The Elephant in the Coreference Room: Resolving Coreference in Full-Length French Fiction Works

Anonymous ACL submission

Abstract

While coreference resolution is attracting more interest than ever from computational literature researchers, representative datasets of fully annotated long documents remain surprisingly scarce. In this paper, we introduce a new annotated corpus of three full-length French novels, totaling over 285,000 tokens. Unlike previous datasets focused on shorter texts, our corpus addresses the challenges posed by long, complex literary works, enabling evaluation of coreference models in the context of long-distance reference chains. We present a modular coreference resolution pipeline that allows for fine-grained error analysis. We show that our approach is competitive with state-of-the-art models and scales effectively to long documents. Finally, we demonstrate its usefulness to infer the gender of fictional characters, showcasing its relevance for both literary analysis and downstream natural language processing tasks.

1 Introduction

Coreference Resolution (CR)—the task of identifying and grouping textual mentions that refer to the same entity (e.g., a person, an organization, a place)—is a fundamental component of natural language processing (NLP). It underpins downstream applications such as information extraction (Yao et al., 2019), text summarization (Liu et al., 2021), and machine translation (Vu et al., 2024). Over the past decades, significant progress has been made in CR, evolving from rule-based multi-sieve systems to end-to-end neural models, encoder-decoder architectures, and large language models based approaches, all contributing to improvements on benchmark datasets (Porada et al., 2024).

These models have long been trained and evaluated solely on generic datasets such as OntoNotes (Hovy et al., 2006). As CR drew attention in other fields, it became evident that models trained on general datasets underperformed when applied to domain-specific tasks. To address this flaw, dedicated datasets have been developed, covering areas such as biomedical (Lu and Poesio, 2021) and encyclopedic data (Ghaddar and Langlais, 2016).

Driven by the availability of extensive digitized collections, literary texts have emerged as a key subject of computational studies and digital humanities (Moretti, 2013). A large part of such research focuses on characters, considered a fundamental aspect of fiction works. The study of characters is essential for analyzing narrative structures, plot development or conducting diachronic studies. More specifically, CR is crucial for applications such as quote attribution (Vishnubhotla et al., 2023), character archetypes inference (Bamman et al., 2014), and social networks extraction (Elson et al., 2010). Additionally, it has been employed to study the representation and behavior of characters according to their gender (van Zundert et al., 2023).

As outlined by Roesiger et al. (2018), literary texts present unique challenges for CR, including character evolution throughout the narrative and the prevalence of dialogues involving multiple participants. They also contain a high proportion of pronouns and nested mentions. Complex narrative structures—such as letters, flashbacks, and sudden narrator interventions—further complicate the task. Additionally, authors often rely on readers' contextual understanding rather than explicit statements, creating ambiguities when linking mentions.

To address these challenges, annotated datasets have been developed, covering multiple languages and genres, from classical novels and fantasy tales to contemporary literature. These resources enable training and evaluating in-domain coreference resolution models, leading to steady performance improvements (Martinelli et al., 2024). Despite visible progress on benchmarks, current state-of-the-art CR models still struggle with full-scale literary texts, limiting usefulness for downstream applications (Vishnubhotla et al., 2023).

A key factor contributing to this limitation lies in the scarcity of fully annotated long documents. Most existing datasets consist of short excerpts or relatively brief texts. Since coreference annotation is labor-intensive and costly, there exists a trade-off between annotating a larger number of short documents or a smaller number of long ones.

We argue that the lack of representative datasets for long literary texts is a major obstacle to effectively scaling CR models. This work aims to bridge this gap, and our contributions are as follows:

- an annotated dataset of character coreference for three full-length French novels spanning three centuries, showcasing the feasibility of combining automatic mention detection with manual coreference annotation.
- A modular CR pipeline scalable to long documents, enabling fine-grained error analysis and achieving competitive with cross-language benchmarks.
- A comprehensive study of the impact of document length on CR performance.
- A case study on character gender inference using CR models.[1]

2 Related Work

2.1 Coreference Models

Coreference resolution has undergone several paradigm shifts (Poesio et al., 2023), evolving from rule-based, linguistically informed models tested on limited examples to data-driven statistical approaches enabled by the creation of large annotated datasets such as those from the Message Understanding Conference (MUC) and the Automatic Content Extraction (ACE) shared tasks (Grishman and Sundheim, 1995; Doddington et al., 2004).

The adoption of neural network-based models, beginning with Wiseman et al. (2015), marked significant progress. The introduction of end-to-end models by Lee et al. (2017, 2018), further advanced CR by jointly detecting mention spans and resolving coreference, eliminating the need for external parsers and handcrafted mention detection models. Building on this foundation, higher-order inference (HOI) strategies and entity-level models were developed to refine entity representations during inference and leverage cluster-level information.

However, as highlighted by Xu and Choi (2020), the performance gains from these strategies have been marginal compared to the substantial improvements achieved by the use of more powerful encoders like ELMo, BERT and DeBERTaV3.

Alternative approaches using encoder-decoder architectures and large language models have been proposed, framing CR as sequence-to-sequence (Hicke and Mimno, 2024) or question-answering (Wu et al., 2020; Gan et al., 2024) tasks. While these methods show promising results, they are computationally intensive and do not scale efficiently to longer documents or resource-constrained scenarios.

