Improving Coreference Resolution through Prompting-based Adversarial Filtering and Data Augmentation

Anonymous ACL submission

Abstract

Coreference resolution is a fundamental task in natural language processing that involves linking different references to the same entity within a text. However, models often struggle to reliably identify referential relationships, particularly in cases involving long contexts or complex modifiers. To address these challenges, this study introduces a data augmentation technique that incorporates adjectival phrases and employs a Prompting-based Adversarial Filtering pipeline. Specifically, we generated and inserted contextually appropriate adjective phrases through the interaction between GPT-4o-mini based Few-shot Prompting and a Discriminative Language Model. These augmentations were then verified for grammaticality and contextual coherence through human evaluation. The resulting synthetic dataset was integrated with the original data to enhance the performance of coreference resolution. Training real-world models with the synthetic dataset led to up to a 1.2% improvement in CoNLL-F1 on the LitBank dataset and up to a 0.4% improvement on the PreCo dataset. Furthermore, the synthetic dataset significantly increased the diversity and complexity of coreference relations. The proposed pipeline represents an important step towards developing coreference resolution models that better capture the linguistic diversity of natural language and demonstrate robustness under challenging conditions.

1 Introduction

Coreference resolution (Karttunen, 1969; Ng and Cardie, 2002) is a fundamental challenge in natural language processing, requiring the accurate identification and linking of multiple mentions referring to the same entity within a document. It plays a crucial role in applications such as pronoun resolution (Zhang et al., 2020), information retrieval, document summarization, question answering, and dialogue systems (Joshi et al., 2020). While recent advances in pre-trained Large Language Models based on the Transformer architecture (Vaswani et al., 2017) have significantly improved performance, challenges remain, particularly in scenarios requiring long-range contextual reasoning or the interpretation of complex lexical structures. Existing datasets for coreference resolution (Pradhan et al., 2013; Chen et al., 2018; Bamman et al., 2019) are often based on relatively simple sentence structures and expressions, constraining the ability of models to learn more linguistically diverse patterns, such as those involving adjectives and adverbial phrases. These more intricate expressions are especially prevalent in literary texts, and the inability to learn them effectively can substantially impair the generalization performance of a model. This limitation is further exacerbated in real-world applications, where models frequently encounter highly modified and contextually complex language, making robust coreference resolution even more challenging.

To address these issues, recent studies have explored data augmentation (Feng et al., 2021) and adversarial filtering (Bras et al., 2020). Data augmentation is a well-established technique that exposes models to a variety of linguistic patterns, reducing their reliance on specific expressions or biased features. Adversarial filtering, in contrast, generates and curates sophisticated example variants, encouraging models to learn linguistic cues and complex relationships that might otherwise be overlooked. There is also growing interest in combining adversarial filtering with data augmentation to systematically adjust dataset difficulty and mitigate model weaknesses (Bhargava and Ng, 2022).

However, existing research often focuses on techniques such as synonym substitution (Pellicer et al., 2023), sentence reordering, and noise injection to generate challenging examples, even within adversarial filtering frameworks. While these techniques are effective for generating difficult-to-distinguish

examples, they fall short in tasks like coreference resolution, where context preservation, referential integrity, and entity recognition are crucial. For instance, a model cannot inherently recognize that "the city" and "the breathtakingly vibrant city" refer to the same entity. Thus, there is a clear need for methods that intentionally introduce multi-layered modifiers, such as adjectives and adverbial phrases, to enrich linguistic features. This approach enables the model to perform coreference resolution based on contextual understanding and referential reasoning rather than relying on simple keyword matching.

To address this issue, we propose a modifier-oriented data augmentation strategy to enable models to learn complex expressions and systematically validate its effectiveness. The main contributions of this study are as follows: (1) To complement the monotonous representation of existing coreference resolution datasets, we introduce examples with modifier phrases to expand learning opportunities for complex coreference relations. (2) We design a Prompting-based Adversarial Filtering pipeline that utilizes GPT-4o-mini (Radford et al., 2018, 2019; Brown et al., 2020) as a Generator Language Model, proposing a data selection method that considers both contextual relevance and difficulty. (3) We construct a synthetic dataset by integrating the augmented dataset with the original data and fine-tuning a pre-trained language model, which significantly improves the F1 score of coreference resolution models. By combining coreference resolution research with data augmentation techniques, this study introduces a novel approach that simultaneously enhances model performance and data quality.

