Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1191–1196
Marseille, 11–16 May 2020

An Annotated Dataset of Discourse Modes in Hindi Stories

Swapnil Dhanwal[1], Hritwik Dutta[1], Hitesh Nankani[1], Nilay Shrivastava[1], Yaman Kumar[4], Junyi Jessy Li[2], Debanjan Mahata[3], Rakesh Gosangi[3], Haimin Zhang[3], Rajiv Ratn Shah[1], Amanda Stent[3]
MIDAS Lab, IIIT-Delhi[1], University of Texas at Austin, USA [2],
Bloomberg, USA[3], Adobe Systems, Noida, India[4]
ykumar@adobe.com, jessy@austin.utexas.edu, swapnild.co@nsit.net.in, rajivratn@iiitd.ac.in
{duttahritwik,hiteshnankani1987,nilayshrivastava1729}@gmail.com
{hzhang449,dmahata,rgosangi,astent}@bloomberg.net

Abstract

In this paper, we present a new corpus consisting of sentences from Hindi short stories annotated for five different discourse modes *argumentative*, *narrative*, *descriptive*, *dialogic* and *informative*. We present a detailed account of the entire data collection and annotation processes. The annotations have a very high inter-annotator agreement (0.87 k-alpha). We analyze the data in terms of label distributions, part of speech tags, and sentence lengths. We characterize the performance of various classification algorithms on this dataset and perform ablation studies to understand the nature of the linguistic models suitable for capturing the nuances of the embedded discourse structures in the presented corpus.

Keywords: Discourse modes, Hindi, Low resource language

1. Introduction and Background

Discourse in the context of linguistics is defined as exploitation of language features by speakers to express what they are talking about, indicating relationship to what they have already talked about, change of topic and relationship between states, events, beliefs etc, in a given mode of communication (Webber et al., 2012). The discourse structures could be present in the form of a single sentence or span across multiple sentences. Understanding discourse structures and relationships between them in text could be useful for many natural language processing tasks such as summarization (Li et al., 2016), question answering (Verberne et al., 2007), natural language generation (Williams and Reiter, 2003), anaphora resolution (Hirst, 1981), textual entailment (Hickl and Bensley, 2007), and machine translation (Li et al., 2014).

With the release of Penn Discourse Treebank (Prasad et al., 2008), there has been an increasing interest from the scientific community to study discourse relations holding between eventualities in text in different languages such as Arabic (Al-Saif and Markert, 2010), Chinese (Zhou and Xue, 2012), Czech (Mladová et al., 2008), Italian (Tonelli et al., 2010), Tamil (Rachakonda and Sharma, 2011), Turkish (Zeyrek et al., 2010), and Hindi (Oza et al., 2009). Very few studies exists that aims to identify different discourse modes (Smith, 2003) from written text.

Hindi language is one of the 22 official languages of India and is among the top five most widely spoken languages in the world[1]. In spite of its wide usage it is still considered as one of the low resource languages by NLP practitioners, which necessitates the creation of new resources and tools for computational linguists that enables them to understand the underlying nuances of the language using natural language processing techniques. The syntax and semantics of Hindi is often different from other high resource languages like English. Dependency of the meaning of expressions on word order, morphological variations, and spelling variations makes Hindi an interesting language to study and also pose additional challenges for linguistic modelling (Kumar et al., 2019).

Tripathi et al. (Tripathi et al., 2016), created a Hindi corpus of 1960 sentences extracted from children stories that are annotated with discourse modes for improving storytelling experience using TTS systems. They annotated an already existing story speech corpus (Sarkar et al., 2014) for three discourse modes - *dialogue*, *narrative* and *descriptive*. The main motivation of their work was to develop an automated discourse mode identification system at sentence level that could be further used for enhancing the output of a TTS system by improving the performance of the prosody models.

Motivated by the efforts of (Tripathi et al., 2016) we present a new dataset consisting of high quality annotations of five discourse modes - *argumentative*, *narrative*, *descriptive*, *dialogic* and *informative* for sentences extracted from Hindi short stories also collected by us. We further show the presence of *argumentative* and *informative* modes that were not annotated previously by the referred prior work and do not limit the corpus to the genre of children stories. We provide a detailed analysis of the presented corpus and also train baseline models for automatic identification of discourse modes at the level of sentences. As previously mentioned discourse structures finds its uses in various NLP tasks. We believe that our dataset has potential uses in many of them, and particularly in improving TTS systems for storytelling.