Ultimately, the development and evaluation of CR models remain deeply tied to the availability of annotated datasets, which continue to drive the direction of research in this field.

2.2 Existing Datasets

While MUC and ACE laid the foundation for coreference datasets, OntoNotes has since become the primary benchmark for CR. Published in 2006 (Hovy et al.) and regularly updated, OntoNotes has been used in the CoNLL shared tasks (Pradhan et al., 2011, 2012). Its latest version (Weischedel et al., 2013) spans multiple languages (English, Chinese and Arabic), and genres, including conversations, news, web, and religious texts. The English part contains 1.6M tokens across 3,943 documents, averaging 467 tokens per document. OntoNotes does not contains singleton mentions—those that do not corefer with any other mention.

The growing interest for large literature corpora has driven the development of dedicated annotated datasets. The late 2010s saw the emergence of the first literary CR datasets, beginning with DROC (Krug et al., 2018), including samples from 90 German novels annotated with character coreference chains. With over 393,000 tokens (averaging 4,368 tokens per document), DROC remains the largest literary CR dataset to date. The RiddleCoref dataset (van Cranenburgh, 2019) followed, covering excerpts from 21 contemporary Dutch novels, though it is not publicly available due to copyright restrictions. Bamman et al. (2020) released LitBank, consisting of the first 2,000 tokens from 100 English novels. This dataset covers six entity categories (persons, faculties, locations, geopolitical, organizations and vehicles). Other datasets include FantasyCoref (Han et al., 2021), KoConovel covering 50 full-length Korean short stories (Kim et al., 2024), and LitBank-fr (Mélanie et al., 2024). This last dataset is noteworthy in that it covers longer

[1] All code and data will be made publicly available.

	Lang.	Domain	Doc.	Tokens	Tokens / Doc.	
					Avg.	Max.
Annotated Datasets						
OntoNotes^{en} (Weischedel et al., 2013)	English	Non-literary	3,493	1,600,000	467	4,009
DROC (Krug et al., 2018)	German	Fiction	90	393,164	4,368	15,718
RiddleCoref (van Cranenburgh, 2019)	Dutch	Fiction	21	107,143	5,102	-
LitBank (Bamman et al., 2020)	English	Fiction	100	210,532	2,105	3,419
FantasyCoref (Han et al., 2021)	English	Fantasy	214	367,891	1,719	13,471
KoCoNovel (Kim et al., 2024)	Korean	Fiction	50	178,000	3,578	19,875
LitBank-fr (Mélanie et al., 2024)	French	Fiction	28	275,360	9,834	30,987
Target Datasets						
Standard Ebooks[2]	English	Fiction	770	82,855,210	107,604	1,105,964
Chapitres (Leblond, 2022)	French	Fiction	2,960	240,971,614	81,409	878,645
Contribution						
Ours	French	Fiction	3	285,176	95,058	115,415

Table 1: Comparison of coreference annotation datasets: OntoNotes (English section), fiction datasets, and target datasets across languages.

excerpts of text—averaging 9,834 tokens and up to 30,987 for the longest document.

Despite these resources, extrinsic evaluations reveal that CR models perform poorly on full-length documents (van Zundert et al., 2023). Studies consistently show that performance degrades with increasing document length (Joshi et al., 2019; Toshniwal et al., 2020; Shridhar et al., 2023). This represents a major challenge given that practical applications involve digitized collections such as Project Gutenberg or Wikisource, where documents frequently exceed 90,000 tokens and can reach up to a million as illustrated in Table 1.

While some initiatives annotate entire books, they often diverge from standard guidelines. He et al. (2013) annotated *Pride and Prejudice* but focused solely on proper mentions. Similarly, van Zundert et al. (2023) labeled character aliases across 170 novels, omitting pronouns and noun phrases. Other datasets, such as QuoteLi3 (Muzny et al., 2017) and PNDC (Vishnubhotla et al., 2022), include coreference annotations for speakers and direct speech but lack broader character coverage. To the best of our knowledge, the only CR results reported on a document of substantial length (37k tokens) come from Guo et al. (2023), but they omit singletons, plural mentions, and nested entities.

These observations underscore the need for an annotated corpus of full-length literary documents. Such a resource will enable more robust evaluation and improvement of CR models, addressing the gap between current datasets and intended applications.

[2] standardebooks.org

3 New Dataset

We selected three average-length French novels spanning three centuries, resulting in a total of 285,176 tokens. We chose to annotate coreference for character mentions only for several reasons. First, most downstream tasks in literary NLP focus on characters. Second, previous work shows that characters account for the majority of annotated mentions—83.1% and 83.5% in LitBank and LitBank-fr, respectively. Restricting annotations to character mentions allows us to leverage the 31,570 mentions already annotated in LitBank-fr to train a highly accurate mention detection model.

For consistency and interoperability, we strictly adhere to the annotation guidelines established by Mélanie et al. (2024) for French. We annotate all mentions referring to a character, including pronouns, nominal phrases, proper nouns, singletons and nested entities. Coreference links capture strict identity relations between mentions.