2 Related Works

2.1 Coreference Resolution

Coreference resolution is the task of identifying and linking multiple expressions that refer to the same entity within a text (Karttunen, 1969). It is broadly categorized into entity coreference resolution and event coreference resolution. In this study, we focus on entity coreference resolution, which involves identifying groups of expressions that refer to the same real-world entity (Haghighi and Klein, 2010). Coreference resolution typically comprises two key steps: mention detection and mention linking (Pradhan et al., 2012). Mention detection involves identifying expressions in the text that can serve as entity mentions, while mention linking clusters these detected mentions into the appropriate coreference groups, ensuring they refer to the same entity (Lee et al., 2017).

Performance evaluation in coreference resolution primarily relies on the F1 score, a harmonic mean of precision and recall that provides a comprehensive measure of both mention detection and linking performance (Cai and Strube, 2010). Several additional metrics are also employed, including MUC (Vilain et al., 1995), which assesses the degree of overlap between gold-standard clusters and predicted clusters based on coreference links, and B^3 (B-Cubed) (Bagga and Baldwin, 1998), which computes precision and recall at the mention level and applies a weighted averaging scheme. The Constrained Entity Aligned F-measure (CEAFe) (Luo, 2005) evaluates coreference accuracy through a one-to-one mapping between gold-standard and predicted clusters.

Several benchmark datasets are widely used for coreference resolution, including CoNLL 2012 (Pradhan et al., 2012), GAP (Webster et al., 2018), LitBank (Bamman et al., 2019), and WikiCoref (Ghaddar and Langlais, 2016). CoNLL 2012 covers multiple languages, including English, Chinese, and Arabic, and spans various text genres. GAP comprises sentence pairs containing gender-ambiguous pronouns extracted from Wikipedia articles. LitBank provides fine-grained coreference annotations for literary texts, whereas WikiCoref is annotated with both entity types and coreference links from Wikipedia corpora.

Coreference resolution models can be broadly categorized based on their learning paradigms into mention-pair classifiers (Haghighi and Klein, 2010), entity-level models (Clark and Manning, 2016), latent-tree models (Fernandes et al., 2014), and mention-ranking models (Wiseman et al., 2016). More recently, deep learning and transformer-based large language (Vaswani et al., 2017) models have been introduced to further enhance coreference resolution performance. However, challenges remain in handling complex contextual dependencies and modifier phrases.

2.2 Adversarial Filtering

Adversarial filtering is a method designed to address model limitations by increasing dataset complexity. It originates from the concept of adversarial examples, which are intentionally crafted inputs that induce a model to produce incorrect predic-

Dataset	#Train	#Dev	#Test
LitBank	80	10	10
PreCo	36,120	500	500

Table 1: Number of documents in Litbank and PreCo

Dataset	#Best	#Weird	#Worst
Augmented LitBank	215	93	27
Augmented PreCo	4,029	1,371	1,193

Table 2: Number of cases in Litbank and PreCo augmented data

tions (Bras et al., 2020). In other words, subtle variations in input data are introduced to deliberately mislead the model, thereby encouraging it to learn more robust representations rather than relying on superficial patterns. This concept is closely related to adversarial training (Goodfellow et al., 2015), a methodology in which adversarial examples are incorporated into the training process alongside original data to enhance model robustness against input variations. Adversarial training iteratively refines the model, ensuring it remains resilient to perturbed inputs. Recent advancements in adversarial training have demonstrated its effectiveness in various domains (Cheng et al., 2022). DISCOSENSE (Bhargava and Ng, 2022) extends the adversarial filtering framework by introducing Controlled Adversarial Filtering, leveraging discourse connectives to assess commonsense reasoning abilities and generating adversarial distractors to increase evaluation difficulty.

Specifically, we employ GPT-4o-mini (Radford et al., 2018, 2019; Brown et al., 2020) to generate, insert, and replace adjectival phrases in coreference expressions. These modified instances are then filtered using a discriminative language model to construct a more challenging dataset. Through this approach, we aim to simultaneously enhance both the performance and robustness of coreference resolution models by exposing them to more complex linguistic patterns.

3 Methodology

3.1 Task Description

Coreference resolution is the task of identifying and linking multiple mentions of the same entity within a given text (Karttunen, 1969). In this study, we generate an adversarial example dataset by augmenting correctly predicted instances with adjectival phrases. The adversarial dataset is then combined with the original data to construct the final synthetic dataset. By training the model with this synthetic dataset, we aim to enhance coreference resolution performance.

3.2 Dataset Format

OntoNotes Formatting The OntoNotes dataset (Pradhan et al., 2013) is structured as a collection of documents, each containing multiple sentences. Sentences are represented as word-wise partitioned lists, and an entire document consists of an aggregation of these sentence lists. This hierarchical structure facilitates contextualization and enables effective modeling of document-level coreference relationships.

