum of the generic, shared scorer (like in the s2e model) and the expert scorer for the specific category. During training, only the shared scorer and the relevant expert scorer receive gradient updates from the loss function.

3.3 Major Changes from Maverick

While we experimented with numerous architectural changes to the model (see Section 4.1), there are three major changes in our final model.

First, in contrast to all previous encoder-based end-to-end neural coreference architectures, Maverick uses the binary cross-entropy loss with teacher forcing instead of the complex log-marginalized loss function. The log-marginalized loss function was used due to the computationally required pruning of mention and antecedent candidates. We additionally eliminated the need for teacher forcing by limiting the number of mentions predicted early in training to a computationally feasible amount.

Second, instead of the English DeBERTaV3 (He et al., 2021) transformer model, we use the recent ModernGBERT (Ehrmanntraut et al., 2025) model, a transformer trained from scratch on German data based on the ModernBERT architecture (Warner et al., 2024) that allows highly efficient processing of long input sequences.

Third, the Maverick model on English has six linguistically motivated categories for the mention-antecedent scorers. While we adapted the built-in lists of pronouns to German, we could not use all categories due to the lexical ambiguity of German pronouns. Thus, we combined the categories PRON-PRON-C and PRON-PRON-NC for dealing with (in)compatible pronoun mentions based on gender, number, and animacy into a PRON-PRON category for all pronoun-pronoun pairs.

4 Experiments

After defining the evaluation metrics and datasets, we first describe preliminary experiments to select the best model design and training procedure. Subsequently, we detail our main experiments on news and literary datasets to create a total of five models.

Evaluation metrics The de facto standard metric for evaluating coreference is the CoNLL-F1 score, the mean of the MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998) and $CEAF_{\phi_4}$ (Luo, 2005) F1 scores. The LEA metric (Moosavi and Strube, 2016) is a more recent metric that addresses many shortcomings of the above metrics. In this work, we evaluate and compare our models using both the CoNLL-F1 and, whenever possible, LEA.

Datasets We perform experiments on the three datasets introduced in Section 2: TüBa-D/Z release 10.0 (TüBa10) using the same split as in our previous work (Schröder et al., 2021), the German part of the SemEval 2010 challenge (SE10), and DROC. TüBa10 is a large dataset covering 1.8 million tokens. SE10 and DROC are much smaller, with 455,000 and 393,000 tokens, respectively. While TüBa10 and SE10 contain thousands of documents, DROC only contains 90 (albeit much longer documents).

4.1 Preliminary Model Design Experiments

To find the best model for our main experiments, we performed a number of preliminary tests to evaluate model variations with foundation language models, teacher forcing, and mention clustering scorers. We considered both model performance and runtime characteristics as we are striving for an efficient and effective model.

Foundational language model We evaluated three language models to replace the English

Foundation	CoNLL-F1
mdeberta-v3-base	76.7
LLäMmlein2Vec-1B	46.7
ModernGBERT-1B	81.4

Table 1: Performance on the TüBa10 development set of different foundational language models

DeBERTa-V3 model. As the model architecture is designed around a language model supporting contexts long enough to process an entire document at once, we were limited to very few options. None of the foundation models used in our previous work (Schröder et al., 2021) are applicable, so a direct comparison is not possible. Instead, we evaluated three models with support for the German language and processing of long contexts. First, we tested the multilingual variant of DeBERTa V3, with promising results of 76.7 CoNLL-F1 for a small, multilingual model – albeit having a slow inference speed. A larger LLM2Vec (BehnamGhader et al., 2024) encoder created from LLäMmlein (Pfister et al., 2024) was faster but produced subpar results of only 46.7 CoNLL-F1. Finally, the ModernGBERT (Ehrmanntraut et al., 2025) model showed strong results with 81.4 CoNLL-F1 while being fast to compute. Table 1 shows a comparison of the CoNNL-F1 score on the TüBa10 development set. Thus, we choose ModernGBERT-1B[3] as our foundation.

Teacher forcing In preliminary experiments, we compared teaching forcing with standard training on the TüBa10 dataset. While training runs with teaching forcing quickly reduced training loss and increased development scores, training without it eventually but consistently led to slightly better scores on the development set. Thus, we deviate from the original Maverick training procedure and use standard training for TüBa10 and SE10.