Some of the main contributions that we make in this work are:

[1] `https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers`

- Present a new publicly available corpus[2] consisting of sentences from short stories written in a low-resource language of Hindi having high quality annotation for five different discourse modes - *argumentative*, *narrative*, *descriptive*, *dialogic* and *informative*.

- Perform a detailed analysis of the proposed annotated corpus and characterize the performance of different classification algorithms.

The remaining paper is organized as follows. Section 2. explains the annotation process and presents the detailed analysis of the corpus. Section 3. shows the performance of different baseline models on our annotated corpus for the task of identifying discourse modes from Hindi sentences. Finally, Section 4. concludes the key findings and future scope of this work.

2. Annotations and Dataset

2.1. Annotations

Our first step in the annotation process was to identify a list of short stories in Hindi. To this end, we first identified 11 famous authors from the twentieth century Indian literature. We then selected four to five stories from each author which were available in the public domain resulting in a collection of 53 stories[3]. Most of these short stories were originally written in Hindi but some of them were written in other Indian languages and later translated to Hindi.

We chose against crowd-sourcing the annotation process because we wanted to directly work with the annotators for qualitative feedback and to also ensure high quality annotations. We employed three native Hindi speakers with college level education for the annotation task. We first selected two random stories from our corpus and had the three annotators work on them independently and classify each sentence based on the discourse mode taxonomy used in (Song et al., 2017). Song et al (Song et al., 2017) developed their taxonomy based on prior works in linguistics (Smith, 2003). This preliminary task helped the annotators familiarize themselves with discourse modes and also understand the scope of this annotation task. More importantly, this also helped us ascertain feedback about the class labels.

Based on the annotators' feedback we first observed that of five discourse modes used in (Song et al., 2017), *Emotion* was extremely prevalent: most of the sentences in these short stories could be associated with some sort of an emotion. We therefore decided to eliminate *Emotion* category from our taxonomy. Second, the annotators' feedback also helped us realize that we needed to introduce a new discourse mode *Dialogic* to characterize sentences that denote conversations in the stories.

2.2. Discourse modes

Following is a short description of the five discourse modes annotated in this paper. Also included are examples for each mode in Hindi and their corresponding English translations.

Narrative: Narrative sentences typically include description of entities performing particular actions and how they are anchored to the timeline of the story.

> शेर को अपने पास आते देख राज झील की ओर भाग लगा।
> *Raj started running towards the lake after seeing the lion approach him.*

Argumentative: Argumentative statements make a claim or a comment on a subject and often support their validity to convince the counter-parties. These statements might include personal opinions of the characters, exclamations, and rhetorical questions.

> मेंरे हिसाब से वे क्रिकेट खेलना नहीं जानते|
> *According to me, they do not know how to play cricket.*

Descriptive: Descriptive statements aim to portray specific locations in a story with the help of language thus creating a mental picture in the reader's mind. The main difference between descriptive and narrative statements is that the latter relates events or entities temporally.

> वृक्षों पर अजीब हरियाली है, खेतों में कुछ अजीब रौनक है, आसमान पर कुछ अजीब लालिमा है।
> *The trees are lush green, the fields are shining, and the sky is glowing red.*

Informative: Informative sentences, as the name suggests, present information about an entity or a situation to help the reader.

> औसत फ्लू के मौसम में लगभग 100 बच्चे मर जाते हैं।
> *On an average, 100 kids die during the flu season.*

Dialogic: As mentioned before, this class was introduced to capture conversational dialogues in the story and it does not include first person thoughts of characters.

> हामिद भीतर जाकर दादी से कहता है—तुम डरना नहीं अम्माँ, मै सबसे पहले आऊँगा।
> Hamid says to his grandmother - don't worry amma, I will go first.

2.3. Annotation task

The annotation guidelines contained a brief introduction to discourse modes, the five discourse modes with explicit definitions, examples, and counter-examples. The short stories were first processed to identify sentence boundaries (Boroș et al., 2018). The annotators were instructed to work on one sentence at a time

[2] https://github.com/midas-research/hindi-discourse

[3] A list of all the annotated short stories and the corresponding authors is provided in supplementary material. We scraped the raw text of all these stories from this website: http://www.hindikahani.hindi-kavita.com/