Cluster Structure A coreference cluster is defined as a set of mention offsets that refer to the same entity. Each offset specifies the start and end indices of a particular word or phrase within a document, uniquely identifying its occurrence. Offsets corresponding to the same coreference relation are grouped into clusters, allowing the model to learn and distinguish different coreference relationships.

Augmented Descriptive Phrase Structure In this study, we leverage a generative language model to expand the scope and complexity of the dataset by incorporating descriptive phrases into coreferential noun phrases. For instance, if the noun phrase "the city" appears in a sentence, an adjectival phrase such as "the beautiful city" is introduced to enhance linguistic diversity while preserving the coreference relationship.

Details on each raw data and each augmented data are provided in Table 1 and Table 2. Table 1 presents the number of train, development, and test cases from the LitBank and PreCo datasets used for model fine-tuning, while Table 2 shows the distribution of datasets obtained by augmenting the train datasets from LitBank and PreCo. The three evaluation criteria are described in Section 3.5.

3.3 Datasets

LitBank (Bamman et al., 2019) is an annotated dataset comprising 100 works of English literature, widely utilized in NLP and computational humanities. It specializes in literary texts, containing documents with long contextual spans and complex narrative structures. These characteristics enable a more sophisticated evaluation of coreference res-

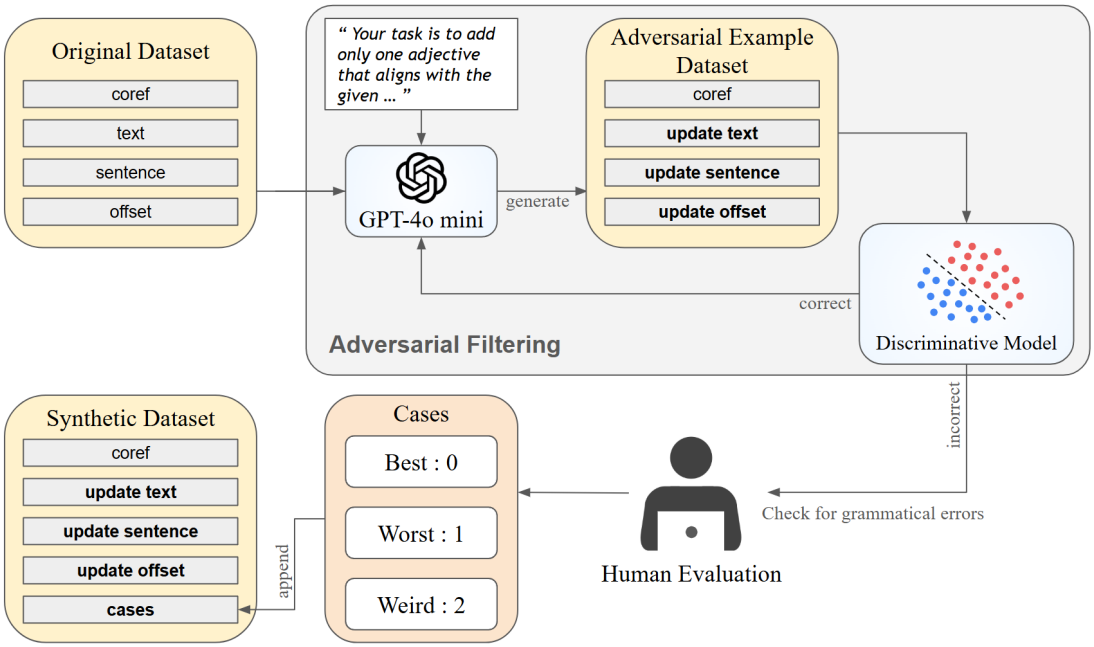

Figure 1: Overall pipeline. The gray rectangle represents the Prompting-based Adversarial Filtering process. If the discriminative model succeeds in making a prediction, the process repeats; otherwise, the data is collected and moved to the human evaluation phase.

olution models that must process long-range dependencies. Unlike general-domain texts such as conversational transcripts or news articles, literary texts are distinguished by their stylistic diversity, frequent use of metaphors, and long-range dependencies. Because of these features, LitBank is particularly well-suited for assessing a model's long-range inference capabilities and anti-forgetting performance in long documents with intricate coreference structures. It is frequently employed in research for character tracking, event extraction, relationship modeling, and literary analysis.