3.1 Mentions Detection Model

While Mélanie et al. (2024) report strong results for mention detection, we opted to retrain our own model. Our approach builds on a stacked BiLSTM-CRF architecture inspired by Ju et al. (2018), leveraging contextual token embeddings from CamemBERT$_{\text{LARGE}}$ (Martin et al., 2020). We achieved an improvement of 4.99 in F1-score on the test set from LitBank-fr (Table 2). To assess generalization performance, we also conducted a leave-one-out cross-validation (LOOCV). Details of the model architecture and hyperparameters are available in the Appendix A.

Model	P	R	F1	Support
Mélanie et al. (test set)	85.0	92.1	88.4	4,061
Ours (test set)	**91.29**	**95.59**	**93.39**	4,061
Ours (LOOCV)	90.72	93.52	92.05	31,570

Table 2: Comparison of mention detection performance.

Average Mentions / Doc.	13,178
Singletons Ratio	1.15%
Coreference Chains / Doc.	159
Average Mentions / Chain	82
Maximum Mentions / Chain	4,932
Average Entity Spread (tokens)	17,529
Maximum Entity Spread (tokens)	115,369

Table 3: Dataset statistics summary.

Coreference annotation is usually carried out in two stages: annotating the mention spans, then linking mentions referring to the same entity together. Given our model's 92.05 F1-score, we consider its performance sufficient to automate the first operation, significantly reducing annotation time.

3.2 Coreference Annotation

Coreference annotation is performed manually, building upon the automatically detected mentions. A single trained annotator reviews the text, assigns entity identifiers to each mention, corrects errors from the mention detection step, deleting spurious mentions, adding missed ones, and adjusting incorrect boundaries. This process ensures that both the mentions and the coreference are considered gold annotations at the end.

Several coreference annotation tools have been developed in recent years (Stenetorp et al., 2012; Yimam et al., 2013; Vala et al., 2016; Muzny et al., 2017). We use SACR, an open-source, browser-based interface developed by Oberle (2018). This tool meets our requirements, allowing efficient processing of long texts, tracking a large number of entities and handling nested mentions.

In practice, mention detection errors are rare and mainly involve difficult cases, such as ambiguous mentions (animals with agentivity, appositions, reflexive pronouns), nested mentions and other edge cases. This confirms the feasibility of leveraging automatic mention detection to accelerate coreference annotation. We estimate the manual annotation of a 100,000-token text to take around 30 to 40 hours.

3.3 Dataset Statistics

Table 3 summarizes key statistics from our dataset. The entity spread refers to the distance between the first and the last mention of an entity (Toshniwal et al., 2020). This metric highlights a key specificity of literary texts, characters can be referred to thousands times over several hundred pages, comprising thousands of tokens.

Another important metric for characterizing coreference is the distance to the nearest antecedent (Han et al., 2021). For each mention, we locate the previous mention belonging to the same coreference chain and measure the difference in terms of mention positions. Bamman et al. (2020) analyzed the distribution of distance to nearest antecedent for proper nouns, noun phrases and pronouns. We replicate their experiment and report similar results. While 95% of pronouns appear within 7 mentions of their last antecedent, this distance can reach up to 270 mentions for proper nouns and noun phrases. This observation calls for distinct handling of pronouns, common, and proper nouns during coreference resolution. We also notice that the last 1% of proper and common noun mentions exhibit a distance of over 1,700 mentions, presenting a significant challenge for coreference resolution. The full distribution of antecedent distances can be found in the Appendix B.

3.4 Corpus Merging

Since we followed the guidelines from Mélanie et al. (2024), the newly annotated dataset is fully compatible with the character annotations from the LitBank-fr dataset. It allows us to merge the two datasets, resulting in a combined dataset containing 31 documents and 71,105 character mentions.

This merged dataset becomes the largest annotated literary coreference dataset in terms of tokens (560,536), average document length (18,081 tokens), and maximum document length (115,415 tokens). Unless otherwise specified, all results presented in this paper pertain to this merged corpus, which we refer to as *Long-LitBank-fr*.

4 Coreference Resolution

Several coreference resolution pipelines are available off-the-shelf, such as the *CoreferenceResolver* module from Spacy[3], Fastcoref (Otmazgin et al., 2022) and AllenNLP (Gardner et al.,

[3] https://spacy.io/api/coref

2018). BookNLP (Bamman et al., 2020), is a pipeline performing, among other, mentions detection and coreference resolution for English. A French adaptation, BookNLP-fr, was developed by Mélanie et al. (2024) and trained on the LitBank-fr dataset. The BookNLP pipelines implement an end-to-end coreference resolution model (Ju et al., 2018), which makes them impractical to modify and conduct detailed error analysis.

Diverging from recent trends of end-to-end architectures, we propose to implement coreference resolution as a modular pipeline, facilitating the study of each component's role and enabling fine-grained error analysis.

4.1 Pipeline Description

The mention-pair-based coreference resolution pipeline is composed of the following modules :

Mention Detection: We use the mention detection module discussed previously. We retrained it on the merged corpus, achieving an increase of 2.31 points in F1-score (94.33). As mention detection significantly impacts overall CR performance, we make it possible to bypass the errors introduced by this module by using gold mentions as input to the mention-pair encoder.

