BenCoref: A Multi-Domain Dataset of Nominal Phrases and Pronominal Reference Annotations

Shadman Rohan[1], Mojammel Hossain[1], Mohammad Mamun Or Rashid[2], Nabeel Mohammed[1]

[1]North South University [2]Jahangirnagar University
shadman.rohan, mojammel.hossain, nabeel.mohammed@northsouth.edu
mamunbd@juniv.edu

Abstract

Coreference Resolution is a well studied problem in NLP. While widely studied for English and other resource-rich languages, research on coreference resolution in Bengali largely remains unexplored due to the absence of relevant datasets. Bengali, being a low-resource language, exhibits greater morphological richness compared to English. In this article, we introduce a new dataset, BenCoref, comprising coreference annotations for Bengali texts gathered from four distinct domains. This relatively small dataset contains 5200 mention annotations forming 502 mention clusters within 48,569 tokens. We describe the process of creating this dataset and report performance of multiple models trained using BenCoref. We expect that our work provides some valuable insights on the variations in coreference phenomena across several domains in Bengali and encourages the development of additional resources for Bengali. Furthermore, we found poor crosslingual performance at zero-shot setting from English, highlighting the need for more language-specific resources for this task. The dataset is available at [1].

1 Introduction

Coreference resolution is the task of identifying all references to the same entity in a document. This task originally started as a sub-task of information extraction. The Message Understanding Conferences (Grishman and Sundheim, 1996) first introduced three tasks, collectively referred to as SemEval, designed to measure the deeper understanding of any information extraction (IE) system. One of these three tasks proposed in the event was coreferencial noun phrase identification.

The Automatic Content Extraction (ACE) Program (Doddington et al., 2004) was the first major initiative that created a large dataset with entity, event and relation annotations. This project revealed some major complexities behind creating such dataset. Some of the significant challenges reported by the annotators include the coreference of generic entities, use of metonymy, characterization of Geo-Political Entity, distinguishing certain complex relations, and recognizing implicit vs. explicit relations.

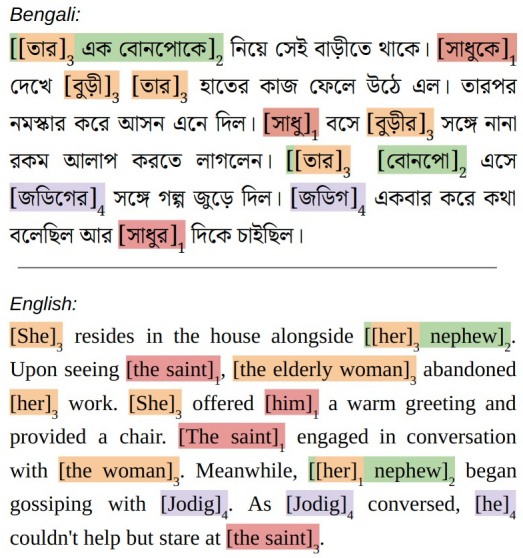

Figure 1: BenCoref annotations with color-coded Coreference chains.

Since then coreference resolution, anaphoric & cataphoric relation identification, event reference detection has been studied widely. As a result, large datasets like ACE (Doddington et al., 2004), Ontonotes (Pradhan et al., 2012), WikiCoref (Ghaddar and Langlais, 2016), and LitBank (Bamman et al., 2020) were made public. Some datasets, like ACE (Doddington et al., 2004), and Ontonotes, expanded this task beyond English to include more languages, like Arabic, and Chinese.

This coreference resolution task has shown potential in improving many downstream NLP tasks

[1]codes used to generate the results along with data is available at: https://github.com/ShadmanRohan/BenCoref

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 104–117

like machine translation (Miculicich Werlen and Popescu-Belis, 2017; Ohtani et al., 2019), literary analysis (Bamman et al., 2014), question anwering (Morton, 1999), text summarization (Steinberger et al., 2007), etc. However, Bengali, despite being a popular lanaguage, has seen very little work is this direction due to lack of public datasets.

Figure 1 shows a sample from our dataset with each color representing an unique entity. The main contributions of this work are:

- We introduce a new Bengali coreference annotated dataset, consisting of 48,569 tokens collected from four diverse domains. Our dataset creation process is shared along with the annotators' guidelines, which we believe is the first of its kind for Bengali coreference annotation.
- We characterize the behaviour and distribution of nominal and pronominal coreference mentions across the four domains with necessary statistics. Furthermore, we report the performance of an end-to-end neural coreference resolution system that was solely built using our data.
- We empirically demonstrate the necessity for more language-specific datasets, particularly for low-resource languages, by comparing our results with zero-shot cross-lingual learning from English.

1.1 Related Datasets

To the best of our knowledge, no coreference dataset in Bengali exists. Most of the works related to Bengali (Sikdar et al., 2013; Senapati and Garain, 2013; Sikdar et al., 2015) uses data from ICON2011 shared task which was never publicly shared.

Most of the major coreference datasets are in English. OntoNotes (Pradhan et al., 2012) is a well-annotated and large dataset with over 1.6M words. This dataset does not contain any singleton mention. Later, LitBank (Bamman et al., 2020) was published that is almost 10 times larger than OntoNotes (12.3M words).

2 Challenges in Bengali

One of the main challenges we faced was the absence of preexisting coreference annotation guidelines tailored for the Bengali language. To overcome this obstacle, we adapted the OntoNotes coreference annotation guideline to suit our objectives. This highlighted several distinctive linguistic characteristics of Bengali, such as zero anaphora, non-anaphoric pronouns, and case-marking, that needs to be carefully considered when preforming co-reference annotation in Bengali. Each of this is discussed with more details and examples in Figure 9 in the Appendix.