PreCo (Chen et al., 2018) is a large-scale English dataset derived from middle and high school reading comprehension test questions. The sentences are primarily extracted from educational reading comprehension passages and encompass a range of reference structures, from simple pronominal expressions to more complex coreference patterns. Due to its large-scale composition, PreCo contains a substantial number of sentences and words, making it well-suited for training large-scale models with high data requirements. Because the dataset is sourced from middle and high school reading materials, many sentences follow standardized grammatical structures. However, some also exhibit complex polysemy and syntactic patterns, contributing to diversity in sentence structure. Additionally, PreCo includes examples of varying difficulty, allowing for comprehensive model evaluation. This enables assessments ranging from basic pronoun resolution to more advanced tasks such as learning compound reference structures.

These two datasets differ in domain characteristics and text structures, making them complementary in evaluating the generalization performance of coreference resolution models. Specifically, LitBank emphasizes long-range inference and linguistic complexity in literary texts, whereas PreCo focuses on scale and a diverse range of difficulty levels derived from educational texts.

3.4 Prompting-based Adversarial Filtering

The data augmentation pipeline proposed extends the concept of adversarial filtering to coreference resolution, emphasizing the interaction between a discriminative language model and a generator language model. This pipeline is designed to enhance the generalization of model performance and robustness by incrementally introducing adversarial examples, such as descriptive phrases, into the coreference resolution dataset through the generator model. The modified dataset is then processed by the discriminative model, which filters the generated data to regulate quality and adjust difficulty levels.

Discriminative models predict coreference relationships from input data and compare them

to gold-standard annotations to identify instances where the model already makes correct inferences. In this study, we employ Maverick-mes (Martinelli et al., 2024) as the discriminative model. The generator model increases the complexity of the dataset by adding or replacing descriptive phrases before coreference expressions. The newly generated examples are then validated by the discriminative model. For this purpose, GPT-4o-mini is utilized as the generator model. To ensure that the generator model accurately determines the appropriate placement and integration of descriptive phrases, we provide explicit examples within the prompts to facilitate the generation of more natural and contextually appropriate adjectival phrases. Furthermore, we develop an automated pipeline to generate modified data based on the prompts, which is subsequently validated and filtered using the discriminative model. Figure 1 illustrates the complete process of Prompting-based Adversarial Filtering. Starting with the original dataset, the generator model inserts appropriate descriptive phrases before coreference expressions.

3.5 Human Evaluation

We conducted a human evaluation to assess the quality of the data generated by the Few-shot Prompt-based Adversarial Filtering process. This evaluation was essential to directly verify the grammatical correctness, semantic appropriateness, and relevance of coreferencence of the augmented data. Three researchers performed the evaluation based on predefined criteria, systematically reviewing all augmented datasets produced through the adversarial filtering process. The assessment focused on the grammatical completeness, semantic relevance of descriptive phrases to nouns, and overall quality of data transformation. The evaluation criteria are presented in Table 7 in the Appendix.

4 Experiments

4.1 Models

Maverick-incr is a coreference resolution model based on the Shift-Reduce Paradigm (Clark and Manning, 2016) that incrementally updates the clusters formed in the previous step. The model processes text sequentially and determines whether newly emerged mentions can be linked to existing clusters. If a mention can be included in an existing cluster, it is merged. Otherwise, a new cluster is created to maintain the coreference relationship. Unlike traditional sentence-by-sentence approaches, Maverick-incr favors real-time and sequential processing, making it particularly well-suited for coreference resolution in streaming data or interactive environments where incremental inference is required.

Maverick-s2e is a coreference resolution model based on the Coarse-to-Fine method (Lee et al., 2017). This approach consists of two steps: mention extraction and mention-antecedent classification. In the mention extraction step, the model identifies potential mentions in the text that can be part of a coreference chain. In the next step, the hidden state corresponding to the start and end tokens of an antecedent candidate mention is compared to classify whether it refers to the same entity. Mentions identified as coreferential are grouped into clusters. This two-step approach improves inference efficiency by first narrowing down candidate mentions before applying a more refined classification, avoiding the need for computationally expensive contextual processing.

Maverick-mes follows the same Coarse-to-Fine-based structure as Maverick-s2e but introduces a Multi-Expert Scorer instead of a Mention-Pair Scorer to refine linguistic pattern recognition. Specifically, it defines six linguistic synchronization categories—PRON-PRON-C, PRON-PRON-NC, ENT-PRON, MATCH, CONTAINS, and OTHER—, determines which category a mention belongs to, and computes a score for each category to form clusters. This approach enhances coreference resolution by pre-typing linguistic features such as pronoun-pronoun agreement, noun phrase-pronoun relations, and partial inclusion relationships.