Considered Antecedents: To address the quadratic complexity of considering all antecedents, recent approaches introduce hyperparameters to uniformly limit the number of considered antecedents (Thirukovalluru et al., 2021; Wu et al., 2020). Inspired by Bamman et al. (2020) and supported by our observations regarding antecedent distance, we adopt a mention-type-specific approach. We limit the number of antecedents to 30 for pronouns and 300 for proper and common nouns.

Mention Pair Encoder: Mention-pairs are encoded by concatenating the representations of the two mentions with a feature vector that includes attributes such as gender, grammatical person, and the distance between the mentions. For multi-token mentions, the representation is calculated as the average of the first and last tokens embeddings.

Mention Pair Scorer: Encoded mention-pairs are passed into a feedforward neural network trained to predict whether two mentions refer to the same entity. Additional details about the features, model architecture and training parameters are provided in the Appendix C.

Antecedent Ranker: Following Wiseman et al. (2015), candidate antecedents are ranked according to their predicted scores. During inference, the highest-scoring antecedent is selected unless all scores fall below a 0.5 threshold, in which case the null antecedent is assigned.

Entity Clustering: The default strategy for linking mentions into entity clusters is to scan the document from left to right, each new mention is either merged into the cluster of its best-ranked antecedent or left as a standalone entity. Coreference chains are defined as the set of mentions in a cluster.

We explore additional strategies to address specific challenges and improve overall performance.

Handling Limited Antecedents: Limiting the number of considered antecedents can lead to split coreference chains. A common strategy in literary texts is to link all matching proper nouns at the document level, along with their derivatives. While previous works have been using hand-crafted sets of aliases to link proper mentions (Bamman et al., 2020), we leverage local mention-pairs scoring to perform coreference resolution at the document scale. Let's say that all local predictions involving mentions of "Sir Ralph Brown" and "Raphael" are coreferent, we propagate this decision to all mention-pairs at the global scale, bridging the gap between a mention and an antecedent that would otherwise be out of the range of locally considered antecedents.

Leveraging Non-Coreference Predictions: While most mention-pair models focus on positive coreference links, the cross-entropy loss used during training involves that they are equally trained to predict non-coreference. We propose leveraging high-confidence non-coreference predictions to prevent later incorrect cluster merging. Mention-pairs containing a coordinating conjunction, such as "[Ralph] and [Mr. Delmare]", are a strong indication of non-coreference between these two mentions, which can be used to prevent the merging of these two entities at document level. This approach is combined with an "easy-first" clustering strategy (Clark and Manning, 2016), which processes mentions in order of confidence rather than left-to-right, thus delaying harder decisions.

The addition of these two strategies is refered to as the *easy-first, global proper mentions coreference approach*. Its effectiveness is evaluated in subsequent experiments.

4.2 Evaluation Metrics

We evaluate CR performance using MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and $CEAF_e$ (Luo, 2005) scores. For overall performance assessment we report the average F1-score of the three metrics which we refer to as the CoNLL F1-score (Pradhan et al., 2012). We use the scorer implementation by Grobol.[4]

4.3 Document Length

While Poot and van Cranenburgh (2020) investigated the impact of document length on CR by truncating documents to different sizes, we adopt a splitting approach. This allows us to evaluate CR performance on more text excerpts.

Given a target sample size of L tokens, we first select all documents from our corpus that exceed this length. Each selected document is then split into non-overlapping samples, each containing L tokens. CR is performed independently on each sample, and the results are averaged across all samples of a given document. To compute the overall CR score, we calculate the macro-average across all retained documents.

4.4 Coreference Resolution Results

4.4.1 Mention-Pairs Scorer Results

The mention-pairs scorer, evaluated using leave-one-out cross-validation with gold mention spans, achieved an overall accuracy of 88.10%. As shown in Table 4, performance disparities between classes reflect the underlying class imbalance, with significantly higher precision and recall for non-coreferent pairs (class 0). Notably, most errors occurred for mention pairs where the scorer's confidence is low (~ 0.5) (see Appendix D). As we use the highest ranked antecedent strategy, not all scorer decisions are used during entity clustering, mitigating the number of wrong decisions considered.

Coref.	P	R	F1	Support
0	92.31	93.18	92.74	5.52M (82%)
1	68.49	65.62	67.02	1.25M (18%)

Table 4: Mention-pairs scorer performance on Long-LitBank-fr corpus. Precision (P), Recall (R).

4.4.2 Highest Ranked Antecedent

After sorting antecedents, the correct antecedent was predicted in 88.05% of cases, highlighting the effectiveness of this approach. Errors occurred for 8,496 mentions (11.95%). In 1,478 cases (2.08%), the range of considered antecedents is too narrow, leaving true antecedents out of reach. For these mentions, the null antecedent is assigned approximately half the time, while an unrelated antecedent is assigned in the other half.

In 7,018 cases (9.87%), the true antecedent is within reach, but the model incorrectly assigned a different antecedent in nearly 90% of instances. In the remaining 10%, the null antecedent is wrongly predicted.

The additional global proper mentions coreference strategy aims at reducing both types of errors, by bridging the gap between proper mentions and their long distance antecedent, and by limiting clustering of mentions that are believed to be distinct from local mention-pair scores.

4.4.3 Entity Clustering Strategies

The global proper mentions strategy leads to an overall gain in performance measured by CoNLL F1-score of 2 points. We observe a slight drop for MUC, but a significant improvement on both B^3 and $CEAF_e$, suggesting an enhancement of the CR both from a mention-level and entity-level.

