[Lions: 1] and [Tigers: 2] and [Bears: 3], Oh My! Literary Coreference Annotation with LLMs

Rebecca M. M. Hicke
Department of Computer Science
Cornell University
rmh327@cornell.edu

David Mimno
Department of Information Science
Cornell University
mimno@cornell.edu

Abstract

Coreference annotation and resolution is a vital component of computational literary studies. However, it has previously been difficult to build high quality systems for fiction. Coreference requires complicated structured outputs, and literary text involves subtle inferences and highly varied language. New language-model-based seq2seq systems present the opportunity to solve both these problems by learning to directly generate a copy of an input sentence with markdown-like annotations. We create, evaluate, and release several trained models for coreference, as well as a workflow for training new models.

1 Introduction

Coreference annotation and entity recognition are key tasks for performing a wide variety of textual analyses. They provide important information about texts as well as serving as the foundation for many more complicated forms of analysis. Particularly within the digital humanities (DH), these tasks are often essential for performing large-scale studies of corpora (e.g. Underwood et al. (2018); Papalampidi et al. (2019); Brahman and Chaturvedi (2020)). However, coreference annotation is considerably more difficult than many binary classification tasks. First, coreference requires nuanced understanding of text, which has been beyond the capabilities of previous NLP. Second, coreference requires structured output, such as marking spans for entity mentions and coreferent mentions, which has previously required custom software.

Generative large language models (LLMs) have recently demonstrated a capacity to solve both problems (Bohnet et al., 2023; Zhang et al., 2023). By leveraging massive pretraining collections and billions of parameters, they can identify the subtle, nuanced patterns of language. In addition, they can generate text that matches specific text markup formats. This capability suggests that non-expert users may be able to use "out of the box" LLMs to generate complicated marked-up text simply by providing examples of the desired input and output. While we evaluate this process by comparing with existing custom-built coreference systems, we emphasize that the potential impact of this process extends to a much broader class of markup.

To explore the promise of fine-tuning generative LLMs for coreference annotation, we evaluate the capabilities of several models to perform coreference annotation on sentences extracted from literary texts. Previous research has shown that literary texts have unique characteristics (Bamman et al., 2020) that make it difficult to adapt generalized NLP models to literary settings. Zhang et al. (2023) achieve high performance on the LitBank corpus when data from the corpus is included in the fine-tuning dataset; we seek to further explore the capabilities of a model adapted specifically for literary coreference.

In this work, we find that a fine-tuned `t5-3b` model significantly outperforms a state-of-the-art neural model for literary coreference annotation (Otmazgin et al., 2022). In addition, we speculate on the ability of these models to perform more complicated, abstract annotation tasks (e.g. identifying character relationships) given its performance on this task.

Specifically, in this work we contribute:

- A high-performing fine-tuned LLM and supporting code that can be used to perform coreference annotation on literary data.[1]
- An analysis of which LLMs are best suited as foundation models for coreference annotation.
- An examination of these models strengths and weaknesses for coreference annotation.

[1] https://huggingface.co/rmmhicke/t5-literary-coreference

Proceedings of LaTeCH-CLfL 2024, pages 270–277
March 22, 2024

2 Related Work

Many researchers have used neural networks (Lee et al., 2017; Clark and Manning, 2016; Dai et al., 2019) or encoder-only transformer models like the BERT models as the basis for coreference systems (Ye et al., 2020; Joshi et al., 2019; Otmazgin et al., 2022; Wu et al., 2020). These methods are multi-step and perform entity recognition and coreference annotation separately. Some studies have explored using generative LLMs for coreference, but they generally either fine-tune on auxiliary tasks (Mullick et al., 2023) or use zero- or few-shot prompting (Le et al., 2022; Le and Ritter, 2023).

Bohnet et al. (2023) fine-tune two sizes of mT5 (`xl` and `xxl`) to output coreference annotations for multi-lingual data. They annotate speaker interactions fed to the model one sentence at a time. The model outputs either link or append actions, which are used to annotate the coreference clusters in the next model input. Similarly, Zhang et al. (2023) find that seq2seq models such as T5 perform well when directly fine-tuned to output sentences annotated for coreference in a format similar to markdown. Like Zhang et al. (2023) and unlike Bohnet et al. (2023), we fine-tune a model to directly produce inline coreference annotations. Unlike both papers, we do not attempt to link annotations between sentences. We also focus specifically on literary coreference annotation, which Zhang et al. (2023) include but do not foreground, and compare encoder-decoder models to decoder-only models. Finally, we perform a qualitative examination of the fine-tuned models' strengths and weaknesses.

