

# Segmenting Scientific Abstracts into Discourse Categories: A Deep Learning-Based Approach for Sparse Labeled Data

Soumya Banerjee  
soumyaBanerjee@outlook.in  
National Digital Library  
Indian Institute of Technology  
Khragpur, India

Debarshi Kumar Sanyal  
debarshisanyal@gmail.com  
National Digital Library  
Indian Institute of Technology  
Khragpur, India

Samiran Chattopadhyay  
samirancju@gmail.com  
Department of IT  
Jadavpur University  
Kolkata, India

Plaban Kumar Bhowmick  
plaban@cet.iitkgp.ac.in  
Center for Education Technology  
Indian Institute of Technology  
Khragpur, India

Partha Pratim Das  
ppd@cse.iitkgp.ac.in  
Department of CSE  
Indian Institute of Technology  
Khragpur, India

## ABSTRACT

The abstract of a scientific paper distills the contents of the paper into a short paragraph. In the biomedical literature, it is customary to structure an abstract into discourse categories like BACKGROUND, OBJECTIVE, METHOD, RESULT, and CONCLUSION, but this segmentation is uncommon in other fields like computer science. Explicit categories could be helpful for more granular, that is, discourse-level search and recommendation. The sparsity of labeled data makes it challenging to construct supervised machine learning solutions for automatic discourse-level segmentation of abstracts in non-bio domains. In this paper, we address this problem using transfer learning. In particular, we define three discourse categories – BACKGROUND, TECHNIQUE, OBSERVATION – for an abstract because these three categories are the most common. We train a deep neural network on structured abstracts from PubMed, then fine-tune it on a small hand-labeled corpus of computer science papers. We observe an accuracy of 75% on the test corpus. We perform an ablation study to highlight the roles of the different parts of the model. Our method appears to be a promising solution to the automatic segmentation of abstracts, where the labeled data is sparse.

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → **Document structure**.

## KEYWORDS

structured abstract, deep learning, LSTM, transfer learning

## 1 INTRODUCTION

The abstract of a research paper is a short, succinct description of the content of the paper. In a few lines, it conveys the information that is subsequently revealed in detail over multiple pages supplemented with figures, tables, and references to existing works. In biomedical literature, including research papers, review articles, and clinical practice guidelines, it is a common practice to have structured abstracts [10]. They typically follow the IMRaD format, i.e., INTRODUCTION, METHODS, RESULTS, and DISCUSSION. Most medical journals indexed in PubMed conform to this style [8].

Structured abstracts can help researchers to refer to their regions of interest quickly, label documents more effectively, assist the indexing process, and help in data mining [8]. Recently researchers have designed deep network models to automatically segment unstructured abstracts in PubMed leveraging the large volume of structured abstracts already available [3–5]. However, structured abstracts are uncommon in other disciplines like computer science, although, arguably, the same benefits may be reaped if structured abstracts were available (see, e.g., [1]).

In this paper, we investigate if sentences in an abstract can be labeled with discourse categories using machine learning methods even if the labeled data is sparse. We take abstracts in computer science (CS) as a case study. We categorize sentences in a CS abstract into three classes: BACKGROUND, TECHNIQUE, and OBSERVATION. We adopt a deep learning model for sequential sentence classification pretrained on structured abstracts from PubMed. We prepare a small corpus of hand-labeled abstracts in CS. We fine-tune the model on a subset of the corpus and test it on the remainder. Fig. 1 shows an abstract from *IEEE Transactions on Learning Technologies* in which each sentence is labeled with one of the three classes. Both hand-annotated golden labels and the predictions done by the model are indicated. We observe an accuracy of 75% on our test corpus, which is quite promising, given the limited amount of golden data. In brief, our contributions are

1. We propose a simplified discourse structure for CS abstracts.
2. We prepare a hand-labeled corpus of structured CS abstracts.
3. We use transfer learning to automatically classify sentences in CS abstracts into the above discourse categories.

The code and the datasets are available publicly<sup>1</sup>. The rest of the paper is structured as follows. Section 2 describes the discourse categories for CS abstracts and the rationale behind choosing this structure, Section 3 introduces the datasets, Section 4 describes the machine learning model, Section 5 evaluates of the model, and finally Section 6 concludes the paper.