Moreover, we discovered that existing annotation software is ill-equipped to manage Bengali text, occasionally leading to inaccurate rendering and unstable character display. This underscores the importance of advancing normalization techniques and standardization of Bengali digital representation.

3 Data Domain Description

The Bengali language can be braoadly categorized into two primary literary dialects, namely "Shadhubhasha" and "Choltibhasha." "Shadhubhasha" was commonly used by Bangla writers and individuals in the 19th and early 20th centuries, while "Choltibhasha" is currently the more prevalent and colloquial dialect. This dataset contains both domains of Bengali text, with story and novel texts sourced from copyright-free books of the 19th and 20th centuries, and biography and descriptive texts obtained from modern sources, primarily in "Choltibhasha." A brief description of each domain is given below:

3.1 Biography

A biography presents a comprehensive account of an individual's life, character, accomplishments, and works, spanning from birth to death or the present time. Although the number of references per document in biographical texts is comparable to other genres, they primarily focus on a single subject throughout the entire narrative. Additionally, the dialect employed in biographies in BenCoref is typically "Choltibhasha."

3.2 Descriptive

By descriptive text we refer to wikipedia-like articles. They cover a broad range of subjects that span various fields, such as technology, professions, travel, economics, and numerous related subtopics. These comprehensive texts try to accurately portray and convey holistic information about real-world objects or experiences.

3.3 Story

BenCoref is primarily composed of short stories, each with a word count of 1000 words or less, which was an arbitrary decision. These stories typically feature 3-4 characters on average. The language used in the stories varies, with some being exclusively in "Shadhubhasha," while others use a mix of "Shadhubhasha" and "Choltibhasha."

3.4 Novel

The Bengali novels in our dataset typically consist of more than 1200 words and feature an average of over 5 characters. These novels primarily employ "Shadhubhasha". The next segment discusses the coreference behaviour across each domain in more detail.

4 Domain Specific Coreference Behaviour Characterization

In this section, some statistics is presented to better understand the coreference phenomenon across each domain. Each coreference cluster may refer to different type of entities, like an object, people, location or event. An arbitrary design choice was made to not explicitly mark the type of entity.

We start by analyzing the mean and standard deviation between mentions across the domains. Table 1 shows that biographies and novels exhibit a low standard deviation but have noticeably different mean distance between mentions. On the other hand, stories and descriptive texts fall in the middle, exhibiting a similar coreference distribution. For mentions that span more than one token, only the first token was used for calculation.

Categories	Mean	Std. Dev
Novel	29.17	3.70
Story	24.10	8.46
Biography	15.67	3.81
Descriptive	22.35	5.42

Table 1: Mean and Std. Deviation of distance between mentions in each domain.

The majority of texts in BenCoref belong to the stories domain, while the biography domain has the smallest contribution. The distribution of mentions, clusters, and tokens across the categories in BenCoref is presented in Figure 4.

Figure 2 depicts the distribution of cluster size across each domain. The cluster size refers to the total number of mentions in each coreference chain. It is worth noting that singletons were not annotated in BenCoref. The story domain has the highest number of coreference chains with two mentions only. Since the story domain contributes the most data to the dataset, this may be a contributing factor to its high frequency in each cluster size. Besides story, the descriptive domain also seems to have more larger coreference chains.

Figure 3 compares the spread of coreference chains in each domain, where the spread refers to the token-level distance between the beginning and end of a coreference chain. A general trend can be observed that as the size of the coreference chain increases, its corresponding frequency decreases in each domain.

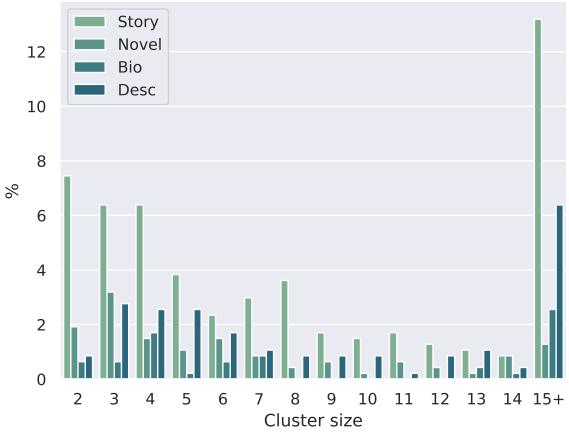

Figure 2: Cluster size comparison between Story, Novel, Biography and wiki-like Descriptive domain.

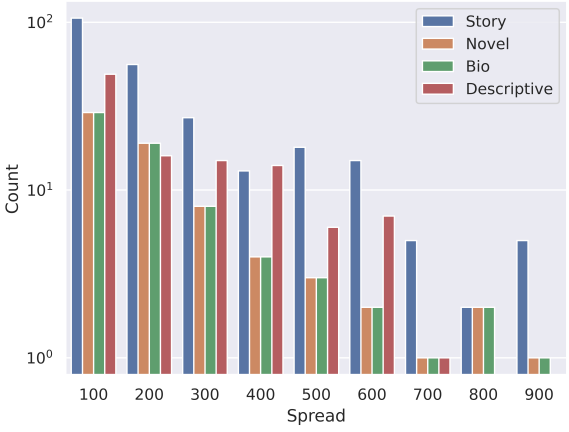

Figure 3: Spread in BenCoref across each domain. The spread is measured by the token level distance between the first and last mention of an entity.