Coreference annotation has been applied to a wide variety of domains, such as movie screenplays (Baruah and Narayanan, 2023), biomedical journals (Cohen et al., 2017), and fiction (Bamman et al., 2020). Coreference annotation for literary texts in a variety of languages has also received a great deal of attention (Poot and van Cranenburgh, 2020; Schröder et al., 2021; Han et al., 2021; Krug et al., 2015; Roesiger et al., 2018). However, to our knowledge no work has yet focused on fine-tuning and evaluating generative LLMs specifically for literary coreference.

3 Data & Methods

Our training data is drawn from the LitBank corpus (Bamman et al., 2019), which includes 100 novels written in English before 1923 representing a mix of "high literary style... and popular pulp fiction" (Bamman et al., 2019, p. 2139). The mixture of publication dates and styles included in the corpus means that we are able to train and evaluate models for a variety of sentence styles. Human coreference annotations are available for the first ~2,000 tokens of each text for people, natural locations, built facilities, geo-political entities, organizations, and vehicles (Bamman et al., 2020).

Table 1: Example of an input-output pair used during fine-tuning. In the output, entities are surrounded by brackets and the association cluster is labeled as an integer.

Input	**Output**
Carl thrust his hands into his pockets, lowered his head, and darted up the street against the north wind.	[Carl: 1] thrust [his: 1] hands into [his: 1] pockets, lowered [his: 1] head, and darted up [the street: 2] against the north wind.

We created a subset of the LitBank corpus containing coreference-annotated sentences from the 92 novels with at least 50 annotated sentences. We standardized the formatting of each sentence by hand in an attempt to regularize punctuation. Then, we created an input and output version of each sample (see Table 1) where the input is the plain sentence and the output contains formatted coreference annotations. These were used to fine-tune and evaluate each model.

We withhold five novels entirely from the training dataset and include all sentences (at least 50) drawn from these novels in the test set. From each of the remaining 87 novels, we include 40 sentences in the training dataset, 2 sentences in the validation set, and the remaining sentences (at least 8) in the test dataset. The final dataset had 3,480 sentences for training, 174 sentences for validation, and 4,560 sentences for testing.

We then fine-tuned different sizes of three LLMs to perform literary coreference annotation: four sizes of T5 (`small`, `base`, `large`, `3b`) (Raffel et al., 2020), three sizes of mT5 (`small`, `base`, `large`) (Xue et al., 2021), and five sizes of Pythia (`70m`, `160m`, `410m`, `1b`, `1.4b`) (Biderman et al., 2023). mT5 is included to inform future research on multilingual coreference. Because we are interested in supporting users with low access to hardware accelerators, models are included only if they can be fine-tuned on a single GPU. The parameters used for fine-tuning can be found in Appendix A.

Finally, as a baseline we evaluate three spaCy-based coreference annotation systems: `fastcoref`

Table 2: Results for entity recognition and coreference. T5 has the best performance, particularly the 3B scale. FastCoref is a non-seq2seq baseline. The multilingual mT5 model is similar but not as good as T5, while the decoder-only Pythia family fails to add any annotations, correctly repeating only inputs with no annotations.

Model	Ent. F1	Coref. F1	Average Edit Distance	Exact String Match
Baselines				
`fastcoref`	50.86	40.46	—	—
`neuralcoref`	41.30	29.68	—	—
`coreferee`	35.04	2.81	—	—
T5				
`t5-3b`	**91.03**	**80.16**	**0.1**	**70.72**
`t5-large`	85.37	71.81	0.44	60.42
`t5-base`	83.74	61.35	1.74	47.76
`t5-small`	58.01	35.96	7.36	26.05
mT5				
`mT5-large`	81.90	58.78	1.77	42.39
`mT5-base`	70.14	47.70	5.12	15.81
`mT5-small`	0.41	0.24	149.79	0.0
Pythia				
`pythia-1.4b`	0.0	0.0	1077.79	9.06
`pythia-1b`	0.0	0.0	789.7	9.04
`pythia-410m`	0.0	0.0	1492.55	8.68
`pythia-160m`	0.0	0.0	1617.35	7.89
`pythia-70m`	0.0	0.0	1054.16	6.91

(Otmazgin et al., 2022) using the `LingMess` model (Otmazgin et al., 2023), huggingface's `neuralcoref` (which is based on Clark and Manning (2016)), and Explosion AI's `coreferee`. We do not include the performance of the BookNLP system, the most-used tool for literary coreference annotation, as it is also trained on LitBank and has likely seen some of our test sentences. However, we include a comparison of BookNLP and our fine-tuned `t5-3b` model's performance on two books not in LitBank: Virginia Woolf's *Orlando* and Radclyffe Hall's *The Well of Loneliness*, which entered the public domain in 2024.