Attention for mention representation Our model represents mentions using their start token and end token representation as produced by the language model. We experimented with adding a third component, an attention-weighted average of all tokens that are part of a mention span. In preliminary experiments on the TüBa10 dataset, we saw no consistent quality difference on the development set when comparing models with/without our additional span representation. As the additional representations lead to increased computation, as evidenced by slower training and inference, it results in decreased efficiency. Consequently, we kept the pure start and end token representation of mentions without the additional representation in our models.

Mention clustering scorers We compared a single generic mention clustering scorer with the five specialized scorers in preliminary experiments on the TüBa10 dataset. Using the additional scorers consistently leads to better CoNLL-F1 scores on the development set: ≈ 78 versus ≈ 81. Thus, we decided to use the stronger, albeit more complex model, as it only requires a small amount of additional compute resources compared to the overall required compute.

4.2 Main Experiments

News articles We train three models independently on TüBa10 and SE10. For the latter, we train two variants: with or without singletons. In each setup, we use the best model configuration found in our preliminary experiments. We use roughly the same hyperparameter settings as Martinelli et al. (2024): Different learning rate of 2×10^{-4} for the head and 1×10^{-5} for the language model, learning rate warm-up over 3 epochs, weight decay of $1e-3$, and up to 300 epochs limited by early stopping.

Literary texts For our DROC experiments, we follow the training setup proposed in our previous work (Schröder et al., 2021), finetuning first on TüBa10 and subsequently on the much smaller DROC. Additionally, we randomly reinitialize the head that creates mention probabilities given the start and end token embeddings. This change is motivated by the fact that DROC's annotations for mention spans differ from those in TüBa10 and that we, likely as a result, faced convergence issues when training from checkpoints trained on TüBa10. We rely on the model's capability to generalize to context lengths that were not seen in pre-training and only perform one forward pass per document. Unlike TüBa10, DROC only considers mentions of characters rather than all entities; we rely on the additional fine-tuning to instill this difference in the model.

5 Results & Error Analysis

Table 2 contains the results of our main experiments, the test set F1 scores across all evaluated

[3] https://huggingface.co/LSX-UniWue/ModernGBERT_1B

	$CEAF_{\phi_4}$	MUC	B^3	CoNLL	LEA
			TüBa10		
Ours	80.32	84.64	80.45	81.80	78.41
C2F	77.09	82.23	77.05	78.79	74.67
			SE10 with singletons		
Ours	89.28	79.67	89.81	86.26	82.65
MLA*				86.40	
			SE10 without singletons		
Ours	76.14	80.20	75.37	77.23	72.98
C2F	73.47	77.77	72.14	74.46	68.89
LA*				77.80	
			DROC with singletons		
Ours	58.14	91.80	71.33	73.75	67.77
C2F				61.66	
Incr.				64.72	
			DROC without singletons		
Ours	57.36	91.15	69.69	72.73	68.66
C2F				65.50	

* uses additional training data & vast compute resources

Table 2: F1 Performance of our system on the test sets of TüBa10, SE10 and DROC (with and without singletons) compared to our previous coarse-to-fine (C2F) and incremental (Incr.) models (Schröder et al., 2021) as well as the mention-link-append (MLA) & link-append (LA) models by Bohnet et al. (2023).

datasets. Our model shows a strong performance increase, compared to our coarse-to-fine (C2F) model (Schröder et al., 2021), of 3 CoNLL-F1 and almost 4 LEA-F1 points on the TüBa10 dataset. On SE10, our model almost ($\Delta < 0.6$ CoNLL-F1 points) reaches the results of the sequence-to-sequence models by Bohnet et al. (2023) while being orders of magnitude faster to compute[4]. Compared to the C2F model, our model achieves an increase of approximately 3 CoNLL-F1 and 4 LEA-F1 points. The largest improvement is seen on DROC, where our model outperforms our previous state-of-the-art models (Schröder et al., 2021) by 9 F1 points, reaching a 73.75 CoNLL-F1 score (in the setup including singletons). This is a substantial improvement in the context of coreference resolution, where the rate of improvement has decreased in recent years. We attribute this improvement to the model's capability to handle the long DROC documents in one forward pass rather than slicing the texts into segments and performing individual forward passes. In such a setting, our model architecture is most favorable over the C2F model while still offering competitive runtime performance.