Strategy	MUC	B^3	$CEAF_e$	CoNLL
Left to Right (Baseline)	**94.66**	64.39	61.28	73.44
Easy-first Global Proper CR	94.47 (-0.19)	**69.26** (+4.87)	**62.58** (+1.30)	**75.44** (+2.00)

Table 5: Coreference resolution for Long-LitBank-fr corpus. Average F1-scores. Gold mentions.

These scores reflect the average performances of this strategy on the full Long-LitBank-fr corpus (averaging 18,081 tokens per document). However it is best suited to long texts that present both the risk of out-of-reach antecedent, and sufficient local evidence on proper mentions-pairs to propagate document-wide decisions. In the following section we examine the impact of document length on coreference resolution performances.

4.4.4 Influence of Document Length

From Figure 1, we observe that the overall CR performance decreases with document length. Much of the performance loss is observed in the lower range. This is critical for literary CR, and might

[4] https://github.com/LoicGrobol/scorch

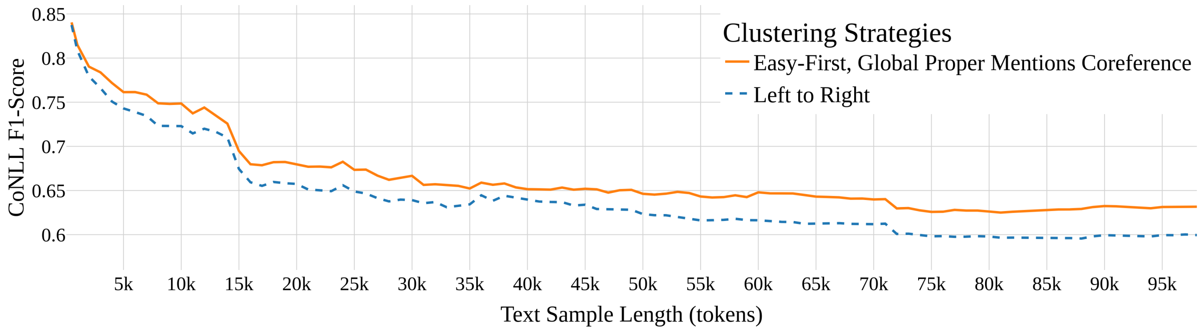

Figure 1: Impact of document length on coreference resolution performance. Gold mentions.

well explain why coreference resolution models trained and evaluated on documents of limited length (2k to 10k tokens), have been deceiving when used for downstream tasks on full length documents (∼ 90k tokens).

The proper mentions global coreference strategy consistently outperform the vanilla left-to-right method. Performance gains is mostly negligible for short documents (< 2k tokens), but becomes significant and stable beyond, reaching +3 points on the CoNLL F1-score. This shows the effectiveness of our approach for handling CR in longer documents.

4.4.5 Comparison to Other Benchmarks

While comparing results across different corpora and languages can be challenging, we chose to do so in order to benchmark the performance of our pipeline with existing systems. Given that document length is critical in CR, we ensure that all model comparisons are conducted on corpora with similar average token counts.

For French coreference resolution, our new pipeline significantly outperforms the model proposed by Mélanie et al. (2024) on their test set. In cross-corpus and cross-language benchmarks, our model consistently surpasses existing baselines, with performance gains strongly correlated with document length—reaching an improvement of +23 points on texts averaging 37,000 tokens. The only case where our model falls short of state-of-the-art results is in comparison to Thirukovalluru et al. (2021). This is due to their use of SpanBERT, a high-performing encoder, well-suited for CR. Given the scarcity of French pretrained models and the absence of a SpanBERT equivalent, future work should explore using larger multilingual models to bridge this gap.

Additionally, Table 6 illustrates the impact of using predicted mentions as input to the scorer, leading to an overall performance drop of 7%, this result is consistent with previous publications.

While this experiment reveals performance limitations exacerbated by document length, the commonly used CR metrics (MUC, B^3, $CEAF_e$) have been criticised for presenting systematic flaws. Alternative metrics such as LEA (Moosavi and Strube, 2016) and BLANC (Recasens and Hovy, 2011) have been proposed as better aligned with linguistic intuitions. Others have argued for extrinsic evaluation methods (O'Keefe et al., 2013; Vishnubhotla et al., 2023), where CR is assessed based on its contribution to downstream tasks, such as classification, which are often easier to evaluate.

We examine the usefulness of our CR pipeline for predicting the gender of fictional characters.