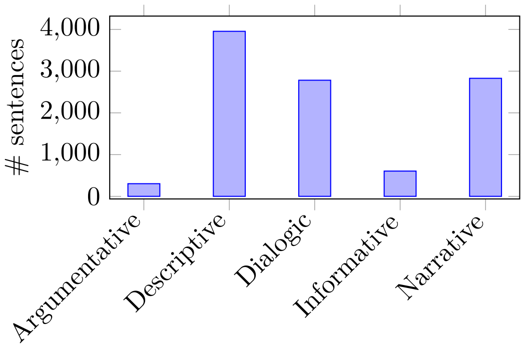

Figure 1: Number of samples per discourse mode

but they did have access to the preceding and succeeding sentences to help ascertain better context. For each, sentence the annotators were asked to identify the most salient discourse mode, and when applicable a secondary discourse mode. Following is an example of a sentence with two discourse modes (*Narrative* and *Argumentative*):

> शायद उसकी समझ में यह बात एकाएक साफ हो गई कि उसकी बेटी भी इतने दिनों में बड़ी हो गई होगी, और उसके साथ भी उसे अब फिर से नई जान-पहचान करनी पड़ेगी।
> *Maybe it was clear to him that his daughter must have grown up now, and that he would have to reacquaint with her again.*

The 53 stories contained 10,472 sentences and all sentences were annotated by the three annotators. We evaluated the annotations in terms of inter-annotator agreements using Krippendorff' s alpha (K-alpha) (Krippendorff, 2011) which is more robust than simple agreement measures because it accounts for chance correction and class distributions. We observed strong inter-annotator agreements (K-alpha of 0.87) and per recommendations in (Artstein and Poesio, 2008) conclude that the annotations are of good quality. We chose a straightforward majority decision for label aggregation: if two or more annotators agreed on a discourse mode for a sentence. In cases where there was no agreement between the annotators, they met in person to discuss and assign the final label.

2.4. Dataset statistics

Figure 1 shows a distribution of the number of sentences for each discourse mode. There is fair amount of class imbalance in this domain with the most prevalent class *Descriptive* having 3,954 samples, and the two low prevalence classes (*Informative* and *Argumentative*) having 605 and 303 samples respectively. Of the 10,472 sentences, 1,504 (14.3%) had two discourse modes. The most frequently co-occurring labels were *Narrative* and *Dialogic*: 719 sentences were labeled with both these discouse modes.

Figure 2 shows a distribution of number of sentences for all the stories: the shortest story has 67 sentences, the longest has 704 sentences, and the average length of the story is 197 sentences. Figure 3 shows a distribution of number of tokens per sentence for each discourse mode. Of the five discourse modes, *Informative* had the longest sentences (23.6 tokens per sentence) and *Argumentative* had the shortest sentences (14.3 tokens per sentence).

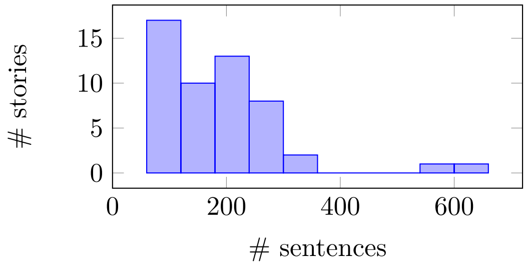

Figure 2: Distribution of # sentences per story

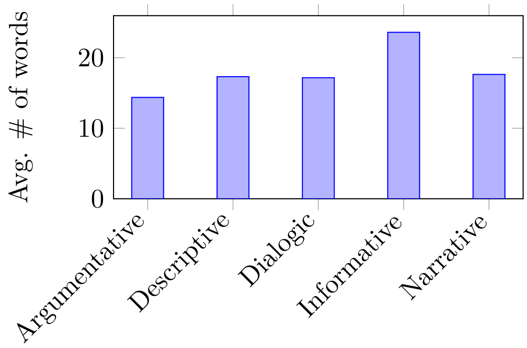

Figure 3: Average number of words per sentence

We also analyzed the data in terms of parts of speech tags (Boroș et al., 2018). The results are summarized in Table 1. Following are a few interesting observations about this data. *Informative* sentences have the most nouns (5.63), proper nouns (1.56), and numerals (1.2) per sentence. This is expected because *Informative* sentences typically provide information about various entities and characters in the story. *Dialogic* sentences have the most punctuations: 2.93 per sentence. *Narrative* sentences have the most verbs: 2.55 per sentence.