Figure 4: (Right) Distribution of Clusters, Mentions, and Tokens across the categories.

An additional "index.csv" file is included with in the dataset, which serves as an index to all the documents included, organized by title and author. A partial view of this file is presented in Appendix Figure 6.

5 Methodology

We used BnWikiSource[2] and Banglapedia(Islam et al., 2003) as sources of copyright-free Bengali text for our dataset. Banglapedia was used for biographies and wiki-like descriptive texts. The dataset creation process is discussed in detail in the following paragraphs.

5.1 Annotation Phase

The WebAnno annotator (Eckart de Castilho et al., 2016) was the chosen tool for annotation. To accommodate WebAnno's limited capacity to work with large texts, the articles were partitioned according to the Table 2. Each partition ends in a complete sentence and any incomplete portion of a sentence were moved to the next fragment. The partition size was chosen arbitrarily to reduce the number of data fragments. In Appendix Figure 5, a screenshot of the WebAnno interface used during this phase is displayed. A post annotation sample is provided in Figure 7 in Appendix from the Biography domain.

Since there is no existing guideline for coreference annotation, the annotators were initially instructed to annotate the noun phrases and its coreferences, which were predominantly pronouns. The primary noun phrase references were tagged as "entity" and their corresponding coreferences were

[2]https://bn.wikipedia.org

Tokens	Partitions
<699	1
<1000	2
>1000	3

Table 2: Documents with greater than 699 tokens and less than 1000 tokens were divided into 2 parts and the ones with more than 1000 tokens were divided into 3.

tagged as "ref". While determining what forms an entity is an important linguistic problem, it is not the primary challenge we are trying to address in our work. Annotators were free to mark any token or span that they considered an entity. After the annotation phase was completed, the data was exported and the character-level annotations were converted to token-level annotations. For every exceptional cases encountered, a new rule was established and enforced during further annotation of the dataset. The rules are further discussed in the next section.

5.2 Annotation Strategy/Guideline

This coreference annotation guide (refer to A in the Appendix) was prepared concurrently with the annotation phase to ensure consistency throughout the annotation process. We mirrored the overall structure of the OntoNotes annotation guidelines, tailoring them to our specific use case.

Initially, we did not impose any specific restrictions on the definition of an entity during the annotation process. The annotators were instructed to annotate any span they deemed as an entity. However, this approach resulted in an annotator bias, with a strong focus on nominal and pronominal mentions. Subsequently, we made the decision to prioritize and concentrate solely on these types of mentions.

Furthermore, as part of our design decision, we chose to not tag singleton mentions. Consequently, any singletons were removed during the post-annotation processing phase.

5.3 Annotation Criteria:

The general rule used for annotation is to annotate mentions in any form, including nested mentions or those referring to multiple entities. The characterization of mention and coreference link types was conducted after annotating the entire dataset. Annotating coreference link types was kept optional due to the significant training required for the task. This

strategy was followed the accelerate the annotation process.

The rules with corresponding examples are illustrated in a more detailed manner in Figure 10 in the Appendix. Furthermore, the coreference link types have been categorized into two groups, namely identical and apposite, and they have been discussed in detail in 11 and 12 in the Appendix. However, the task of annotating coreference link types is currently pending and will be addressed in future work.

While this guideline is incomplete and limited in scope, it can play an impotant role in encouraging the next generation of coreference datasets in Bengali. The OntoNotes coreference guideline(gui, 2007) is currently in its 7th edition which is a strong indication that the first attempt on making a such guideline would be imperfect and will require further revisions. It may take several iterations before we can have a robust guideline for coreference annotation in Bengali.

5.4 Inter-Annotator Agreement

The OntoNotes strategy was roughly employed to assess interannotator agreement in this work. Specifically, two annotators independently annotated the documents, and only in cases of disagreement, a third annotator was consulted to arrive at a final decision. These ultimate annotations were deemed as the gold standard annotations.

Based on the adjudicated version as the ground truth, the individual annotations in our dataset achieved an average MUC score of 78.3 on the combined dataset. while the combined inter-annotator MUC score was 67.6.

However, it is important to acknowledge that the process of resolving disagreements was not adequately documented and will be addressed in greater detail in future endeavors.

6 Experiments

We took an end-to-end neural network based modeling approach. The following section discusses the algorithm, followed by the experimental setup, evaluation strategy and analysis of results.

We used the 300-dimensional Fasttext and Glove embeddings (Grave et al., 2018) as words representations. To generate contexual representations the embeddings were passed through a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) for some experiments and a variation of the popular transformer-based (Vaswani et al., 2017) pretrained model, BERT(Devlin et al., 2019), for other experiments. For the task of coreference resolution, the contextual representations from these base models were passed on to a span ranking model-head, originally proposed in (Lee et al., 2018). For the crosslingual experiment, a multilingual BERT was finetuned on the OntoNotes dataset.

For hyperparameter optimization, we tuned the maximum number of words in a span(s), maximum number of antecedents per span(a), and coref layer depth(CL).

6.1 Experimental Setup

The data was separated into train and dev set on a ratio of 95% by 5%. An additional test set was carefully prepared, completely disjoint from the train and dev set, that contains 37 documents. An overview of the dataset given in Table 3

	categories	documents	mentions	clusters
train + dev	biography	17	421	38
	descriptive	36	1157	108
	novel	13	601	78
	story	56	3021	278
test	biography	10	303	22
	descriptive	9	290	33
	novel	3	191	15
	story	15	697	53

Table 3: Dataset distribution

For evaluating our system, we used the CONLL-2012 official evaluation scripts which calculates four metrics: Identification of Mentions, MUC, B3 and CEAF. The following section analyzes the performance of our model.