Corpus	**Model**	**Mentions**	**Tokens / Doc**	**MUC**	**B^3**	**CEAFe**	**CoNLL**
LitBank-fr (test-set)	Mélanie et al. 2024	Gold	2,000	88.0	69.2	71.8	76.4
LitBank-fr (test-set)	Ours	Gold	2,000	**92.43**	**70.67**	**75.59**	**79.56**
LitBank (*English*)	Bamman et al. 2020	Gold	2,105	88.5	72.6	76.7	79.3
LitBank-fr (LOOCV)	Ours	Gold	2,105	**91.93**	**74.6**	**75.35**	**80.63**
LitBank (*English*)	Bamman et al. 2020	Predicted	2,105	84.3	62.73	57.3	68.1
LitBank (*English*)	Thirukovalluru et al. 2021	Predicted	2,105	**89.50**	**78.21**	**67.59**	**78.44**
LitBank-fr (LOOCV)	Ours	Predicted	2,105	84.58	74.77	63.30	73.21
KoCoNovel (*Korean*)	Kim et al. 2024	Predicted	3,578	71.06	57.33	44.19	57.53
Long-LitBank-fr (LOOCV)	Ours	Predicted	3,578	**88.31**	**68.79**	**47.17**	**68.09**
G. Orwell, *Animal Farm*	Guo et al. 2023	Predicted	37,000	-	-	-	36.3
Long-LitBank-fr (LOOCV)	Ours	Predicted	37,000	**92.79**	**52.35**	**32.89**	**59.34**

Table 6: Comparison of CR performance with other work on literary coreference with predicted and gold mentions.

5 Gender Prediction Case study

As mentioned, studies gravitating around character gender have attracted substantial attention from computational humanities researchers (Underwood et al., 2018). A critical part of such studies lies in the ability to accurately predict the gender of as many character mentions as possible in order to get representative results.

Early works relied on heuristics to infer gender from explicit clues (he, Mrs, the old man), achieving high precision (90%) but lower recall (30-50%). This is due to the high proportion of ambiguous mentions in literary texts involving first and second person pronouns, indefinite pronouns, as well as ambiguous nouns. Recent works leverages CR for broader gender prediction (Vianne et al., 2023).

5.1 Data Preparation

We use the *Long-Litbank-fr* corpus. Starting with the 71,106 character mentions, we discard singletons (2.74%) and plural mentions (9.84%). We manually annotate the gender of the remaining 62,162 mentions at the entity level. We adopt a binary approach to gender (male, female). Works of fiction are subject to play on characters' gender, such as gender revelation or asymmetry of knowledge between characters. To assign character gender we adopt the omniscient perspective (Kim et al., 2024), refering to the knowledge one have at the end of the entire book. After annotation, we discard chains whose gender cannot be annotated with certainty, leaving us with 804 entities and 61,852 mentions (86.99% of all mentions).

5.2 Prediction Pipeline

To predict the gender of character mentions we implement a multi-stage solution:

Heuristic rules: assign gender based on heuristics from explicit gender clues (pronouns, noun phrases, articles, adjectives).

First-name database: determine the gender of proper mentions using a statistical database of first names given to children born in France between 1900 and 2023.[5]

Coreference propagation: resolve coreference, compute the male/female ratio of processed mentions, and assign the majority gender to all mentions within the coreference chain.

[5]Database from the French National Institute of Statistics and Economic Studies (*INSEE*).

We compare our results with those of Naguib et al. (2022) who used a similar combination of heuristic rules and CR to infer character gender.

5.3 Case Study Results

Coreference resolution significantly improves recall compared to rule-based methods. While heuristics achieve high precision (>98%), they suffer from low recall (37-47%), reflecting the significant number of mentions whose gender cannot be inferred without additional context.

Our approach outperforms the baseline by leveraging sophisticated heuristic rules, a first-names database, and a more effective CR pipeline. Although CR slightly reduces precision—a consequence of clustering errors—the substantial recall gain makes it a robust method overall.

	Male		**Female**	
	P	**R**	**P**	**R**
Baseline Naguib et al. 2022	95.00	45.00	97.00	58.00
Heuristic Rules	**99.77**	36.97	**98.85**	46.67
+ First-name data	99.77	38.35	98.82	47.41
+ Coreference	95.35	**91.55**	90.37	**93.40**

Table 7: Mentions gender prediction performance. Precision (P), Recall (R).

6 Conclusion

We highlight critical limitations in coreference resolution (CR) for literary texts, particularly the scarcity of representative datasets, limiting the possibility to train and evaluate models tailored for literary computational studies. To bridge this gap, we release an annotated corpus of character coreference chains for three full-length French novels spanning three centuries (285,000+ tokens). We introduce a modular CR pipeline tailored for long documents, integrating global coreference propagation for proper nouns and an easy-first clustering approach. After carrying out a detailed error analysis of each component, we study the impact of document length on overall coreference performance. Our approach is competitive with existing state-of-the-art models, demonstrating good performance on longer texts. To demonstrate practical value, we apply it to character gender inference, significantly improving recall over rule-based baselines while maintaining high precision, and outperforming other CR-based approach. This study underscores the need for robust datasets and well-evaluated models to advance literary CR research.

Limitations

While our dataset is among the largest annotated literary datasets in terms of tokens (285,000), it is limited by the fact that it only contains three documents. This implies that it does not encompass the full diversity of time periods, literary movements, and genres within French literature. This limitation may impact the generalizability of the coreference resolution (CR) models trained on this dataset. The proposed *Long-LitBank-fr* corpus resulting from the concatenation with the *LitBank-fr* dataset mitigates this issue by increasing diversity and improving the potential for model generalization.

Another limitation is that we focused solely on annotating coreference chains for characters. Some downstream applications may require resolving coreference for other entity types (e.g., geographical entities, events). Since our annotations are restricted to characters, a model trained exclusively on this data may not easily transfer to tasks involving other entity types. In such cases, enriching the annotations would be necessary for broader applicability.