In a qualitative error analysis (see Appendix A for two examples), we find that many errors in news texts are caused by mentions that are not annotated as such in the gold data. Moreover, during training, the model is punished for those arguably correct predictions, leading to recall issues elsewhere. We suspect that major improvements may require data curation or active learning techniques. For literary texts, a different picture presents itself. The model especially struggles with long coreference connections, but at times also fails at more local resolution tasks, presumably due to the relative complexity of literary texts. We also observe an issue that is presumably caused by the difference in mention conventions across datasets: our model sometimes predicts very long mentions, which are common in TüBa10 but not seen in DROC.

Inference efficiency On the DROC test set, inference using our old C2F model (Schröder et al., 2021) takes about 6 seconds while the incremental model takes 1 minute and 4 seconds. Our new approach, on the other hand, takes roughly 36 seconds (all measured on a single A6000 GPU).

6 Conclusion

We present new state-of-the-art models for coreference resolution on German news and literary texts that are efficient to compute for practical applications. Our models substantially improve over the previous state-of-the-art models on the TüBa/DZ (+3 F1) and DROC (+9 F1) datasets. Further, our models are comparable in performance on the SemEval-2010 dataset to the current state of the art – by models that use additional training data and are impractically slow to compute, even with the required vast computational resources (8 TPUs-v4 for inference, 128 TPUs-v4 for training). We provide our models alongside a ready-to-use Python library, making them easy to integrate into research software. In the future, we want to combine our models with document slicing and merging techniques like hierarchical coreference resolution (Gupta et al., 2024) to process book-length documents.

Limitations

We do not see any major limitations in our models. While it might not always produce perfectly correct results according to the human gold labels, it

[4]While we cannot reliably provide the exact speedup, we estimate a factor between 80 and 325 given the use of our single A6000 with 150 bfloat16 TFLOPS taking 36 seconds on 71,000 tokens versus their 8 TPUs-v4 with 275 bloat16 TFLOPS per TPU taking 30 minutes inference on 160,000 tokens depending on unknown implementation details such as parallelizability across TPUs.

certainly manages to produce highly usable, reasonable results that likely outperform the coreference resolution quality of a non-expert annotator, since annotating coreference chains is a challenging task, even for humans.

For documents longer than the intended encoder context length (8192 tokens for ModernGBERT), we rely on the model to still perform accurate predictions. An alternative is to slice a document into chunks shorter than the maximal encoder context length – preferably at sentence ends. This approach may yield improved performance.

Although our model is highly efficient computationally compared to most other high-quality coreference resolution models, there is a limit on the absolute document length due to memory and compute constraints on GPUs. While the mention extraction step scales well with long documents (linear computational complexity), the mention clustering has quadratic runtime and space complexity depending on the number of extracted mentions. As such, it might not be feasible to use our model on whole books, unless mentions are made scarce, by – for example – limiting them to only the characters in a story instead of all linguistically possible mentions.

Ethical considerations

Purely relying on machine learning models to solve a task always introduces room for issues. The models may exhibit (unknown) biases based on their design or inherent in their training data. The foundation models used in our work have been trained on a vast amount of texts that likely have not been verified manually. However, our models were trained on datasets that were manually annotated and curated (Telljohann et al., 2004; Krug et al., 2018). During manual evaluation of our model's predictions on texts from the datasets as well as new texts, we did not encounter any obvious biases, such as missing or wrong predictions based on gender, etc. Coreference models have been known to exhibit biases towards gender and other stereotypes; we do not attempt to resolve such biases in this work (Zhao et al., 2018). Automation with machine learning can enable misuse, although clearly harmful application fields for coreference resolution are not obvious to us. Downstream usage of our models will need to consider potential biases in the model whenever operating on the data from our automated approach.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, Philadelphia, Pennsylvania, USA.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. ModernGBERT: German-only 1b encoder model trained from scratch. *Preprint*, arXiv:2505.13136.