4.2 Evaluation Metric

4.2.1 MUC(Mention-Unicon Cross)

MUC(Vilain et al., 1995) is a metric that evaluates coreference resolution based on the precision and recall of coreference links. Calculated by comparing the number of links between clusters and assessing how accurately the predicted cluster connections align with the gold standard clusters.

$$MUC_{Precision} = \frac{TP}{TP + FP}$$

$$MUC_{Recall} = \frac{TP}{TP + FN}$$

$$MUC_{F1} = 2 \cdot \frac{MUC_{Precision} \cdot MUC_{Recall}}{MUC_{Precision} + MUC_{Recall}}$$

- TP (True Positives): Correctly predicted links in coreference clusters.
- FP (False Positives): Predicted links that do not exist in the gold standard clusters.
- FN (False Negatives): Links that exist in the gold standard clusters but are missing in the predictions.

4.2.2 B-Cubed (B³)

B³ (Bagga and Baldwin, 1998) evaluates coreference resolution by measuring the precision and recall of individual mentions and computing a weighted average to assess how consistently each mention is assigned to the correct cluster. A model achieves a high score only if it excels in both accurate classification (precision) and error-free retrieval (recall) of mentions.

$$B^3_{\text{Precision}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_i \cap G_i|^2}{|C_i|}$$

$$B^3_{\text{Recall}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|C_i \cap G_i|^2}{|G_i|}$$

$$B^3_{\text{F1}} = 2 \cdot \frac{B^3_{\text{Precision}} \cdot B^3_{\text{Recall}}}{B^3_{\text{Precision}} + B^3_{\text{Recall}}}$$

- C_i: Predicted cluster containing the i-th mention.
- G_i: Gold cluster containing the i-th mention.
- N: Total number of mentions.
- $|C_i \cap G_i|$: Number of mentions shared between the predicted and gold clusters.

4.2.3 CEAFe(Constrained Entity Alignment F-Measure)

CEAFe(Luo, 2005) evaluates coreference resolution based on a one-to-one mapping between clusters. If a gold-standard cluster is split into multiple predicted clusters or merged into a single predicted cluster, the score penalization is significant.

$$\text{Similarity}(C, G) = \sum_{(c,g) \in \text{Optimal Matching}} \phi(c, g)$$

$$\text{CEAFe}_{\text{Precision}} = \frac{\text{Similarity}(C, G)}{|C|}$$

$$\text{CEAFe}_{\text{Recall}} = \frac{\text{Similarity}(C, G)}{|G|}$$

$$\text{CEAFe}_{\text{F1}} = 2 \cdot \frac{\text{CEAFe}_{\text{Precision}} \cdot \text{CEAFe}_{\text{Recall}}}{\text{CEAFe}_{\text{Precision}} + \text{CEAFe}_{\text{Recall}}}$$

$$\phi(c, g) = \frac{2 \cdot |c \cap g|}{|c| + |g|}$$

- C: Set of predicted clusters.
- G: Set of gold clusters.
- $|C|$: Number of predicted clusters.
- $|G|$: Number of gold clusters.
- $\phi(c, g)$: Similarity between a predicted cluster c and a gold cluster g.

4.2.4 CoNLL-2012 F1 Score

CoNLL-2012 (Pradhan et al., 2012) F1 Score is calculated as the mean of three F1 scores.

$$\text{CoNLL-2012}_{\text{F1}} = \frac{\text{MUC}_{\text{F1}} + B^3_{\text{F1}} + \text{CEAFe}_{\text{F1}}}{3}$$

4.3 Setup

We utilized DeBERTa-v3(He et al., 2023) as the document encoder for the discriminative language model. DeBERTa improves upon the existing BERT architecture by introducing a disentangled attention mechanism and enhances contextual understanding through an improved lexical embedding method. For optimization, Adafactor (Shazeer and Stern, 2018) was employed with weight decay set to 0.01. The learning rate was configured as 3e-4 for linear layers and 2e-5 for the pretrained encoder. The LitBank dataset was trained for 300 epochs due to its small training size, which results in slow improvements in validation performance per epoch. In contrast, the PreCo dataset, which contains approximately 40,000 training samples, was trained for only 5 epochs as the large dataset size facilitates faster convergence. All training was conducted on an RTX 4090 GPU with 24GB of VRAM.