3. Experiments

We split the data randomly into three sets: 60% (6283 sentences) for training, 15% (1571 sentences) for validation, and 25% for testing (2618 sentences). We trained six different standard machine learning algorithms: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), and XGBoost (XGB). We experimented with four types of representations: bag of words, token-level n-grams (2 and 3 grams), tf-idf representation of tokens, and character-level n-grams (2 and 3 grams). The results are summarized in terms of accuracy and F1-scores in Table 2. We observed that character-level n-grams obtained the best performance across all the models. The LR model obtained the best F1-score for 3 out of the 5 classes and also the best macro-averaged

	ADJ	ADP	ADV	AUX	CCONJ	DET	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	VERB
Argumentative	0.64	1.73	0.17	1.25	0.36	0.47	2.90	0.21	0.92	1.64	0.39	1.61	0.26	1.78
Descriptive	0.86	2.60	0.26	1.99	0.42	0.49	3.74	0.25	0.66	1.43	0.84	1.63	0.16	1.96
Dialogic	0.65	1.80	0.24	1.54	0.22	0.37	3.09	0.22	0.81	1.65	1.10	2.93	0.20	2.33
Informative	1.39	4.39	0.39	2.53	0.45	0.85	5.63	1.42	0.84	1.44	1.56	0.18	0.35	2.16
Narrative	0.75	2.83	0.31	1.38	0.40	0.40	3.77	0.29	0.53	1.39	1.04	1.84	0.16	2.55

Table 1: POS tag analysis: Each entry in the table is the average number of words tagged as the column label. ADJ: Adjective, ADP: Adposition, ADV: adverb, AUX: auxiliary, CCONJ: Coordinating Conjunction, DET: Determiner, INTJ: Interjection, NOUN: Noun, NUM: Numeral, PART: Particle, PRON: Pronoun, PROPN: Proper Noun, PUNCT: Punctuation, SCONJ: Subordinating Conjunction, SYM: Symbol, VERB: verb

Model	Argumentative	Descriptive	Dialogic	Informative	Narrative	Macro-F1
NB	0.250	0.663	0.615	0.562	0.656	0.549
LR	**0.274**	0.688	**0.618**	**0.727**	0.667	**0.594**
SVM	0.066	0.000	0.149	0.195	0.009	0.083
RF	0.151	0.614	0.484	0.496	0.591	0.467
XGB	0.247	**0.695**	0.585	0.647	**0.687**	0.572

Table 2: Model performances: F1 score for each dialogue mode, and the macro averaged F1.

F1. The XGB models obtained the best F1-score for *Argumentative* and *Informative* classes. All the models obtained the worst performance on the lowest prevalent class: *Argumentative*. Though *Informative* has relatively few samples, some of the models obtained very good performance on this class.

We conducted a second experiment where the training data was over-sampled to ensure uniform distribution for all the five discourse modes resulting in 11,855 samples (2,371 per class). In this experiment, in addition to the five shallow models, we also trained one deep learning model (CNN-BiLSTM) where individual words were represented using pre-trained FastText (Bojanowski et al., 2017) embeddings. The results are summarized in Table 3. The over-sampling process improved the overall performance of all the models but most significantly for the SVM model. The LR model is still the best obtaining best F1-score for 3 out of 5 classes and also the best macro-averaged F1.

We did not see a significant improvements from deep learning models. This is likely because of lack of embeddings for a large portion of the vocabulary in the dataset. We conducted a small ablation study where in we evaluated the CNN-BiLSTM model on the same dataset but the words in the sentences were randomly shuffled. This variation reduced the performance of the model from a macro-f1 of 0.613 to 0.268. We then evaluated with half the words from each sentence in the test dataset. This variation reduced the performance to 0.547. Based on these number, we expect that deep learning models would do much better with embeddings that have greater vocabulary coverage and also those which account for contextual information. Further investigation is necessary to establish this hypothesis.

We further analyzed the predictions made by the LR model in terms of label confusions (see Table 4). We observe most confusion between *Narrative* and *Descriptive* categories. The model misclassified 116 *Descriptive* samples as *Narrative* and 73 samples the other way. We also observe that *Argumentative* category has a very low precision with a lot of samples from *Descriptive* and *Dialogic* categories misclassified as *Argumentative.*

4. Conclusion and Future Work

In this paper, we presented a new publicly available corpus containing sentences from Hindi short stories annotated by humans for five different discourse modes: *Argumentative*, *Descriptive*, *Dialogic*, *Informative*, and *Narrative*. Motivated by prior work of annotating discourse modes for children stories we extend our annotation process to short stories from a variety of genres. We provided a detailed description of the datasets along with performances of machine learning models for the task of identifying discourse modes from Hindi sentences. Our experiments led us to few interesting observations. We could not get the best performance using the deep learning model trained on the data yet we saw the importance of word order, context and sequence by performing few ablation studies. We believe that these findings are contradictory, and hope to investigate further in the future. As a future work we would also like to use the presented corpus to see how it could be further used in certain downstream tasks such as emotion analysis, machine translation, textual entailment, and speech sythesis for improving storytelling experience in Hindi language.

