An Annotated Dataset and Automatic Approaches for Discourse Mode Identification in Low-resource Bengali Language

Salim Sazzed
Old Dominion University
Norfolk, VA, USA
ssazz001@odu.edu

Abstract

The modes of discourse aid in comprehending the convention and purpose of various forms of languages used during communication. In this study, we introduce a discourse mode annotated corpus for the low-resource Bengali (also referred to as Bengali) language. The corpus consists of sentence-level annotation of three discourse modes, *narrative*, *descriptive*, and *informative* of the text excerpted from a number of Bengali novels. We analyze the annotated corpus to expose various linguistic aspects of discourse modes, such as class distributions and average sentence lengths. To automatically determine the mode of discourse, we apply CML (classical machine learning) classifiers with n-gram based statistical features and a fine-tuned BERT (Bidirectional Encoder Representations from Transformers) based language model. We observe that fine-tuned BERT-based model yields better results than CML classifiers. Our created discourse mode annotated dataset, the first of its kind in Bengali, and the evaluation, provide baselines for the automatic discourse mode identification in Bengali and can assist various downstream natural language processing tasks.

1 Introduction

Discourse is the notion of conversation that is expressed through language. Based on Webber et al. (2012), discourse indicates the relationship between states, events, or beliefs manifested within one or multiple sentences in a given mode of communication. Understanding discourse structures and identifying relationships between various modes can help downstream natural language processing tasks including text summarization (Li et al., 2016), question answering (Verberne et al., 2007), anaphora resolution (Hirst, 1981), and machine translation (Li et al., 2014).

The modes of discourse, also referred to as rhetorical modes, represent the variety, conventions, and purposes of the dominant types of language used in communication (both oral and written). The discourse modes have high importance while writing composition because they attribute to several factors that would affect the quality and coherence of a text. The combination and interaction of various discourse modes make a text organized and unified (Smith, 2003). To give an example, the writer may start an expressing an event through narration, then provide details regarding using descriptive modes and establish ideas with argument. Discourse modes have also importance in rhetorical research as they are closely related to rhetoric (Connors, 1981) that provides guidelines for effectively expressing content.

Researchers categorized modes of discourse into various categories (Rozakis, 2003; Song et al., 2017; Dhanwal et al., 2020). Based on Rozakis (2003), discourse modes can be classified into four categories, *narration*, *description*, *exposition*, and *argument*. Narration mode primarily focuses on governing the progression of the story by presenting and connecting events; *exposition* mode instructs or explains; the *argument* aims to provide a convincing or persuasive statement; *description* tries to provide detailed mentions of characters, objects, and scenery, in a figurative language. Song et al. (2017) categorized the mode of discourse into five categories, *narration*, *exposition*, *description*, *argument* and *emotion* expressing sentences in narrative essays, while Dhanwal et al. (2020) annotated discourse mode of short story into *argumentative*, *narrative*, *descriptive*, *dialogic*

and *informative* categories. Although a piece of text can be labeled as a specific mode of discourse, it is not uncommon to have text snippets with multiple modes of discourse Song et al. (2017) where one of them possesses the dominant role.

Although discourse structure and mode have a significant role in various downstream natural language processing tasks, research in this area is largely unexplored in Bengla. Although Bengali is the 7th most spoken language in the world [1], NLP resources are scarcely available except few areas such as sentiment analysis (Sazzed and Jayarathna, 2019; Sazzed, 2020) or inappropriate textual content detection (Sazzed, 2021a,b,c). Regarding discourse analysis, only a limited number of works performed research (Chatterjee and Chakraborty, 2019; Banerjee, 2010; Sarkar and Chatterjee, 2013; Das and Stede, 2018; Das et al., 2020). However, to the best of our knowledge, no study related to automatic discourse mode identification has been carried out yet. Thus, in this study, we introduce an annotated dataset and present a set of techniques for the automatic identification of discourse modes.

Following the rough guidelines provided by Smith (2003) and Dhanwal et al. (2020) for discourse mode annotation, we manually categorize a dataset of 3310 sentences from Bengali Novels into various discourse modes. The sentences are annotated in three modes of discourse, *narrative*, *descriptive* and *informative*. For automatic identification of the discourse mode, we extract word n-gram based features from the text and then employ several classical machine learning (CML) classifiers such as logistic regression (LR), support vector machine (SVM), random forest (RF). In addition, the transformer-based multilingual BERT language model is leveraged and fine-tuned for discourse mode determination. We observe that the multilingual BERT model yields better performance than the CML classifiers, although the difference is not substantial compared to LR or SVM.