Datasets	Model	MUC	B^3	CEAFe	CoNLL-F1
Original LitBank	Maverick-incr	85.5	73.0	68.6	75.7
Original LitBank	Maverick-s2e	87.1	74.2	66.2	75.8
Original LitBank	Maverick-mes	86.6	75.2	65.6	75.8
Augmented LitBank	Maverick-incr	84.1	71.0	67.2	74.1
Augmented LitBank	Maverick-s2e	87.0	75.0	66.1	76.0
Augmented LitBank	Maverick-mes	86.3	74.6	63.6	74.8
Synthetic LitBank	Maverick-incr	**85.8**	**73.6**	**71.2**	**76.9**
Synthetic LitBank	Maverick-s2e	**87.1**	**75.9**	**68.1**	**77.1**
Synthetic LitBank	Maverick-mes	**87.3**	**76.2**	**66.9**	**76.8**

Table 3: Performance of four evaluation metrics for the Maverick model on the LitBank dataset and the augmented LitBank dataset.

Datasets	Model	MUC	B^3	CEAFe	CoNLL-F1
Original PreCo	Maverick-s2e	89.2	88.5	84.6	87.4
Original PreCo	Maverick-mes	88.6	88.2	84.6	87.1
Synthetic PreCo	Maverick-s2e	**89.5**	**89.0**	**85.3**	**87.9**
Synthetic PreCo	Maverick-mes	**89.0**	**88.6**	**84.8**	**87.4**

Table 4: Performance of four evaluation metrics for the Maverick model on the PreCo dataset and the augmented PreCo dataset.

5 Results

In this section, we compare and analyze the MUC, B^3, CEAFe, and CoNLL-F1 scores of Maverick-incr, Maverick-s2e, and Maverick-mes models trained on the LitBank and PreCo datasets. Each metric evaluates coreference resolution from a different perspective: link-based (MUC), mention-level (B^3), and cluster alignment (CEAFe). By examining these metrics, we assess how the inclusion of adjectival phrases in the synthetic dataset contributes to performance improvements. Detailed performance results are presented in Table 3 and Table 4. Performance is compared across models, with the best results highlighted in bold.

5.1 Performance Comparison

5.1.1 MUC (Link-based Evaluation)

MUC score increased more than other metrics in Maverick-mes for both LitBank and PreCo, by 0.7% and 0.4%, respectively. The MUC metric is determined by the number of correctly identified coreference cluster links, and since the augmented data improves qualitatively rather than quantitatively, this metric is particularly suited for mes models. Specifically, Maverick-mes, which relies on part-of-speech-based features, links coreference to pronouns or proper nouns, making it more responsive to the nature of the augmented data. In contrast, Maverick-incr, which merges or splits coreference clusters, evaluates coreference relations qualitatively, while Maverick-s2e, which identifies the beginning and end of coreference words, primarily evaluates them quantitatively. As a result, the MUC metric demonstrates that the augmented data is more effective for mes models than for incr and s2e models.

5.1.2 B-Cubed (Mention-based Evaluation)

B^3 score increased by 1.8% and 0.5% on Maverick-s2e for LitBank and PreCo, respectively. This improvement is attributed to the fact that B^3 performs better when mentions within each cluster are matched exactly. Since the augmented dataset includes descriptive phrases, it is evident that the s2e model, which constructs clusters by identifying the start and end of coreference mentions, achieves higher B^3 scores. Similarly, the mes model, which shares a structural similarity with s2e, exhibits a greater increase in B^3 performance than the incr model on LitBank.

5.1.3 CEAFe (one-to-one Cluster Alignment)

CEAFe outperformed the Maverick-incr model on LitBank by 2.6% and the Maverick-s2e model on PreCo by 0.7%. As a metric that measures the

similarity between predicted and correct clusters, CEAFe is particularly well-suited for incr, which incrementally expands each cluster, and s2e, which effectively aligns coreference mentions with the correct cluster with high probability. These results indicate that the model improves its ability to predict clusters closer to the ground truth by learning complex cases through the augmentation of descriptive phrases.

5.2 Discussion

The CoNLL-2012 F1 score, which represents the average of the three coreference resolution metrics, increases across all synthetic datasets and models. While the improvement ranges from 0.3% to 0.5% for PreCo, it exceeds 1.0% for LitBank, which has a smaller dataset size. Notably, the largest performance gains are observed in B^3 and CEAFe, suggesting that the inclusion of descriptive phrases has a significant impact on coreference resolution.

However, training on the augmented dataset alone may lead to a decline in performance. For the augmented PreCo dataset, we did not conduct an ablation study due to the relatively small number of augmented cases compared to the original data. However, for LitBank, we trained models exclusively on the augmented data and observed that Maverick-s2e performed worse than the original dataset across all metrics, except for a 0.8% improvement in B^3. This degradation occurs because the augmented dataset contains only high-quality cases, excluding many coreference clusters that involve pronouns or proper nouns. Descriptive phrases that modify pronouns or proper nouns are often grammatically ill-formed, leading to their rejection by human evaluators. Consequently, the exclusion of these cases resulted in inferior performance when training on the augmented dataset alone compared to the original dataset.