## 2 DISCOURSE CATEGORIES

Abstracts in biomedical papers in PubMed are structured into five classes: BACKGROUND, OBJECTIVE, METHOD, RESULT, and

<sup>1</sup><https://github.com/soumyaxyz/abstractAnalysis>

Blending Digital and Face-to-Face Interaction Using a Co-Located Social Media App in Class	
Improving face-to-face (f2f) interaction in large classrooms is a challenging task as student participation can be hard to initiate.	BACKGROUND BACKGROUND
Thanks to the wide adoption of personal mobile devices, it is possible to blend digital and face-to-face interaction and integrate co-located social media applications in the classroom.	BACKGROUND BACKGROUND
To better understand how such applications can interweave digital and f2f interaction, we performed a detailed analysis of real-world use cases of a particular co-located social media app: SpeakUp.	TECHNIQUE TECHNIQUE
In a nutshell, SpeakUp allows the creation of temporary location-bound chat rooms that are accessible by nearby users who can post and rate messages anonymously.	TECHNIQUE OBSERVATION
We find that the use of co-located social media is associated with an increase in content-related interaction in the class.	OBSERVATION OBSERVATION
Furthermore, it is associated with an increase in the perceived learning outcomes of students compared to a control group.	OBSERVATION OBSERVATION
We further provide design guidelines to blend digital and f2f interaction using co-located social media in the classroom based on 11 case studies covering over 2,000 students.	OBSERVATION OBSERVATION

Figure 1: An abstract, from *cs.TLT* dataset, with golden and predicted ( correct and erroneous ) labels.

CONCLUSION. As similar segmented CS abstracts are not available, we employed human annotators to prepare our own datasets. However, the annotators found it difficult to consistently annotate the sentences of CS abstracts into five distinct classes similar to those of the PubMed abstracts. Our discussions with the annotators revealed that the OBJECTIVE is hardly explicated; rather, the same sentence mixes the OBJECTIVE and the METHOD. These findings motivated us to adopt a compressed discourse structure where such sentences are labeled as TECHNIQUE. If a sentence is clearly an OBJECTIVE, it has been labeled as BACKGROUND. Similarly, CS abstracts typically report some qualitative or quantitative findings without always making a general comment. Such sentences have been labeled as OBSERVATION. There can be a rare fourth sentence label, CODE, but the sentences in this class can be easily detected with regular expressions in a preprocessing step. Therefore, we remove such lines from the corpus, before feeding the abstracts to the machine learning model. Thus, a CS abstract is segmented into three classes: BACKGROUND, TECHNIQUE, and OBSERVATION. The mapping between the two structures is shown in Table 1.

BACKGROUND	BACKGROUND
OBJECTIVE	
METHOD	TECHNIQUE
RESULT	OBSERVATION
CONCLUSION	

Table 1: Mapping between the discourse categories in biomedical and computer science papers.

### 3 DATASET

We introduce five new datasets of segmented abstracts in this paper. First, we introduce the *PubMed-non-RCT* corpus containing abstracts of articles from PubMed that do not report randomized control trial (RCT). Like its RCT-analog in [4], it contains 20k abstracts, structured into the five classes already described.

Dataset	Classes	Train	Validation	Test
<i>PubMed-non-RCT</i>	5	15k (165681)	2.5k (26992)	2.5k (24054)
<i>cs.NI</i>	3	110 (1224)	-	40 (460)
<i>cs.TLT</i>	3	110 (928)	-	40 (326)
<i>cs.TPAMI</i>	3	110 (901)	-	40 (326)
<i>cs.combined</i>	3	330 (3053)	-	120 (1112)

Table 2: Dataset summary. Columns 3–5 show the number of abstracts in each subset, with the total number of lines in bracket.

We created four corpora of CS abstracts: (1) *cs.NI*: 150 abstracts on Networking and Internet Architecture from arXiv, (2) *cs.TLT*: 150 abstracts from the journal *IEEE Transactions on Learning Technologies*, (3) *cs.TPAMI*: 150 abstracts from the journal *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and (4) *cs.combined*: the aggregation of the above three CS corpora. Unlike *PubMed-non-RCT*, these abstracts were not available in structured format. So, we first segmented each CS abstract into sentences with the spaCy<sup>2</sup> library. Each sentence in *cs.TLT* and *cs.TPAMI* abstracts was then labeled with one of the three classes by two independent annotators who are CS engineering graduates and, therefore, familiar with the domains. The inter-annotator agreement was found to be very high with Cohen’s Kappa [2] of 0.87 and 0.91, respectively. In case of *cs.NI*, the dataset was distributed among senior CS undergraduates. Subsequently, the annotations were reviewed by a CS doctoral scholar. Disagreements were resolved through mutual discussion. In a few instances, the sentence segmentation generated with spaCy was erroneous. The annotators corrected them manually. The requirement of domain experts precluded the generation of very large labeled datasets in CS. Table 2 summarizes the datasets and how each is divided into train, validation, and test subsets.