6.2 Results and Analysis

category	model	parameters	pre.	rec.	f1
biography	c2f+Glove	s=30, a=50, CL=2	93.83	65.34	77.04
	c2f+Fasttext	s=20, a=50, CL=2	96.51	64.02	76.98
	BERT-base	s=30, a=50, CL=2	94.22	86.13	90.00
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	7.14	4.62	5.61
story	c2f+Glove	s=30, a=50, CL=2	73.78	58.96	65.55
	c2f+Fasttext	s=20, a=50, CL=2	74.80	54.08	62.78
	BERT-base	s=30, a=50, CL=2	83.91	65.85	73.79
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	7.40	3.73	4.96
novel	c2f+Glove	s=30, a=50, CL=2	78.00	40.83	53.60
	c2f+Fasttext	s=20, a=50, CL=2	87.50	43.97	58.53
	BERT-base	s=30, a=50, CL=2	85.41	64.39	73.43
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	8.51	4.18	5.61
descriptive	c2f+Glove	s=30, a=50, CL=2	66.39	27.93	39.32
	c2f+Fasttext	s=20, a=50, CL=2	72.16	24.13	36.17
	BERT-base	s=30, a=50, CL=2	82.95	50.34	62.66
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	7.47	5.51	6.34

Table 4: Identification of mentions

			B^3			MUC			$CEAF_{\phi4}$			Avg		
category		parameters	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	f1	Pre.	Rec.	F1
biography	c2f + Glove	s=30, a=50, CL=2	84.52	44.74	58.51	92.26	64.41	76.05	55.33	40.24	46.60	77.37	49.80	60.39
	c2f + Fasttext	s=20, a=50, CL=2	89.09	43.99	58.90	95.74	64.05	76.75	61.24	36.19	45.49	82.02	48.08	60.38
	BERT-base	s=30, a=50, CL=2	85.37	73.59	79.04	93.79	86.12	89.79	57.48	49.64	53.27	78.88	69.78	74.03
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	4.28	0.25	0.46	0.67	0.35	0.46	1.22	2.93	1.72	2.06	1.18	0.88
story	c2f + Glove	s=30, a=50, CL=2	46.92	23.95	31.72	63.41	44.40	52.23	20.24	36.99	26.16	43.52	35.11	36.70
	c2f + Fasttext	s=20, a=50, CL=2	47.31	22.80	30.77	65.23	42.54	51.50	23.49	34.02	27.79	45.34	33.12	36.69
	BERT-base	s=30, a=50, CL=2	54.64	31.62	40.06	74.46	53.88	62.52	28.80	40.23	33.57	52.63	41.91	45.38
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	2.32	0.25	0.45	1.42	0.62	0.86	2.00	2.59	2.26	1.91	1.15	1.19
novel	c2f + Glove	s=30, a=50, CL=2	49.87	7.98	13.77	59.45	25.00	35.20	16.37	26.61	20.27	41.90	19.86	23.08
	c2f + Fasttext	s=20, a=50, CL=2	60.30	10.45	17.82	72.72	31.81	44.26	23.36	27.74	25.37	52.13	23.33	29.15
	BERT-base	s=30, a=50, CL=2	43.55	33.93	38.14	71.31	52.27	60.32	34.98	32.80	33.85	49.95	39.67	44.10
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	3.54	0.25	0.47	2.66	1.13	1.59	1.90	2.49	2.15	2.70	1.29	1.40
descriptive	c2f + Glove	s=30, a=50, CL=2	48.33	11.91	19.12	58.24	20.62	30.45	31.16	26.83	28.83	45.91	19.79	26.13
	c2f + Fasttext	s=20, a=50, CL=2	58.32	9.74	16.70	66.66	18.67	29.17	29.98	20.81	24.57	51.65	16.41	23.48
	BERT-base	s=30, a=50, CL=2	62.81	26.88	37.65	76.62	45.91	57.42	46.12	28.18	34.99	61.85	33.66	43.35
	M-BERT(Zero-Shot)	s=30, a=50, CL=2	2.01	0.56	0.88	1.16	0.77	0.93	2.30	3.00	2.60	1.82	1.44	1.47

Table 5: Performance on test data. The main evaluation metric is the average F1 score of MUC, B^3, and $CEAF_{\phi4}$. The best scores are highlighted.

The performance of the model was reasonable given the size of our dataset. As neural networks tend to achieve optimal performance with larger datasets, we hypothesize that our results could be enhanced by expanding our dataset. The model demonstrated good performance in identifying individual mentions, as evidenced by the scores presented in Table 4. However, we observed a decrease in performance during the second phase of clustering the mentions, as shown in Table 5. This highlights the challenge of accurately identifying coreference clusters, particularly in languages with complex sentence structures and a high degree of lexical ambiguity. Further innovation is needed to address these challenges and improve the overall performance of coreference resolution models.

Upon closer inspection one recurring problem was discovered. The model failed to do basic common sense reasoning on long coreference clusters, often breaking it up into several clusters. As demonstrated in Figure 8 in Appendix, the model failed to merge clusters 0 and 1, which should have been a single cluster.

Furthermore, it can be observed that the coreference resolution model performs significantly better on the biography domain as compared to other domains. The relatively low mean and standard deviation of the distance between mentions reported in Table 1 may have contributed to this result. However, despite forming the major portion of the dataset, the story domain did not show any significant improvement. The high standard deviation in distance between mentions reported in Table 1 for the story domain may have contributed to this lack of improvement. Qualitative analysis is needed to investigate the underlying causes of this performance gap.