[1] https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world

1.1 Contributions

The main contributions of this study can be summarized as follows-

- We create a Bengali discourse mode corpus by collecting and annotating texts from a number of Bengali novels. Currently, no discourse mode annotated dataset is available in Bengali; therefore, a key contribution of this study is the development of such a resource that is publicly available for researchers [2].
- We analyze the annotated corpus to reveal attributes of text representing various discourse modes.
- We employ CML classifiers with n-gram based statistical features and a fine-tuned pre-trained language model for automatically identifying various modes of discourse.

2 Data Annotation and Collection

The data collection process starts with identifying a set of novels from Bengali literature. We select six 20th-century Bengali novels **গোলমেলে লোক, পথের পাঁচালি, আরণ্যক,পটাশগড়ের জঙ্গ,নন্দিত নরকে, হিমু)** written by three famous Bengali novelists, *'Shirshendu Mukhopadhyay'*, *'Bibhutibhushan Bandyopadhyay'*, and *'Humayun Ahmed'*. Unlike English, the electronic versions (i.e., eBooks) of Bengali books are hardly available as eBooks are not popular among Bengali readers. Moreover, we notice that most of the eBooks available in PDF format were created by scanning images of the print versions; therefore, they are not suitable for text extraction. We find a website that provides a set of Bengali fiction in EPUB format. From there, we manually download the abovestated six Bengali novels and extract the text for annotation.

Three native Bengali speakers with university-level education perform the annotation. Annotating the mode of discourse in a piece of text (i.e., sentence) is often challenging since a sentence may have multiple modes, or the distinction is often not obvious. Thus annotators are provided a set of online

[2] https://github.com/sazzadcsedu/DiscourseBangla.git

resources and guidelines from a number of publications.

The discourse modes are selected based on the existing works of Song et al. (2017) and Dhanwal et al. (2020). Song et al. (2017) categorized modes of discourse into five categories, *narration*, *exposition*, *description*, *argument* and *emotion* in narrative essays, while Dhanwal et al. (2020) annotated discourse modes into *argumentative*, *narrative*, *descriptive*, *dialogic* and *informative* categories. As our annotated content (i.e., excerpted sentences of Bengali novels) are more similar to the content (i.e., short stories) of Dhanwal et al. (2020), our annotated discourse modes are more similar to their annotation. However, we notice that the presence of the argumentative mode in a fictional novel is rare as instead of establishing any opinion, a novel tells a story in chronological order. Besides, it is observed that the dialogic category itself does not comprise any new mode. Instead, it echoes the narrative or descriptive or other modes from a third-person point of view; thus, we do not include it as a separate mode.

2.1 Discourse Modes

In this study, the following three discourse modes are considered for annotation.

Narrative: Narrative sentences relate to entities performing particular actions, often in chronological order as a part of storytelling.

Bengali: সর্বজয়া ছেলের কাণ্ড দেখিয়া অবাক হইয়া রহিল

English Translation: "Sarvajaya was surprised to see the boy's actions"

Descriptive: Descriptive statements illustrate specific entities with some kind of description so that reader can imagine this in his mind. It enables readers to visualize characters, settings, and actions. For example, it tells how entities look, sound, feel, taste, and smell.

Bengali: একমাথা ঝাঁকড়া ঝাঁকড়া চুল, ভারি শান্ত, সুন্দর চোখমুখ, কুচকুচে কালো গায়ের রং।

English Translation: "She has curly hair, heavy, calm, beautiful eyes, and a sleek black complexion"

Informative: Informative sentences provide information regarding entities or circumstances.

Bengali: এটা পটাশগড়ের এক রাজা বানিয়েছিল।

English Translation: It was made by a king of Potashgarh.

2.2 Annotation Task

The annotation guidelines consist of the formal and informal descriptions of three different types of discourse modes, examples of various modes with the explanation, and examples of co-occurrence of various modes with mode dominance. Although the annotation is performed at the sentence level, the annotators are instructed to consider the surrounding sentences to get a better idea about the context of the sentence for better annotation. In case of the presence of multiple modes in a sentence, the annotators are asked to determine the most dominant discourse mode based on the provided guidelines and their own judgment and label accordingly.

2.3 Annotation and Dataset Statistics

The final dataset consists of 3310 sentences annotated by the three annotators, where two annotators label all the sentences and the third annotator acts only if there is any disagreement between the first two annotators for any case. Note that to include varied types of events and description sentences are randomly selected from the various sections of the novels by annotators (around 50% by each of the annotators). We observe an annotator agreement of 0.78 based on a Cohen's kappa (Cohen, 1960) for the label assignment between the first two annotators.

Table 1: Statistics of various discourse modes in the annotated corpus

Classifier	#Sentence	#Words/ Sentence
Narrative	2282	14.62
Descriptive	782	23.43
Informative	246	11.73

Table 1 depicts the distributions of various modes of discourse in the annotated dataset. As shown in Table 1, the annotated dataset is class imbalanced. We notice that the most dominant mode in the novels is *narrative* since the progression of a novel involves a lot of narrative events. Overall, almost 70% of the sentences in the annotated corpus represent nar-

rative mode. The descriptive mode has 782 instances, while the informative mode is less prevalent and has only 246 samples.