Talika Gupta, Hans Ole Hatzel, and Chris Biemann. 2024. Coreference in long documents using hierarchical entity merging. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 11–17, St. Julians, Malta. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

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

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, USA. 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, USA. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 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)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2024. LLäMmlein: Compact and competitive German-only language models from scratch. *Preprint*, arXiv:2411.11171.

Ian Porada, Xiyuan Zou, and Jackie Chi Kit Cheung. 2024. A controlled reevaluation of coreference resolution models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 256–263, Torino, Italia. ELRA and ICCL.

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*, pages 492–501, Cambridge, Massachusetts, USA. 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).

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 2021)*, pages 170–181, Düsseldorf, Germany. KONVENS 2021 Organizers.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, page 45–52, Columbia, Maryland, USA. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

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.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

A Qualitative Error Analysis

0 **Berliner Vogel stoppt [ICE aus Bonn]$_1$**
0 **[Berliner Vogel]$_4$ stoppt [ICE aus [Bonn]$_5$]$_1$**

1 Ein Vogel hat am Samstag nachmittag in Berlin **[den aus Bonn kommenden ICE 846]$_1$** gestoppt.
1 **[Ein Vogel]$_4$** hat am Samstag nachmittag in Berlin **[den aus [Bonn]$_5$ kommenden ICE 846]$_1$** gestoppt.

2 Wie der Bundesgrenzschutz mitteilte, war nach einem betrieblichen Halt im Bahnhof **[Charlottenburg]$_2$** die Oberleitung gerissen.
2 Wie der Bundesgrenzschutz mitteilte, war nach einem betrieblichen Halt im **[Bahnhof Charlottenburg]$_2$** **[die Oberleitung]$_6$** gerissen.

3 Nach Angaben **[der Bahn AG]$_3$** hatte sich ein Vogel in der Oberleitung verhakt.
3 Nach Angaben **[der Bahn AG]$_3$** hatte sich **[ein Vogel]$_4$** in **[der Oberleitung]$_6$** verhakt.

4 Weitere Einzelheiten sind noch nicht geklärt.
4 Weitere Einzelheiten sind noch nicht geklärt.

5 Im Reiseverkehr kam es laut **[Bahn AG]$_3$** zu Verspätungen von 30 bis 40 Minuten.
5 Im Reiseverkehr kam es laut **[Bahn AG]$_3$** zu Verspätungen von 30 bis 40 Minuten.

6 Nach Angaben **[der Bahn AG]$_3$** wurde niemand verletzt.
6 Nach Angaben **[der Bahn AG]$_3$** wurde niemand verletzt.

7 Die Reisenden konnten auf einen anderen ICE umsteigen.
7 Die Reisenden konnten auf einen anderen ICE umsteigen.

8 Die Züge zwischen den Berliner Bahnhöfen Zoo und **[Charlottenburg]$_2$** konnten nur noch eingleisig fahren.
8 Die Züge zwischen den Berliner Bahnhöfen Zoo und **[Charlottenburg]$_2$** konnten nur noch eingleisig fahren.

9 Die Regionalzüge wurden umgeleitet.
9 Die Regionalzüge wurden umgeleitet.

Figure 2: Visualization produced by our analysis tool for a single document from the TüBa10 test comparing the gold annotations with the prediction of our final model. The document is shown in pairs of sentences where the first sentence is from the gold annotations and the second is the system's prediction. In the gold annotations, multiple mentions of a bird *Vogel* (cluster 4), the city of *Bonn* (cluster 5), and a reference to the overhead wire *Oberleitung* (cluster 6) were missed. Another tiny detail is the first mention of the train station *Charlottenburg* in sentence 2. While the gold annotations only annotated the word *Charlottenburg*, our system's prediction span the whole noun phrase *Bahnhof Charlottenburg* which is arguably correct since the second mention of it in sentence 8 explicitly refers to the train stations *Berliner Bahnhöfen Zoo und Charlottenburg* and not to the locality of Berlin also called *Charlottenburg*.

4 [Ich]$_2$ erfuhr unterwegs, daß [du]$_3$ eilig nach einem [Wundarzt]$_4$ geschickt habest – [ich]$_2$ will doch nicht hoffen, daß [dir]$_3$ oder den [Kindern]$_5$ etwas zugestoßen ist?
4 [Ich]$_2$ erfuhr unterwegs, daß [du]$_3$ eilig nach einem [Wundarzt]$_{44}$ geschickt habest – [ich]$_2$ will doch nicht hoffen, daß [dir]$_3$ oder den [Kindern]$_{45}$ etwas zugestoßen ist?