The zero-shot crosslingual experiment demostrated that coreference knowledge doesn't easily transfer through multilingual training. This clearly demonstrates the need for language specific datasets. Some studies (Novák and Žabokrtskỳ, 2014) report developing projection techniques to improve crosslingual coreference resolution. There maybe scope for further work in this direction.

7 Conclusion

This paper presented BenCoref, the first publicly available dataset of coreference annotations in Bengali. The creation process and annotation guidelines were described in detail to facilitate future work in this area. We then used the dataset to develop an end-to-end coreference resolution system and reported its performance across different domains. Our findings indicate that a lower mean and standard deviation of token-distance between mentions may lead to better results, but further experiments on other datasets are needed to confirm this hypothesis. We also observed a higher tendency for breakage in longer coreference chains.

Our zero-shot cross-lingual experiment demonstrated that coreference knowledge does not easily transfer through multilingual training, highlighting the importance of language-specific datasets. While some studies (Novák and Žabokrtskỳ, 2014) have reported success in developing projection techniques to improve cross-lingual coreference resolution, further research is required to explore this area.

Acknowledgements

We would like to express our sincere gratitude to everyone who contributed to this research. We would also like to thank our colleagues and collaborators for their valuable insights and feedback throughout the project. We are especially grateful to our annotators for their hard work and dedication to ensure the quality of this dataset. We also appreciate the feedback and suggestions from the anonymous reviewers, which helped to improve the quality of this manuscript. No financial assistance was provided by any organization for the completion of this project.

References

2007. *Ontonotes English Coreference Guidelines.*

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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).

Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Abbas Ghaddar and Philippe 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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479.

Sirajul Islam et al. 2003. Banglapedia. *National Encyclopedia, Asiatic Society of Bangladesh, Dhaka*.

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.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Using coreference links to improve spanish-to-english machine translation. Technical report, Idiap.

Thomas S Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*.

Michal Novák and Zdeněk Žabokrtskỳ. 2014. Cross-lingual coreference resolution of pronouns. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 14–24.

Takumi Ohtani, Hidetaka Kamigaito, Masaaki Nagata, and Manabu Okumura. 2019. Context-aware neural machine translation with coreference information. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 45–50.

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.

Apurbalal Senapati and Utpal Garain. 2013. Guitar-based pronominal anaphora resolution in bengali. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 126–130.

Utpal Kumar Sikdar, Asif Ekbal, Sriparna Saha, Olga Uryupina, and Massimo Poesio. 2013. Adapting a state-of-the-art anaphora resolution system for resource-poor language. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 815–821.

Utpal Kumar Sikdar, Asif Ekbal, Sriparna Saha, Olga Uryupina, and Massimo Poesio. 2015. Differential evolution-based feature selection technique for anaphora resolution. *Soft Computing*, 19(8):2149–2161.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

A Appendices

			B^3			MUC			$CEAF_{\phi4}$			Avg		
Category		Parameters	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1
Biography	c2f + Fasttext	s=30, a=250, CL=2	93.43	45.94	61.59	97.88	65.83	78.72	65.62	41.76	51.04	85.64	51.18	63.78
	c2f + Fasttext	s=20, a=50, CL=3	91.63	45.20	60.54	96.84	65.48	78.13	65.87	41.92	51.23	84.78	50.87	63.30
	c2f + Fasttext	s=10, a=50, CL=3	90.98	47.81	62.68	96.92	67.25	79.41	77.44	42.24	54.66	88.54	52.44	65.58
Story	c2f + Fasttext	s=30, a=250, CL=2	48.07	22.95	31.07	65.78	38.81	48.82	21.98	33.73	26.62	45.28	31.83	35.50
	c2f + Fasttext	s=20, a=50, CL=3	53.08	21.70	30.81	67.02	38.50	48.91	22.17	34.02	26.84	47.42	31.41	35.52
	c2f + Fasttext	s=10, a=50, CL=3	53.82	19.35	28.47	66.76	34.62	45.60	21.92	34.77	26.89	47.50	29.58	33.65
Novel	c2f + Fasttext	s=30, a=250, CL=2	47.06	7.32	12.67	61.03	26.70	37.15	16.04	23.06	18.92	41.38	19.03	22.91
	c2f + Fasttext	s=20, a=50, CL=3	65.10	8.54	15.10	71.42	25.56	37.65	24.96	28.08	26.42	53.83	20.73	26.39
	c2f + Fasttext	s=10, a=50, CL=3	57.71	5.82	10.57	63.79	21.02	31.62	17.03	23.42	19.72	46.18	16.75	20.64
Descriptive	c2f + Fasttext	s=30, a=250, CL=2	56.43	9.88	16.82	61.84	18.28	28.22	25.72	20.01	22.51	48.00	16.06	22.52
	c2f + Fasttext	s=20, a=50, CL=3	53.78	7.64	13.37	58.92	12.84	21.08	28.76	19.17	23.01	47.15	13.22	19.15
	c2f + Fasttext	s=10, a=50, CL=3	52.41	7.46	13.06	58.06	14.00	22.57	24.47	19.03	21.41	44.98	13.50	19.01

Table 2: Some additional results on the Model's performance.

Figure 5: A screenshot from WebAnno(Eckart de Castilho et al., 2016) during annotation phase. In this example, the highlighted words are marked as mentions and same color indicate the mentions belong to the same cluster. A colored line joins the highlighted words creating a chain forming a single cluster.