4 Results

One advantage of seq2seq LLMs is their ability to produce complicated, structured output *as text* without the need for complex structured prediction model architectures. This means that we can use "off the shelf" transformers and fine-tune them using standard methods to produce coreference annotations. The problem with directly generating marked-up text, however, is that the generated output might not be purely additive: it may change the words in addition to adding annotations.

We therefore evaluate each fine-tuned LLM using four metrics. We measure the fidelity of the output with average Levenshtein distance between the input sentence and the model output stripped of all coreference annotation. We measure annotation accuracy using F1 scores for entity recognition and coreference annotation. Finally, we record whether there is an exact string match between the human-annotated output and the model output (not including leading or trailing spaces). This metric is measured as a percentage of sentences instead of a percentage of entities, as the F1 scores are.

We expect that the entity and coreference F1 scores will be an underestimate of the true model performance. This is for several reasons. The first is that we only count entities as labeled when the cleaned model output (stripped of entity and coreference annotation) is exactly the same length as the input string. Additionally, we only count exact entity or coreference matches. Thus, a match is not made when an entity is selected but misspelled (e.g. `[Helene's: 1]` is produced instead of `[Helen's: 1]`) or when a different substring is selected to represent an entity (e.g. `[the study behind the dining-room: 1]` is selected instead of `[the study: 1]`). The metric also only counts a coreference annotation as correct if the exact same index is used to identify the cluster. Therefore, if an extra entity cluster is labeled or missed (e.g. a sentence is annotated `[The lady: 1] in the room picked up [his: 2] hat.` instead of `[The lady: 1] in [the room: 2] picked up [his: 3] hat.`) some annotations may not be counted even though they are technically correctly identified. Finally, there are cases where the annotation of an entity is somewhat subjective and human observers may side with the initial annotations or the model output. For example, the LitBank annotation for the sentence `As it chanced, [Dale: 1] lay face down upon the floor of the loft` does not mark "the loft" as an entity. However, the model output does, and one could argue that this is a location which should be annotated. For these reasons, we expect that the true performance of these models in the eyes of a human evaluator would be higher than it appears given the strict F1 scores reported.

In order to provide more generous accuracy metrics that are comparable across other studies, we also use the `corefud-scorer` developed by Michal Novák and Martin Popel to report the models' performance using seven common coreference evaluation metrics (Table 3). We count singletons and

Table 3: Model results given in common coreference metrics (F1 scores). All Pythia models produce 0.0 F1 in all cases.

Model	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL avg.
Baselines							
`fastcoref`	80.08	56.72	58.31	38.38	56.90	54.00	58.39
`neuralcoref`	52.64	36.08	39.03	24.71	32.35	31.25	37.81
`coreferee`	41.49	29.77	33.76	22.09	25.18	23.46	31.12
T5							
`t5-3b`	**89.19**	**89.21**	**89.20**	**87.20**	**86.29**	**85.23**	**88.53**
`t5-large`	82.14	83.71	83.41	81.19	77.81	77.74	82.35
`t5-base`	71.71	77.73	75.47	72.82	70.76	65.22	74.09
`t5-small`	45.62	55.82	52.03	48.10	41.97	36.47	49.85
mT5							
`mT5-large`	68.26	74.75	72.72	69.98	67.07	61.37	71.00
`mT5-base`	61.38	64.26	62.14	56.83	55.91	49.41	60.82
`mT5-small`	0.08	0.40	0.56	0.60	0.02	0.22	0.36

Table 4: Example of less-successful input-output pairs produced from fine-tuned models. The first was produced by `t5-large`, the second by `pythia-1b`.

Input	Output
He shivered as if he had cold slimy water next his skin.	He shivered as if he had cold slimy water next **to** his skin.
We must go to Athens.	We must go to Athens. go to [Athens: 2]. go to [Athens: 2]: 2]. go to [A [Athens: 2]:]: 2]: 2 to [A [A [A: 2]: 2]: 2]: 2 to [: 2 [A: 2]: 2]: 2 2 2 2 2 to [: 2 [: 2]: 2]: 2 2 2 2 to [: 2 2]: 2 2...

require exact entity matches. Again, for these calculations we only count an entity as labeled if the clean model output is the same length as the input sentence. However, the scorer simplifies cases in which the same entity has been marked twice — transforming `[[he: 1]: 2]` to `[he: 2]` — and does not require exact spelling matches between labeled entities.