We observe that the most frequently co-occurring modes are *narrative* and *descriptive*, as often chronological events are described with some details. We find that over 20% of narrative sentences convey description to some extent. This observation is consistent with the findings of Song et al. (2017). In the presence of multiple discourse modes within the same sentence, it is often challenging to identify the dominant one.

As seen by Table 1, the average sentence lengths of different discourse modes vary to some extent. For example, the lengths of the sentences representing the descriptive mode are much higher than the other two modes. A higher length of descriptive sentences is expected since they elucidate particular entities or events with some details.

3 Machine Learning Based Approaches

3.1 Classical ML Classifier

We employ four classical supervised ML classifiers: logistic regression (LR), support vector machine (SVM), random forest (RF), and extra trees (ET) for determining the modes of the discourse of sentences. For SVM, we apply all three types of kernels, linear, polynomial, and Gaussian radial basis function (RBF). We find the linear kernel performs best for our classification problem (reported results).

The word n-gram features are utilized as input for the CML classifiers. An n-gram is a contiguous sequence of n items from a piece of text. We extract the word-level unigrams and bigrams from the text, compute corresponding tf-idf scores, and then feed those values to the CML classifiers.

For the CML classifiers, the default parameter settings of the scikit-learn (Pedregosa et al., 2011) library are used. A class-balanced weight is set for all CML classifiers.

3.2 Deep Learning Based Classifier

The transformer-based pre-trained contextual embedding such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have achieved state-of-the-art results in various text classification tasks with limited labeled data. As these language models have been trained with a large amount of unlabelled data, they possess contextual knowledge; thus, fine-tuning them utilizing a small amount of problem-specific labeled data can attain satisfactory results.

BERT utilizes the transformer architecture to learn contextual relationships between words (or sub-words) in a piece of text. Before feeding text sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The BERT model then tries to infer the original value of the masked words utilizing the contextual meaning provided by the surrounding non-masked words present in the sequence.

The multilingual BERT (M-BERT) (Devlin et al., 2019) is the multilingual version of BERT, which was pre-trained with the Wikipedia content of 104 languages (Bengali is one of them). It consists of twelve-layer transformer blocks where each block contains twelve head self-attention layers and 768 hidden layers that result in approximately 110 million parameters.

3.2.1 Fine Tuning

We fine-tune M-BERT for categorizing sentences into the three classes, *narrative, descriptive, informative*. Since this is a classification task, we utilize the classification module of the M-BERT. The hugging face library (Wolf et al., 2019) is used to fine-tune M-BERT.

Since the initial layers of M-BERT only learn very general features, we keep them untouched. Only the last layer of the M-BERT is fine-tuned for our binary-level classification task. We only add one layer on top of the M-BERT for classification that acts as a classifier. We tokenize and feed our input training data to fine-tune M-BERT model; Afterward, the fine-tuned model is used for classifying the testing data.

A mini-batch size of 8 and a learning rate of $4*10^{-5}$ are used. The validation and training split ratio is set to 80% and 20%. The model is optimized using the Adam optimizer (Kingma and Ba, 2014), and the loss parameter is set to sparse-categorical-cross-entropy. The model is trained for 3 epochs with early

Table 2: Performance of various approaches for discourse mode prediction

Type	Classifier	Narrative	Descriptive	Informative
		F1/Acc.	**F1/Acc.**	**F1/Acc.**
CML	LR	0.8857/0.9708	0.6796 /0.5896	0.064/0.0333
	SVM	0.8739/0.9787	0.6126/0.4909	0.0328/0.0167
	RF	0.8433/0.9911	0.3773/0.2416	0.0165/0.0083
	ET	0.8458/0.9938	0.4/0.2571	0.0328/0.0167
DL	Multilingual BERT	0.912/0.957	0.66/0.6875	0.0468/0.024

Table 3: An example of the confusion matrix yielded by the LR classifier

Class	Narrative	Descriptive	Informative
Narrative	2213	69	0
Descriptive	337	438	7
Informative	184	50	12

stopping enabled.

3.3 Evaluation Settings

To evaluate the performances of various approaches, 5-fold cross-validation is applied. The 5-fold cross-validation split the dataset into 5-mutually independent subsets. It consists of 5 iterations; in each iteration, one of the new subsets is used as a testing set, and the other two subsets are used as a training set.