6 Conclusion

We propose a new benchmark dataset for coreference resolution models to learn complex and challenging reference relations through data augmentation techniques incorporating descriptive phrases. Specifically, we introduce the Prompting-based Adversarial Filtering technique, which integrates GPT-4o-mini based Few-shot Prompting with Adversarial Filtering, and design the augmentation process to capture diverse linguistic characteristics that are often underrepresented in existing datasets. By leveraging the interaction of a Discriminative Language Model, we generate and insert contextually natural yet challenging descriptive phrases, verify their grammatical and semantic appropriateness through human evaluation, and integrate them with the original data to construct the final synthetic dataset.

The experimental results demonstrate that training models on the synthetic dataset consistently improves performance across all evaluation metrics. In particular, CoNLL-F1 scores increased by up to 1.3% on the LitBank dataset and up to 0.5% on the PreCo dataset, indicating that models trained with augmented data effectively learn linguistic diversity and complex descriptive structures, which are difficult to acquire from conventional datasets. Furthermore, model-specific improvements were observed: Maverick-incr achieved the highest gains in CEAFe, benefiting from its incremental clustering approach, while Maverick-s2e exhibited significant improvements in B^3. These results suggest that synthetic datasets diversify coreference relationships and provide metric-dependent advantages based on the structural characteristics of each model.

However, unlike LitBank, the PreCo dataset contained a relatively smaller proportion of augmented data and initially exhibited higher baseline performance, which somewhat limited the overall performance gains. Nevertheless, we observed consistent improvements of 0.3% to 0.5% across all models, confirming that our data augmentation approach is domain-independent and contributes to performance enhancement across different datasets.

In summary, the Prompting-based Adversarial Filtering pipeline proposed in this study demonstrates its effectiveness in improving the accuracy and generalization of coreference resolution models by mitigating reliance on simple patterns and enhancing linguistic diversity. Future research should focus on scaling up augmentation by increasing both the volume and linguistic variety of descriptive phrases, expanding the range of part-of-speech modifications, and applying this approach to multilingual corpora and various NLP tasks. Such advancements are expected to further improve model robustness across a broader spectrum of linguistic phenomena.

7 Limitations

First, due to computational resource constraints, we were unable to fully evaluate the Maverick-

incr model on the PreCo dataset. This limitation restricted direct performance comparisons between models and hindered precise performance validation. Additionally, we encountered human resource constraints during the human evaluation phase. While human evaluation is essential for ensuring data quality, it is resource-intensive. With only three evaluators, a significant portion of the augmented data could not be manually reviewed. Furthermore, the percentage of augmented data in the PreCo dataset was considerably lower than in the LitBank dataset, potentially limiting the scope of performance improvements observed in PreCo. Future research should expand the dataset to encompass a wider range of linguistic structures and apply our techniques to multiple NLP models to further assess their generality and effectiveness. Increasing the number of human evaluators would enhance the reliability of qualitative assessments, while optimizing the ratio of augmented data would help ensure diversity and balance across the dataset. Despite these limitations, the dataset and augmentation methodology proposed in this study represent a significant step toward improving linguistic diversity and model robustness in coreference resolution. Moreover, this research provides valuable insights for the development of more sophisticated NLP models in the future.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

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.

Prajjwal Bhargava and Vincent Ng. 2022. DiscoSense: Commonsense reasoning with discourse connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10295–10310, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *CoRR*, abs/2002.04108.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, page 28–36, USA. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. 2022. Cat: Customized adversarial training for improved robustness. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 673–679. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2014. Latent trees for coreference resolution. *Computational Linguistics*, 40(4):801–835.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: 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).

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *Preprint*, arXiv:1412.6572.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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

Lauri Karttunen. 1969. Discourse referents. In *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*, Sånga Säby, Sweden.

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.

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.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

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

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. A brief survey and comparative study of recent development of pronoun coreference resolution. *Preprint*, arXiv:2009.12721.

A Example of augmented sentence

Table 5 provides an example of the best case from our augmented LitBank dataset. Although the dataset is organized in OntoNotes format, we present it here in standard sentence format for readability. The underlined words indicate coreference mentions, while bolded words represent augmented descriptive phrases.