	Title	Author Name	Ids	Type
0	অবরোধ-বাসিনী	বেগম রোকেয়া সাখাওয়াত হোসেন	doc_067 to 090	Story
1	কিশোরদের মন	দক্ষিণারঞ্জন মিত্র মজুমদার	doc_054 to 063	Novel
2	মায়াবাঁশী	রবীন্দ্রনাথ মৈত্র	doc_091 to 096	Story
3	টুনটুনির বই	উপেন্দ্রকিশোর রায়চৌধুরী	doc_097 to 117	Story
4	বীরবলের হালখাতা/আমরা ও তোমরা	প্রমথ চৌধুরী	doc_119	Story
...	...	...	...	...
56	টুসু উৎসব	বাংলাপিডিয়া	doc_049	Descriptive
57	খোজা	বাংলাপিডিয়া	doc_050	Descriptive
58	একাদশী	বাংলাপিডিয়া	doc_051	Descriptive
59	ঈদুল ফিত্‌র	বাংলাপিডিয়া	doc_052	Descriptive
60	গায়ে হলুদ	বাংলাপিডিয়া	doc_053	Descriptive

Figure 6: The supplementary datafile index.csv contains an index to all the datapoints

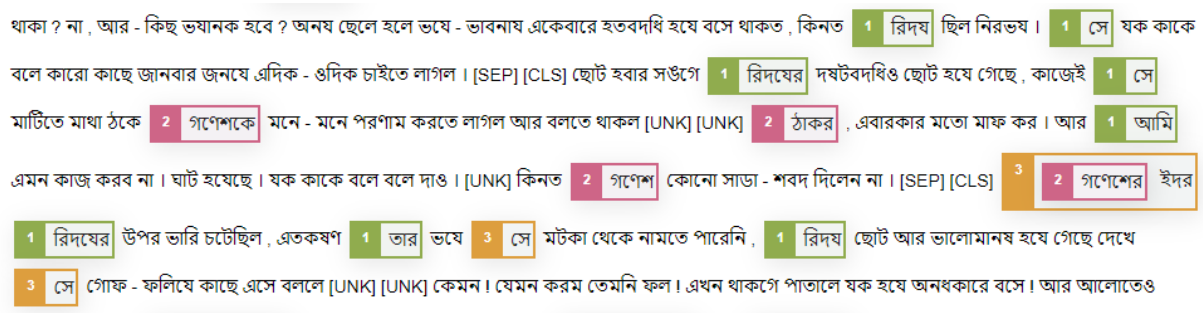

Figure 7: Biography document

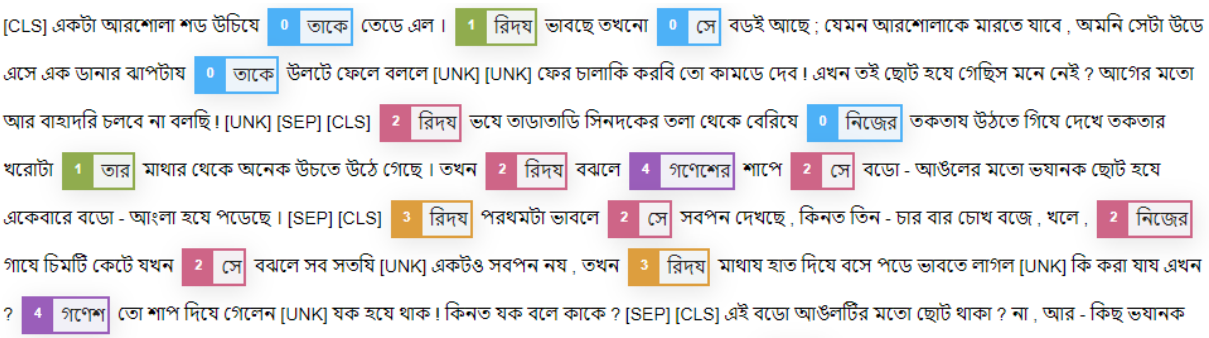

Figure 8: BERT-base model's prediction on a biography document

Category		Parameters	Pre.	Rec.	F1
Biography	c2f+Fasttext	s=30, a=250, CL=2	98.52	66.00	79.05
	c2f+Fasttext	s=20, a=50, CL=3	97.54	65.67	78.50
	c2f+Fasttext	s=10, a=50, CL=3	98.52	66.00	79.05
Story	c2f+Fasttext	s=30, a=250, CL=2	74.20	49.92	59.69
	c2f+Fasttext	s=20, a=50, CL=3	75.16	49.49	59.68
	c2f+Fasttext	s=10, a=50, CL=3	76.52	46.77	58.05
Novel	c2f+Fasttext	s=30, a=250, CL=2	79.00	41.36	54.29
	c2f+Fasttext	s=20, a=50, CL=3	90.12	38.21	53.67
	c2f+Fasttext	s=10, a=50, CL=3	83.75	35.07	49.44
Descriptive	c2f+Fasttext	s=30, a=250, CL=2	74.03	26.55	39.08
	c2f+Fasttext	s=20, a=50, CL=3	70.00	19.31	30.27
	c2f+Fasttext	s=10, a=50, CL=3	68.88	21.37	32.63

Table 7: Additional identification of mention results

Special Considerations in Bengali from Linguistic Perspective

Zero Anaphora: Include implied referents that are not explicitly mentioned in the text. Bangla, like many other South Asian languages, frequently uses zero anaphora, where the subject or object of a sentence is left implicit.
For example, in the sentence "ভাত খাব।" (will eat rice). Here, the subject (who wants to eat) is not explicitly mentioned in the text. This is an example of zero anaphora.