T5 Of the models tested, fine-tuned `t5-3b` achieves the highest performance (Table 2). It exactly reproduces 70.72% of the human-annotated outputs and has F1 scores of 91.03% for entity recognition and 80.16% for coreference annotation. Overall, the T5 models outperform all other model families; Pythia is unable to correctly identify any entities or coreference clusters, each mT5 model underperforms the equivalent T5 model, and all baselines are outperformed by all T5 models except `t5-small` (which is outperformed on coreference by `fastcoref`).

Larger models do better. The smaller T5 models, particularly `t5-small`, struggle to accurately match brackets and parentheses. They also frequently miss nested entities such as `[[her: 2] father: 1]`, randomly neglect to annotate any entities in a sentence comparable to those for which it has relatively high performance, or repeat substrings and brackets at the end of its output. `t5-large` sometimes adds extra words to sentences, often when grammatically intuitive (Row 1, Table 4), replicates only substrings of the original input, or makes other small formatting errors. It also continues to struggle with identifying some of the more complicated multi-word entities and nested entities. Finally, the replication errors for `t5-3b` are mostly formatting errors or the addition / exclusion of single words or small substrings. The output of this model sometimes still includes hallucinated repetitions, but it is very rare. Most of the annotation "errors" made by this model could be judgment calls, or cases where the original annotator had more context. Even this largest model occasionally struggles with matching brackets and nested entities, but this is also extremely rare.

If we examine single word replacements made by `t5-small` — for cases when the cleaned output is the exact length of the input — we find that it struggles with complicated and unusual words (e.g. *bordighera*, *schiaparelli*), names (e.g. *Katharine* is replaced by *Catarine*, *explained*, and *Katarine*), gender (*Mr.* is replaced by *Mrs.* five times), pronouns (*their* is replaced 82 times by 18 unique strings), and language (*however* is replaced by *nevertheless*, *allerdings*, *cepedant*, and *totuşi* and *Winterbourne* becomes *Hierbourne*). `t5-base` and `large` make similar single word replacements, but the translation errors are reduced to changing entities to their spelling in their original language (e.g. *Munich* becomes *München*). The replication errors made by `t5-3b` are almost all misspellings or

Table 5: Sentences with coreference annotation from fine-tuned `t5-3b` and BookNLP. The sentences are drawn from Virginia Woolf's *Orlando* (Rows 1 and 2) and Radclyffe Hall's *The Well of Loneliness* (Rows 3–5).

t5-3b	BookNLP
Rows of chairs with all their velvets faded stood ranged against the wall holding their arms out...	Rows of chairs with all [their: 5] velvets faded stood ranged against the wall holding [their: 5] arms out...
[Fathers: 1] instructed [[their: 1] sons: 2], [mothers: 3] [[their: 3] daughters: 4].	[Fathers: 1] instructed [[their: 1] sons: 2], [mothers: 3] [[their: 1] daughters: 4].
... sat [Stephen: 1] with [her: 1] feet stretched out to the fire and [her: 1] hands thrust in [her: 1] jacket pockets.	...sat [Stephen: 1] with [her: 2] feet stretched out to the fire and [her: 2] hands thrust in [her: 2] jacket pockets.
[Mrs. Williams: 1] glanced apologetically at [her: 2]: 'Excuse 'im, [Miss Stephen: 2], 'e's gettin' rather childish.	[Mrs. Williams: 2] glanced apologetically at [her: 2]:Éxcuse' [i: 1]m, [Miss Stephen: 3],' [e: 4]'s gettin' rather childish.
When one's getting on in years, one gets set in one's ways, and [my: 1] ways fit in very well with [Morton: 2].	When [one: 1]'s getting on in years, [one: 1] gets set in [one: 1]'s ways, and [my: 2] ways fit in very well with [Morton: 3].

changes in the plurality of words. Names also continue to confound the model as do parts of speech.

The CoNLL average coreference score achieved by the fine-tuned `t5-3b` model exceeds that of BookNLP by 9.5%; however, the BookNLP system simultaneously provides coreference annotations for each ∼2,000 word section of novel at once, whereas the T5 model runs on individual sentences. In order to further compare the two systems, we thus ran both on 100 random sentences drawn from two novels excluded from the systems' training data: Virginia Woolf's *Orlando* and Radclyffe Hall's *The Well of Loneliness*. The models produce the same or similar outputs for a large number of sentences and generally provide very plausible annotations. Of the 100 sentences, `t5-3b` only failed to replicate two inputs, both of which were quite long, and one of the replication errors only consisted of a dropped word.

There were some cases in which `t5-3b` appeared to perform better than BookNLP: it was better at identifying when pronouns referred to objects and not people (Row 1, Table 5), it sometimes identified the correct coreference cluster when BookNLP failed (Row 2, Table 5), and in one interesting case it was able to correctly cluster a name and pronouns despite a disconnect between the expected gender

Table 6: A `mt5-large` model fine-tuned only on English has some ability to identify entities in non-English text.