Input (Source Sentence): On either side of this road straggled two uneven rows of wooden buildings ; the general merchandise stores , the two banks , the drug store , the feed store , the saloon , the post-office . On the sidewalk in front of one of the stores sat a little Swede boy , crying bitterly . **Output (Augmented Sentence):** On either side of this road straggled two uneven rows of wooden buildings ; the general merchandise stores , the two banks , the drug store , the feed store , the saloon , the post-office . On the sidewalk in front of one of the **various** stores sat a little Swede boy , crying bitterly .

Table 5: Example of Augmented Sentence.

B Prompt Templeate for Adversarial Filtering

When using GPT-4o-mini to augment descriptive phrases, it is essential to identify coreference mentions in a given sentence and add modifiers only to those mentions. In doing so, the following considerations should be taken into account when selecting modifiers:

- Avoid repeating the same modifier within a sentence.
- Do not use overly generic modifiers.
- Modifiers should not alter the original meaning of the sentence.

The first issue arises from repeating the same word, which can make the sentence structure awkward and potentially grammatically incorrect. Nevertheless, we excluded repeated modifiers during human evaluation to maintain naturalness. The second issue is that overly generic modifiers fail to contribute meaningfully to identifying coreference mentions, contradicting the purpose of our augmentation strategy. To address this, we instructed annotators to select contextually relevant modifiers derived from the given sentence that do not compromise its original meaning. Detailed prompts for modifier generation are provided in Table 6.

Generate Adjective Prompt for GPT-4o-mini

Instructions:
You will be given a sentence in OntoNotes format along with a coreference cluster and its offsets. Your task is to add **only one adjective** that aligns with the given coreference term. The adjective must be placed **immediately before** the term within the sentence.

Guidelines:
1. Identify the words in the sentence that correspond to each offset.
2. Updated Coreference Offsets should be calculated step by step.
3. For each remaining term (starting from the second), add **only one adjective** **immediately before** the term if it adds meaningful context.
4. **Never add articles ('the', 'a')**, only one adjective.
5. Ensure the adjective does not change the sentence's original meaning.
6. **Avoid repeating the same word multiple times in sequence** (e.g., avoid adding 'large' twice in a row like 'large large').
7. Use adjectives that are contextually relevant and meaningful. Avoid using too general adjectives like 'good', 'bad', 'nice', or nonsensical combinations.
8. Adjectives should enrich the meaning or add useful information without making the description redundant or awkward.
9. **If no suitable adjective can be added without disrupting the meaning or creating redundancy, do not add an adjective at all.** The coreference term should remain unchanged in such cases.
10. **NEVER VIOLATE THE OUTPUT TEMPLATE**

Input:
- Sentence: ontonotes_sentence
- Coreference Offsets: offsets
- Coreference Words: words

Output Format:
1. Updated Coreference Words : The modified OntoNotes format sentence with adjectives added.

Example:
Input:
- Sentence: ['Barack', 'Obama', 'is', 'traveling', 'to', 'Rome', '.', 'The', 'city', 'is', 'sunny', 'and', 'the', 'president', 'plans', 'to', 'visit', 'its', 'most', 'important', 'attractions']
- Coreference Offsets: [[5, 5], [7, 8], [17, 17]]
- Coreference Words: [['Rome'], ['The', 'city'], ['its']]
Correct Output:
1. Updated Coreference Words : [['Rome'], ['The', 'picturesque', 'city'], ['its']]
Explanation:
- 'picturesque' was added to 'city' to enrich the description without altering the intended meaning.
- No adjective was added to 'Rome' or 'its' as it was unnecessary.
Now, process the following input.

Table 6: Generate adjective prompt for Adversarial Filtering using GPT-4o-mini

Case	Criteria	Original Sentence	Augment Sentence	Explanation
High-Quality (Best)	The sentence must be grammatically correct while incorporating descriptive phrases that are semantically relevant to the coreferential clusters.	"The man went to the store."	"The diligent man went to the store."	A contextually relevant descriptive phrase, 'diligent,' was added before the coreferential word 'man'.
Unacceptable (Worst)	The augmented descriptive phrases are either grammatically incorrect or not suitable for coreference clusters.	"My name is Jim."	"My name is enchanting Jim."	The descriptive phrase 'enchanting' is inappropriate, making it difficult to establish a coreference cluster with 'Jim'.
Acceptable but Semantically Misaligned (Weird)	The sentence is grammatically correct, but the descriptive phrases or synonyms used as a replacement are semantically inappropriate for the coreferential clusters.	"The cat jumped onto the couch."	"The shiny feline jumped onto the couch."	The adjective 'shiny' is contextually inappropriate for the coreferential word 'cat,' and the original term has been replaced with its synonym 'feline.'

Table 7: Augmented examples classified into best, weird, and worst cases according to evaluation criteria with corresponding explanations.