5. Bibliographical References

Al-Saif, A. and Markert, K. (2010). The leeds arabic discourse treebank: Annotating discourse con-

Model	Argumentative	Descriptive	Dialogic	Informative	Narrative	Macro-F1
NB	0.310	0.751	0.710	0.802	0.692	0.653
LR	0.370	**0.778**	**0.716**	0.826	**0.740**	**0.685**
SVM	0.297	0.732	0.579	0.767	0.673	0.609
RF	0.155	0.716	0.612	0.740	0.649	0.574
XGB	**0.377**	0.755	0.667	0.741	0.719	0.651
CNN-BiLSTM	0.180	0.727	0.637	**0.841**	0.679	0.613

Table 3: Model performances for over-sampled data: F1 score for each dialogue mode, and the macro averaged F1.

	Argumentative	Descriptive	Dialogic	Informative	Narrative
Argumentative	39	15	17	1	9
Descriptive	31	745	87	24	116
Dialogic	45	68	486	10	82
Informative	3	11	3	135	1
Narrative	12	73	74	4	527

Table 4: Confusion matrix for LR model on over-sampled data. Row label is the ground truth and column label is predicted by the model.

nectives for arabic. In *LREC*, pages 2046–2053.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boroș, T., Dumitrescu, S. D., and Burtica, R. (2018). NLP-cube: End-to-end raw text processing with neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 171–179, Brussels, Belgium, October. Association for Computational Linguistics.

Hickl, A. and Bensley, J. (2007). A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176. Association for Computational Linguistics.

Hirst, G. (1981). Discourse-oriented anaphora resolution in natural language understanding: A review. *Computational Linguistics*, 7(2):85–98.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Kumar, Y., Mahata, D., Aggarwal, S., Chugh, A., Maheshwari, R., and Shah, R. R. (2019). Bhaav-a text corpus for emotion analysis from hindi stories. *arXiv preprint arXiv:1910.04073.*

Li, J. J., Carpuat, M., and Nenkova, A. (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.

Li, J. J., Thadani, K., and Stent, A. (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.

Mladová, L., Zikánová, Š., and Hajicová, E. (2008). From sentence to discourse. In *Proc. 6th Int' l Conf. on Language Resources and Evaluation.*

Oza, U., Prasad, R., Kolachina, S., Sharma, D. M., and Joshi, A. (2009). The hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *LREC*. Citeseer.

Rachakonda, R. T. and Sharma, D. M. (2011). Creating an annotated tamil corpus as a discourse resource. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 119–123. Association for Computational Linguistics.

Sarkar, P., Haque, A., Dutta, A. K., Reddy, G., Harikrishna, D., Dhara, P., Verma, R., Narendra, N., SB, S. K., Yadav, J., et al. (2014). Designing prosody rule-set for converting neutral tts speech to storytelling style speech for indian languages: Bengali, hindi and telugu. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 473–477. IEEE.

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

Song, W., Wang, D., Fu, R., Liu, L., Liu, T., and Hu, G. (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.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. K. (2010). Annotation of discourse relations for conversational spoken dialogs. In *LREC.*

Tripathi, K., Sarkar, P., and Rao, K. S. (2016). Sentence based discourse classification for hindi story text-to-speech (tts) system. In *Proceedings of the 13th International Conference on Natural Language*

Processing, pages 46–54.

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P. (2007). Evaluating discourse-based answer extraction for why-question answering.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Williams, S. and Reiter, E. (2003). A corpus analysis of discourse relations for natural language generation.

Zeyrek, D., Demirşahin, I., Sevdik-Çalli, A., Balaban, H. Ö., Yalçinkaya, İ., and Turan, Ü. D. (2010). The annotation scheme of the turkish discourse bank and an evaluation of inconsistent annotations. In *Proceedings of the fourth linguistic annotation workshop*, pages 282–289. Association for Computational Linguistics.

Zhou, Y. and Xue, N. (2012). Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 69–77. Association for Computational Linguistics.