Furthermore, our study is limited to French-language texts, and we did not explore cross-lingual generalization of our pipeline. Expanding the dataset to include full documents in other languages could improve its applicability. This could be achieved through annotation transfer or by leveraging multilingual models, which would help reduce the cost of manual annotation.

Finally, while extrinsic evaluation is not the primary focus of this work, we have only begun to assess our pipeline through its application to character gender inference. A more comprehensive evaluation of the models' suitability for full-document literary analysis would require additional extrinsic assessments, such as network extraction or quote attribution.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *International Conference on Computational Linguistics*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth 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.

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

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 138–147, Uppsala, Sweden. Association for Computational Linguistics.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2016. Wiki-Coref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Dual cache for long document neural coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15272–15285, Toronto, Canada. Association for Computational Linguistics.

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.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Rebecca Hicke and David Mimno. 2024. [Lions: 1] and [Tigers: 2] and [Bears: 3], oh my! literary coreference annotation with LLMs. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 270–277, St. Julians, Malta. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

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.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. 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 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Koconovel: Annotated dataset of character coreference in korean novels.

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]. In *DARIAH-DE Working Papers 27, DARIAH-DE*.

Aude Leblond. 2022. Corpus chapitres.

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.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 12–23, Punta Cana, Dominican Republic. 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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. 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.

Franco Moretti. 2013. *Distant Reading*. Verso, London.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470, Valencia, Spain. Association for Computational Linguistics.

Frédérique Mélanie, Jean Barré, Olga Seminck, Clément Plancq, Marco Naguib, Martial Pastor, and Thierry Poibeau. 2024. Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. *Journal of Computational Literary Studies*, 3(1):1–34.

Marco Naguib, Marine Delaborde, Blandine Andrault, Anaïs Bekolo, and Olga Seminck. 2022. Romanciers et romancières du XIXème siècle : une étude automatique du genre sur le corpus GIRLS (male and female novelists : an automatic study of gender of authors and their characters). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Atelier TAL et Humanités Numériques (TAL-HN)*, pages 66–77, Avignon, France. ATALA.

Bruno Oberle. 2018. Sacr: A drag-and-drop based tool for coreference annotation. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

Tim O'Keefe, Kellie Webster, James R. Curran, and Irena Koprinska. 2013. Examining the impact of coreference resolution on quote attribution. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 43–52, Brisbane, Australia.

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. Association for Computational Linguistics.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9(Volume 9, 2023):561–587.

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.

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.

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.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA. Association for Computational Linguistics.

M. Recasens and E. Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Nat. Lang. Eng.*, 17(4):485–510.

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.

Kumar Shridhar, Nicholas Monath, Raghuveer Thirukovalluru, Alessandro Stolfo, Manzil Zaheer, Andrew McCallum, and Mrinmaya Sachan. 2023. Longtonotes: OntoNotes with longer coreference chains. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1428–1442, Dubrovnik, Croatia. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3921–3931, Online. Association for Computational Linguistics.

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.

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

Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 184–189, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Joris van Zundert, Andreas van Cranenburgh, and Roel Smeets. 2023. Putting dutchcoref to the test: Character detection and gender dynamics in contemporary dutch novels. In *Proceedings of the Computational Humanities Research conference 2023*, pages 757–771. CEUR Workshop Proceedings (CEUR-WS.org). Computational Humanities Research Conference ; Conference date: 06-12-2023 Through 08-12-2023.

Laurine Vianne, Yoann Dupont, and Jean Barré. 2023. Gender Bias in French Literature. In *Conference on Computational Humanities Research CHR2023*, Paris, France. Ariane and Epita and Humanistica.

Marc B. Vilain, John D. Burger, John S. Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Conference on Message Understanding, MUC 1995, Columbia, Maryland, USA, November 6-8, 1995*, pages 45–52. ACL.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848, Marseille, France. European Language Resources Association.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Context-aware machine translation with source coreference explanation. *Transactions of the Association for Computational Linguistics*, 12:856–874.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0.

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.

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.

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.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria. Association for Computational Linguistics.

A Mention Detection Model

The mention detection module consists of two stacked BiLSTM-CRF models, each trained on a different nesting level of mentions. During inference, predicted spans from both models are combined. If two mention spans overlap, the span with the lower prediction confidence is discarded.

BERT embeddings: The raw text is split into overlapping segments of length L (the maximum embedding model context window) with an overlap of $L/2$ to maximize the context available for each token. Each segment is passed through the $\text{CamemBERT}_{\text{LARGE}}$ model, and we retrieve the last hidden layer as the token representations (1024 dimensions). The final token embedding is computed as the average from overlapping segments. We do not fine-tune CamemBERT for this task.

BIOES tag prediction: For each sentence, token representations are passed through the BiLSTM-CRF model, which outputs a sequence of BIOES tags: B-PER (Beginning of mention), I-PER (Inside), E-PER (End), S-PER (Single-token mention), and O (Outside).

A.1 Model Architecture

- **Locked Dropout** (0.5) applied to embeddings for regularization.
- **Projection Layer**: Highway network mapping 1024 $\rightarrow$ 2048 dimensions.
- **BiLSTM Layer**: Single bidirectional LSTM (256 hidden units per direction).
- **Linear Layer**: Maps 512-dimensional BiLSTM outputs to BIOES label scores.
- **CRF Layer**: Enforces structured consistency in predictions.