The F1 score and accuracy of all three classes are reported separately. The F1 score of each class is computed based on its precision and recall scores. Let c represents a particular class and c' refer to all other classes. The TP, FP, and FN for the class c are defined as follows-
TP = both true label and prediction refer a sentence to class c
FP = true label of a sentence is class c', while prediction says it is class c
FN = true label marks a sentence as class c, while prediction refers to it class c'

4 Results and Discussion

Table 2 provides the F1 scores and accuracy of various CML-based classifiers and transformers-based M-BERT model for discourse mode identification.

The results reveal that all the four CML classifiers, LR, SVM, RF, and ET, yield high performance for the narrative class prediction; they achieve F1 scores between 0.84-0.89 and an accuracy of around 97%. For the descriptive class prediction, LR and SVM perform better than the RF and ET; they obtain f1 scores over 0.60 compared to 0.4 scores of decision tree-based classifiers. However, we observe that for informative class prediction all the classifiers perform poorly.

We observe that the performances of CML classifiers are affected by the class distribution of the dataset. Since the narrative class contains close to 70% of the instances in the dataset, the classifiers are biased towards it (Table 3). All the CML classifiers fail to provide an acceptable level of performance for the minor *informative* class even after using class-balanced weights. We also employ SMOTE (Chawla et al., 2002) oversampling techniques for class balancing; however, we do not notice any noticeable performance improvement using SMOTE.

The transformer-based multilingual language model yield slightly better performance than the CML classifiers. For the dominant *narrative* class, it attains an f1 score of 0.912. For other classes, it obtains similar f1 scores of the LR and SVM, around 0.67 and 0.05, respectively. It is noticed that all the classifiers perform poorly for the minor *informative* class prediction.

The results suggest that the transformer-based multilingual BERT model can be effective for discourse mode classification in Bengali text. Although we do not notice signif-

icant improvement compared to CML classifiers in this study, it could be attributed to limited labeled data. With more labeled data incorporated, the improvement could be higher (transformer-based models have shown state-of-the-art performances for various NLP tasks across languages). Low resource language such as Bengali suffers from data annotation issues, as there are not enough resources to create a large labeled dataset. Thus, incorporating a pre-trained model can help address the scarcity of annotated data in the Bengali language to some extent.

5 Summary and Future Work

In this study, we introduce a corpus consisting of sentences level annotation of various modes of discourse. The corpus consists of excerpted text from Bengali novels annotated with three different discourse modes: *narrative*, *descriptive* and *informative*. We provide details of the annotation procedure, such as annotation guidelines and annotator agreements, and investigate the characteristics of various discourse modes. Finally, we employ CML and deep learning-based classification approaches for automatic discourse mode identification. We observe that transformer-based fine-tuned language models yield the best performance. Our future work will expand the size of the corpus and demonstrate the usefulness of discourse mode annotated data for downstream tasks such as automated essay scoring and sentiment analysis in the low-resource Bengali language.

References

Sanjoy Banerjee. 2010. Context in communication: analysis of bengali spoken discourse.

Rajoshree Chatterjee and Jayshree Chakraborty. 2019. Analyzing discourse coherence in bengali elementary choras (children's nursery rhymes). *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 11(3).

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Robert J Connors. 1981. The rise and fall of the modes of discourse. *College Composition and Communication*, 32(4):444–455.

Debopam Das and Manfred Stede. 2018. Developing the bangla rst discourse treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Debopam Das, Manfred Stede, Soumya Sankar Ghosh, and Lahari Chatterjee. 2020. Dimlex-bangla: A lexicon of bangla discourse connectives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1097–1102.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, Debanjan Mahata, Rakesh Gosangi, Haimin Zhang, Rajiv Shah, et al. 2020. An annotated dataset of discourse modes in hindi stories. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1191–1196.

Graerne Hirst. 1981. Discourse-oriented anaphora resolution in natural language understanding: A review. *American journal of computational linguistics*, 7(2):85–98.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in

near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Laurie Rozakis. 2003. *The complete idiot's guide to grammar and style*. Penguin.

Abhishek Sarkar and Pinaki Sankar Chatterjee. 2013. Identification of rhetorical structure relation from discourse marker in bengali language understanding.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource bengali language. In *Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020)*, pages 50–60.

Salim Sazzed. 2021a. Abusive content detection in transliterated bengali-english social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130.

Salim Sazzed. 2021b. Identifying vulgarity in bengali social media textual content. *PeerJ Computer Science*, 7:e665.

Salim Sazzed. 2021c. A lexicon for profane and obscene text identification in bengali. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1289–1296.

Salim Sazzed and Sampath Jayarathna. 2019. A sentiment classification in bengali and machine translated english corpus. In *2019 IEEE 20th international conference on information reuse and integration for data science (IRI)*, pages 107–114. IEEE.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Wei Song, Dong Wang, Ruiji Fu, Lizhen Liu, Ting Liu, and Guoping Hu. 2017. Discourse mode identification in essays. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 112–122.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.