Other common zero anaphora types in Bengali include *Possessive and Reflexive Drop*.
Possessive Drop occurs when the possessive pronoun or possessive marker is omitted, and the possession is understood from the context. Example: "মায়ের কাছে গিয়েছি।" (Went to my mother.) - The possessive pronoun 'আমার' (my) is dropped, but it is understood that the speaker went to their own mother.

Reflexive Drop happens when the reflexive pronoun is dropped, and the reflexive action is understood from the context. Example: "খেলছি।" (I am Playing.) - The reflexive pronoun ‘আমি’ (I am) is dropped, but it is implied that the subject is performing the action on themselves.

Non-Anaphoric Pronouns: Some pronouns in Bengali are inherently non-anaphoric, meaning they rarely behave as anaphoric. These pronouns should not be considered as mentions for coreference.

When considering the sentence "সে নিজের কাজটা করতে পারে।" (He can do his own work.), the reflexive pronoun "নিজের" (own) is non-anaphoric, as it refers to the subject "সে" (he) and emphasizes that he can perform his own work.

Case Marking: Bengali has a rich system of case markers that indicate grammatical relations, which may differ from English. Rules for coreference resolution in English may not directly account for the impact of case markers on coreferential expressions in Bengali.
For example:

মানুষটি ধরেছে বাঘ। (The person has caught the tiger.)
মানুষ ধরেছে বাঘটি। (The tiger has caught the person.)
পুলিশ মারে ডাকাতকে। (The police apprehend the robber.)
পুলিশকে মারে ডাকাত। (The robber apprehend the police.)
মানুষ ধরে বাঘ।
পুলিশ মারে ডাকাত।
However, these are valid sentences in Bengali as well, but they do not clarify the relationship between the entities involved.

Thus, the proper use of case markers in Bengali helps to disambiguate the roles and actions of the different entities involved. However, it is also possible to write valid sentences with ambiguity in Bengali by omitting the use of case markers.

Figure 9: This highlights few key considerations when annotating coreference in Bengali.

BenCoref Co-reference Annotation Guideline

0. General Policy

- Tag mentions according to rules defined in section 1
- Annotating the Coreference Link Types is optional
- Nested Mentions are annotated separately for each mention
    - Example: "রিয়া, যিনি আমার বন্ধুর বোন, আমার বাড়িতে এসেছে।"(Riya, who is my friend's sister, has come to my house.)
      In this sentence, the mention of "আমার" (my), “আমার বন্ধুর” (my friend's) is nested within the mention of "আমার বন্ধুর বোন" (my friend's sister). Tag them all within each mention.
- Multiple Reference: Mentions that refers to multiple chains are annotated multiple times for each chain.
    - Example: “রহিম স্কুলে যায় করিমের সাথে এবং মাঠে খেলে আব্দুলের সাথে। রহিম তার বন্ধুদের সাথেই সময় কাটাতে পছন্দ করে।” (Rahim goes to school with Karim and plays on the field with Abdul. Rahim enjoys spending time with his friends.)
      In this example, the mention "রহিম" (Rahim) refers to two different chains: going to school with Karim and playing on the field with Abdul. To capture these multiple references, the mention "রহিম" would be annotated separately for each chain.

1. Identifying Mentions:

- ***Noun Phrases***: Include all noun phrases that refer to entities, events, or abstract concepts. This includes proper nouns, common nouns, and pronouns.
    - Example: "রহিম বাজারে গেল।" (*Rahim* went to the market.)Here, "রহিম" (Rahim) and "বাজার" (market) are noun phrases that can be considered as mentions.

- ***Pronouns***: Include all pronouns that refer to entities, events, or abstract concepts. This includes personal, demonstrative, and interrogative pronouns.
    - Example: "সে বাজারে গেল।" (He went to the market.) Here, "সে" (he) is a pronoun that can be considered as a mention.

- ***Verbs***: Include verbs that imply the existence of a certain entity or event.
  For example, in the sentence "অর্থনীতি তাৎপর্যপূর্ণভাবে বৃদ্ধি পায়। তবে আমরা যদি এই বৃদ্ধি বজায় রাখতে না পারি শিগগিরই এটি হ্রাস পাবে।" (Economics thrives when it grows consistently. However, if we cannot sustain this growth, it will deteriorate quickly.) Here, "বৃদ্ধি" (thrives) is a verb that implies the existence of an event (playing).
  This “Verb” category of mentions is not included in BenCoref. We expect to include this in future.

2. Coreference Link Types:

- ***Identical***: Link mentions that refer to the exact same entity.
    - Example: "রহিম বাজারে গেল। সে ফল কিনল।" (Rahim went to the market. He bought fruits.) Here, "রহিম" (Rahim) and "সে" (he) refer to the exact same entity.
- ***Appositive***: Link mentions that refer to the same entity, but one mention provides additional information about the entity.
    - Example: "রাম, আমার বন্ধু, বাজারে গেল।" (Ram, my friend, went to the market.) Here, "রাম" (Ram) and "আমার বন্ধু" (my friend) refer to the same entity, but "আমার বন্ধু" (my friend) provides additional information about "রাম" (Ram).

Figure 10: BenCoref Annotation Guideline with examples.

Identical References

- ***Predicate Nominative***: Link a noun phrase and a verb or adjective that refer to the same entity.