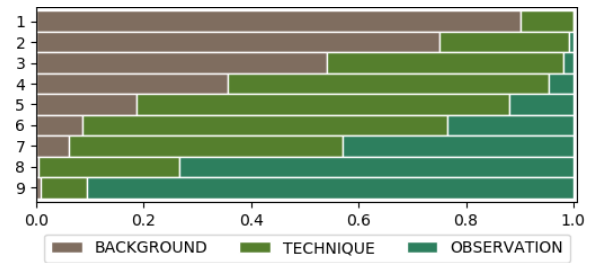


Figure 2: Distribution of labels in *cs.combined*, visualized.

The average number of lines in a CS abstract is approximately nine. Figure 2 shows the distribution of labels in the *cs.combined* corpus. Normalizing all abstracts to nine sentences, the figure demonstrates what fraction of which lines belong to which label. It is apparent that the abstracts in the dataset roughly follow the label sequence: BACKGROUND → TECHNIQUE → OBSERVATION.

<sup>2</sup><https://spacy.io>

## 4 DEEP LEARNING MODEL FOR DISCOURSE CATEGORY IDENTIFICATION

The proposed model utilizes the state-of-the-art sequential sentence classification architecture proposed by Jin and Szolovits [5]. Architecturally, the model is composed of four conceptual layers.

- 1) The token processing layer: This layer, for a sentence of  $n$  tokens  $\{w_1, \dots, w_n\}$ , generates embeddings for each token and produces an  $n \times D$  matrix, where  $D$  is the embedding dimension. Pretrained GloVe word vectors [7] are used to initialize the input token vectors.
- 2) The sentence processing layer: This layer accepts the output from the token processing layer and calculates the encoding for the whole sentence. The sentence encoding is calculated by passing the embedding matrix through a bidirectional LSTM and then applying self-attention on the output.
- 3) The abstract processing layer: Unlike the previous layers, this layer operates on a complete abstract at once. The layer accepts a matrix of  $N$  vectors  $\{s_1, \dots, s_N\}$ , where each  $s_i$  is a sentence embedding. A bidirectional LSTM transforms the input matrix, which is then passed to a single dense layer. The output from this layer corresponds to the per-class score for each sentence.
- 4) The output generation layer: This final layer optimizes the prediction of the abstract processing layer by modeling the transition probability between two consecutive sentence labels with conditional random field (CRF) [9].

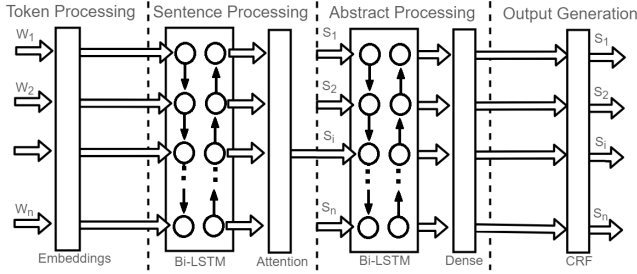


Figure 3: Model architecture.

The model architecture is summarized in Figure 3. During training, Adam optimizer [6] and categorical cross-entropy loss function are applied. The model is first trained on the large *PubMed-non-RCT* corpus of structured biomedical abstracts. The resultant model is then fine-tuned on a much smaller CS dataset, i.e., the training subset of every *cs.\** dataset mentioned in Table 2, and tested on the corresponding test subset.

## 5 EVALUATION OF THE MODEL

In this section, we evaluate the proposed model and validate our claims.

### 5.1 Evaluation

The *PubMed-non-RCT* corpus was converted into a three-class dataset according to the mapping described in Table 1. Then our proposed model was trained on it. Jin and Szolovits reported 92.6% accuracy on their *PubMed 20k RCT* dataset. Our model achieves

a comparable 92.1% accuracy on the biomedical *PubMed-non-RCT* dataset. The model trained on *PubMed-non-RCT* was saved. Next, we evaluated the model on each CS corpus listed in Table 2. For each corpus, we evaluated in 3 ways, as shown in Table 3: (1) *Locally-trained*: The deep net model was trained only on the training subset and tested on the test subset of the CS corpus. (2) *Pre-trained on PubMed*: The model was trained only on the *PubMed-non-RCT* dataset and tested on the test subset of the CS corpus. (3) *Fine-tuned*: The model, pretrained on the *PubMed-non-RCT* dataset, was fine-tuned on the training subset and tested on the test subset of the CS corpus. The results and the inferences are discussed in the following section.