Input	Output
-La condesa de Albornoz -respondió el niño.	[La condesa de Albornoz: 1] -respondió [el niño: 2].
Mes parents ne peuvent plus faire autrement.	[Mes: 1] parents ne peuvent plus faire autrement.
Und vor ihm, in der Ferne da drüben, stiegen die blauen Bergriesen auf.	And vor [ihn: 1], in der Ferne da drüben, stiegen die blauen Bergriesen auf.

of the name and the gender of the pronouns (Row 3, Table 5). However, in some cases BookNLP caught edge cases that `t5-3b` did not: it more accurately identified entities written in dialect (Row 4, Table 5) and it occasionally caught less explicit entities (such as 'one' or 'others') that the model did not (Row 5, Table 5).

Overall, however, the performance of the two models appeared to be largely comparable. Despite this, we still consider the `t5-3b` model's performance to be significant for two reasons. Whereas the BookNLP pipeline required extensive development and would be very labor intensive to replicate for other data genres, fine-tuning the T5 models is simple and adaptable. In addition, the BookNLP pipeline is restricted to performing the tasks for which it has currently been trained; we view this as a promising calibration for the seq2seq models' ability to perform tasks that the LSTM cannot, such as relationship identification and characterization.

mT5 We also tested variations of T5 trained on larger multilingual collections. Although it performs worse than `t5-base` and larger, `mt5-large` reaches relatively high-performance. This performance may be boosted using additional training data, thus making it a viable option for further exploration into multi-lingual coreference annotation. Currently, when fed prompting sentences in German, Spanish, and French the model is able to reproduce sentences and identify some basic entities. However, it struggles with longer sentences and more complicated or opaque entities (Table 6).

Pythia Previous work has only considered encoder-decoder architectures. We evaluate the open-source decoder-only Pythia model family (Biderman et al., 2023). Pythia-based models are frequently able to replicate inputs. However, they usually append extensive hallucination to the replicated input, often adding repeating substrings, brackets, or integers (Row 2, Table 4). They very rarely include any formatting in the replicated text resembling that used for the coreference annotations. Thus, these models are currently unsuable for this task.

5 Conclusion

Fine-tuned generative LLMs show great promise for coreference annotation. They are simple to apply and can be efficiently trained for specific corpora from open-source base models. The errors made by large models in replicating inputs are minor and they are able to capture great complexity in the entities they annotate. In the future, we hope to extend this method to operate on longer contexts. Specifically, we propose to pre-pend all previously identified entities to each successive input. In addition, we believe that the high performance of the large, encoder-decoder models like `t5-3b` suggests that these models may be capable of performing more complex annotations, such as identifying emotional states or power dynamics between characters.

Acknowledgements

We would like to thank Federica Bologna, Katherine Lee, Kiara Liu, Rosamond Thalken, Andrea Wang, Matthew Wilkens, and Gregory Yauney for their thoughtful feedback. This work was supported by the NEH project AI for Humanists and grew out of Mimno's experience as a Visiting Researcher with Michael Collins' group at Google Research.

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, 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, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Sabyasachee Baruah and Shrikanth Narayanan. 2023. Character coreference resolution in movie screenplays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10300–10313, Toronto, Canada.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430, Honolulu, Hawaii, 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.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. 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.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(1):1–14.

Zeyu Dai, Hongliang Fei, and Ping Li. 2019. Coreference aware representation learning for neural named entity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4946–4953.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

Nghia T. Le, Fan Bai, and Alan Ritter. 2022. Few-shot anaphora resolution in scientific protocols via mixtures of in-context experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2693–2706, Abu Dhabi, United Arab Emirates.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers?

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.

Dhruv Mullick, Bilal Ghanem, and Alona Fyshe. 2023. Better handling coreference resolution in aspect level sentiment classification by fine-tuning language models. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 39–47, Singapore.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan.

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.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. 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 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Corbèn Poot and Andreas van Cranenburgh. 2020. A benchmark of rule-based and neural coreference resolution in Dutch novels and news. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 79–90, Barcelona, Spain (online). Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning*, 21(1):5485–5551.

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.

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.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, pages 1–25.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, 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.

A Fine-tuning Parameters

The fine-tuning parameters for each model can be found below. The batch size varied based on model.

Parameter	Value
evaluation_strategy	epoch
learning_rate	2e-5
weight_decay	0.01
save_total_limit	3
num_train_epochs	10
gradient_checkpointing	True