  For example, in "সোহেল হচ্ছে কর্মঠ।" (Sohel is hardworking.), the noun phrase "সোহেল" (Sohel) is linked to the adjective "কর্মঠ" (hardworking) through the copular verb "হচ্ছে" (is). Both the noun phrase and the adjective refer to the same entity, Sohel being hardworking.

- ***Relative Pronoun***: Link a relative pronoun to a noun phrase in a relative clause.

  For example, in the sentence "রহিম, যিনি বাজারে গেল, আমার বন্ধু।" (Rahim, who went to the market, is my friend.), "রহিম" (Rahim) and "যিনি" (who) refer to the same entity. The relative pronoun "যিনি" (who) refers back to the noun phrase "রহিম" (Rahim), indicating that Rahim is the person who went to the market. This construction allows you to provide additional information about the noun Rahim within the sentence.

- ***Part-Whole***: This type of link occurs when one mention is a part of another mention.

  For example, in the sentence "রাহিমের হাত ভালো নয়।" (Rahim's hand is not good.), "রাহিমের হাত" (Rahim's hand) and "রহিম" (Rahim) would be linked as part-whole.

- ***Event*:** This type of link occurs when two or more mentions refer to the same event or occurrence.

  For example, in the sentence "বিয়েটা সুন্দর হয়েছিল। সেই অনুষ্ঠান আমার মনে আছে।" (The wedding was beautiful. I remember that event.), "বিয়েটা" (The wedding) and "সেই অনুষ্ঠান" (that event) refer to the same event, so they would be linked as event coreference.

- ***Set-Member***: This type of link occurs when one mention is a member of a set referred to by another mention.

  For example: "ছাত্ররা খেলাঘরে গেল। রামও ছিল।" (The students went to the playground. Ram was also there.), "রাম" (Ram) is a member of the set "ছাত্ররা" (The students), so they would be linked as set-member.

- ***Bridging***: This type of link occurs when one mention infers or implies the existence of another mention.

  For example, in the sentence "রাম একটি বাড়ি কিনল। ছাদটা সবুজ ছিল।" (Ram bought a house. The roof was green.), "ছাদটা" (The roof) infers the existence of "একটি বাড়ি" (a house), so they can be linked as bridging.

- ***Metonymy***: This type of link occurs when one mention is used to refer to another related mention.

  For example, in the sentence "বাংলাদেশ ক্রিকেট দল প্রস্তুত। বাংলাদেশ ক্রিকেট খেলতে চায়।" (Bangladesh cricket team is ready. Bangladesh wants to play cricket.), "বাংলাদেশ" (Bangladesh) is used to refer to “বাংলাদেশ ক্রিকেট দল”(Bangladesh cricket team), so they would be linked as metonymy.

Figure 11: Identical Reference Types

Appositive References

- **Role Appositive**: This type of link occurs when one mention describes the role or occupation of another mention.

  Example: "রহিম, একজন ডাক্তার, আমার বন্ধু।" (Rahim, a doctor, is my friend.). Here, "রহিম" (Rahim) and "একজন ডাক্তার" (a doctor) refer to the same person, so they would be linked as role appositive.

- **Descriptive Appositive**: This type of appositive provides descriptive information about the noun it is referring to.

  Example: "আমার বন্ধু, রহিম, আসছে।" (My friend, Rahim, is coming.) In this example, the appositive "রহিম" (Rahim) provides descriptive information about the noun "আমার বন্ধু" (my friend).

- **Identifying Appositive**: Identifying appositives are used to provide specific identifying information about a noun, such as its name or title.

  Example: "বাংলাদেশের প্রধানমন্ত্রী, শেখ হাসিনা, সংসদে উপস্থিত ছিলেন।" (Sheikh Hasina, the Prime Minister of Bangladesh, was present in the Parliament.) In this sentence, the appositive "শেখ হাসিনা" (Sheikh Hasina) provides the specific identifying information about the noun phrase "বাংলাদেশের প্রধানমন্ত্রী" (the Prime Minister of Bangladesh).

- **Occupational Appositive**: Occupational appositives provide information about the occupation or profession of a person.

  Example: "আমার বাবা, একজন ডাক্তার, প্রায়শই বাইকে কাজ করেন।" (My father, a doctor, mostly travels by bike.) In this sentence, the appositive "একজন ডাক্তার" (a doctor) provides information about the occupation of the noun phrase "আমার বাবা" (my father).

- **Qualifying Appositive**: Qualifying appositives provide additional qualifying information or characteristics about a noun phrase.

  Example: "তার ছেলে, সুন্দর গলার অধিকারী, গান গায়।" (Her son, a beautiful-voiced boy, sings songs.) In this example, the appositive "সুন্দর গলার অধিকারী" (a beautiful-voiced boy) provides qualifying information about the noun phrase "তার ছেলে" (her son).

- **Geographic Appositive**: Geographic appositives provide information about a specific location or place associated with the noun phrase.

  Example: "বগুড়া জেলা, বাংলাদেশের উত্তর-পশ্চিমাঞ্চলের একটি প্রশাসনিক অঞ্চল, পরিচিত হয় শিক্ষা নগরী হিসেবে।" (Bogra District, an administrative region in the north-western region of Bangladesh, is known as a renowned educational city.) In this sentence, the appositive "বাংলাদেশের উত্তর-পশ্চিমাঞ্চলের একটি প্রশাসনিক অঞ্চল" (an administrative region in the north-western region of Bangladesh) provides geographic information about the noun phrase "বগুড়া জেলা" (Bogra District).

Figure 12: Appositive Reference Types