5 Ruhig sagte [Erna]$_3$: Weder die [Kinder]$_5$ noch [ich]$_3$ bedürfen [seines]$_4$ Beistandes, aber einen [Freund]$_1$ von [uns]$_6$ hat nicht weit von [unserm]$_6$ Hause ein Unfall betroffen, der zwar, dem Himmel sei Dank, nicht bedeutend ist, aber dessen Behandlung denn doch [meine]$_3$ wenigen medicinischen Kenntnisse übersteigt.
5 Ruhig sagte [Erna]$_3$: Weder die [Kinder]$_{45}$ noch [ich]$_3$ bedürfen [seines]$_2$ Beistandes, aber einen [Freund]$_2$ von [uns]$_{46}$ hat nicht weit von [unserm]$_{46}$ Hause ein Unfall betroffen, der zwar, dem Himmel sei Dank, nicht bedeutend ist, aber dessen Behandlung denn doch [meine]$_3$ wenigen medicinischen Kenntnisse übersteigt.

...

19 Alle [seine]$_1$ Einwendungen wurden jedoch durch den Ausspruch des [Arztes]$_4$, daß [er]$_1$ ohne Gefahr der Verschlimmerung [seines]$_1$ Zustandes nicht hinweg gebracht werden könne, widerlegt, und da [Erna]$_3$ [ihn]$_1$ mit der [ihr]$_3$ eigenen, lieblichen, herzgewinnenden Weise bat, [sich]$_1$ doch zu gedulden, und [ihr]$_3$ die Freude zu gönnen, durch [ihre]$_3$ Pflege zu [seiner]$_1$ Erholung beitragen zu dürfen, auch [Linovsky]$_2$, nachdem [er]$_2$ [sich]$_2$ etwas gesammelt hatte, eine etwas wohlwollendere Außenseite zeigte, als vorher, so ergab [er]$_1$ [sich]$_1$, wiewohl ein unglückweissagendes Gefühl in [seiner]$_1$ Seele [ihn]$_1$ mahnte, daß schleunige Flucht heilsamer für [ihn]$_1$ als Bleiben sei.
19 Alle [seine]$_1$ Einwendungen wurden jedoch durch den Ausspruch des [Arztes]$_4$, daß [er]$_1$ ohne Gefahr der Verschlimmerung [seines]$_1$ Zustandes nicht hinweg gebracht werden könne, widerlegt, und da [Erna]$_3$ [ihn]$_1$ mit der [ihr]$_3$ eigenen, lieblichen, herzgewinnenden Weise bat, [sich]$_1$ doch zu gedulden, und [ihr]$_3$ die Freude zu gönnen, durch [ihre]$_3$ Pflege zu [[seiner]$_1$ Erholung beitragen zu dürfen, auch [Linovsky]$_2$, nachdem [er]$_2$ [sich]$_2$]$_2$ etwas gesammelt hatte, eine etwas wohlwollendere Außenseite zeigte, als vorher, so ergab [er]$_1$ [sich]$_1$, wiewohl ein unglückweissagendes Gefühl in [seiner]$_1$ Seele [ihn]$_1$ mahnte, daß schleunige Flucht heilsamer für [ihn]$_1$ als Bleiben sei.

Figure 3: Visualization produced by our analysis tool for part of a single document from the DROC test set (an excerpt from *Erna* by *Charlotte von Ahlfeld*). As above, the document is shown in pairs of sentences where the first sentence is from the gold annotations and the second is the system's prediction. The surgeon (*Wundarzt*, cluster 4) is, many sentences later, referred to as the physician (*Arzt* in sentence 19). The model fails to capture this relationship. The system also fails to resolve the personal pronoun *seines* in sentence 5, which also refers to the surgeon; this example illustrates the model's difficulty with coreference resolution in literary texts, even in a local context. Another effect we can observe is that the model sometimes predicts much longer mentions than the dataset typically contains (see sentence 19); we suspect that this is caused by the pre-training on TüBa10, but note that we did reset the mention heads.