A.2 Model Training

- **Data Splitting**: Leave-One-Out Cross-Validation (LOOCV) with an 85%/15% train-validation split.
- **Batch Size**: 16 sentences per batch.
- **Optimization**: Adam optimizer ($lr = 1.4 \times 10^{-4}$, weight decay = 10^{-5}).
- **Learning Rate Scheduling**: ReduceLROnPlateau (factor = 0.5, patience = 2).
- **Average Training Epochs**: 20.
- **Hardware**: Trained on a single 6GB Nvidia RTX 1000 Ada Generation GPU.

B Nearest Antecedent Distribution

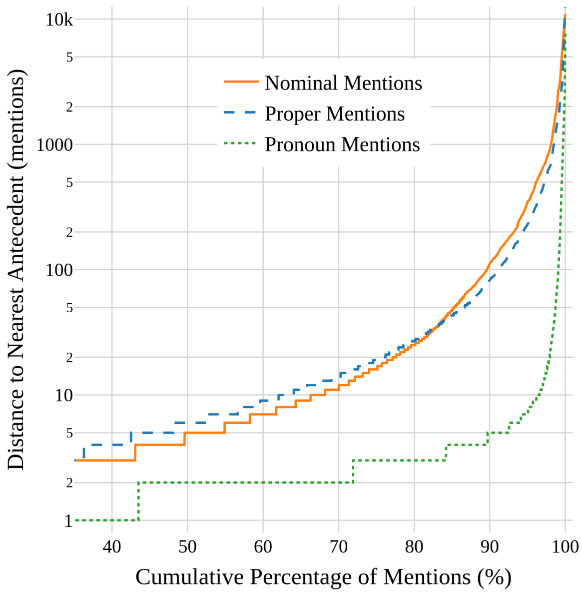

Figure 2: Distance to nearest antecedent for mentions of different type.

C Coreference Resolution Model

C.1 Model Architecture

- **Model Input**: 2,165-dimensional vector, composed of concatenated:
  - **CamemBERT embeddings**: Maximum context embeddings for both mentions (2 × 1,024 = 2,048 dimensions).
  - **Mention Features** (106 dimensions):
    - Mention length.
    - Position of the mention's start token in the sentence.
    - Grammatical category (pronoun, common noun, proper noun).
    - Dependency relation of the mention's head (one-hot encoded).
    - Gender (one-hot encoded).
    - Number (one-hot encoded).
    - Grammatical person (one-hot encoded).
  - **Mention Pair Features** (11 dimensions):
    - Distance between mention IDs.
    - Distance between start and end tokens of mentions.
    - Sentence and paragraph distance.
    - Difference in nesting levels.
    - Ratio of shared tokens between mentions.
    - Exact text match (binary).
    - Exact match of mention heads (binary).

    * Match of syntactic heads (binary).
    * Match of entity types (binary).

- **Hidden Layers**:
  - Three fully connected layers.
  - 1,900 hidden units per layer with ReLU activation.
  - Dropout rate of 0.6 for regularization.
- **Final Layer**:
  - Linear layer mapping from 1,900 dimensions to a single scalar score.
  - Output: Continuous value between 0 (not coreferent) and 1 (coreferent).

C.2 Model Training

- **Data Splitting**: Leave-One-Out Cross-Validation (LOOCV) with an 85%/15% train-validation split.
- **Batch Size**: 16,000 mention-pairs per batch.
- **Optimization**: Adam optimizer ($lr = 4.0 \times 10^{-4}$, weight decay = 10^{-5}).
- **Antecedent Candidates**:
  - 30 for pronouns.
  - 300 for common and proper nouns.
- **Hardware**: Trained on a single 6GB Nvidia RTX 1000 Ada Generation GPU.

D Mention-Pairs Scorer Error Distribution

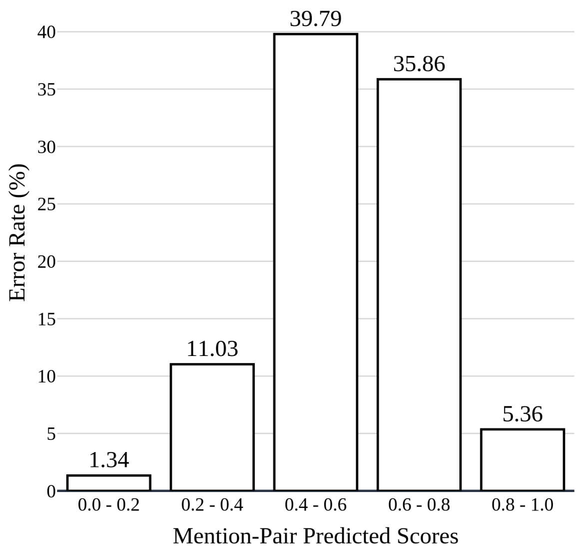

Figure 3: Error Rate by Mention-pair Predicted Score Range.

E Annotated Dataset Details

Year	Author	Text	Tokens
1731	Antoine-François Prévost	*Manon Lescaut*	71,219
1832	George Sand	*Indiana*	115,415
1923	Delly	*Dans les ruines*	98,542

Table 8: Annotated Dataset Details