### 5.2 Results and Analysis

Table 3 summarizes the accuracy of the locally trained model, the pre-trained model, and the fine-tuned model on all the four test subsets. Note that the accuracy of the model on a dataset is calculated as the percentage of sentences that are correctly labeled across all abstracts in the dataset.

Dataset	Details	Accuracy
<i>cs.NI</i>	Locally trained	54.08
	Pre-trained on PubMed	29.46
	Fine-tuned	<b>65.22</b>
<i>cs.TLT</i>	Locally trained	61.98
	Pre-trained on PubMed	56.87
	Fine-tuned	<b>79.45</b>
<i>cs.TPAMI</i>	Locally trained	71.28
	Pre-trained on PubMed	50.15
	Fine-tuned	<b>83.44</b>
<i>cs.combined</i>	Locally trained	41.73
	Pre-trained on PubMed	39.53
	Fine-tuned	<b>75.18</b>

Table 3: Summary of results.

We can observe that transfer learning with fine-tuning ('Fine-tuned') provides a significant improvement over the 'locally trained' model. We also observe that without fine-tuning ('Pre-trained on PubMed'), transfer learning performs worse than local training. This indicates that the pretrained model is insufficient. It also gives an estimate of how dissimilar the particular CS corpus is compared to the biomedical corpus; if the CS corpus is similar to the biomedical corpus, further fine-tuning would be unnecessary. The results indicate that *cs.TLT* and *cs.TPAMI* have modest similarity to *PubMed-non-RCT* while *cs.NI* has hardly any similarity as the 29.46% accuracy is as bad as a purely random guess. Nevertheless, we observe that, even when the datasets are completely dissimilar, transfer learning with fine-tuning on a marginal amount of labeled data can provide more than a 10% accuracy boost.

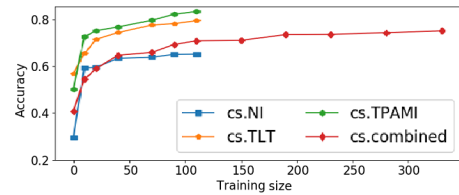


Figure 4: Effect of training size on accuracy.

Figure 4 demonstrates the effect of training size on the accuracy of the model on the four CS corpora. We observe that, initially the

accuracy increases rapidly with the training size. For comparatively larger size of the training corpus, accuracy still increases, albeit more gradually.

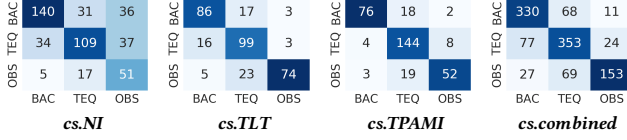


Figure 5: Confusion matrices.

Figure 5 presents the confusion matrices corresponding to the best runs (i.e., for the *Fine-tuned* model) on the test subsets of the four CS corpora. Table 4 presents the per-class precision, recall, F1-score, and support, corresponding to the same runs.

Dataset	Label	P	R	F	S
cs.NI	BACKGROUND	78.21	67.63	72.54	207
	TECHNIQUE	69.43	60.56	64.69	180
	OBSERVATION	41.13	69.86	51.78	73
cs.TLT	BACKGROUND	80.37	81.13	80.75	106
	TECHNIQUE	71.22	83.90	77.04	118
	OBSERVATION	92.50	72.55	81.32	102
cs.TPAMI	BACKGROUND	91.57	79.17	84.92	96
	TECHNIQUE	79.56	92.31	85.46	156
	OBSERVATION	83.87	70.27	76.47	74
cs.combined	BACKGROUND	76.04	80.68	78.29	409
	TECHNIQUE	72.04	77.75	74.79	454
	OBSERVATION	81.38	61.45	70.02	249

Table 4: Results expressed in terms of (P)recision, (R)ecall, (F)1 and (S)upport.

It is clear that the relatively poor accuracy for *cs.NI* is the result of the failure of our model to identify the sentences belonging to the class OBSERVATION. In general, abstracts tend to follow the pattern BACKGROUND → TECHNIQUE → OBSERVATION. However, as reported by the annotators, the abstracts in *cs.NI* often deviate from this trend. Moreover, *cs.NI* is quite imbalanced, with significantly fewer instances of OBSERVATION. All these factors contribute to the model’s difficulty with the *cs.NI* dataset. However, based on our observation in Fig 4, we expect with a larger volume of training data, these limitations can be overcome.

Details	Accuracy on cs.NI test subset
Fine-tuned on original cs.NI	65.22
Fine-tuned on noisy and augmented cs.NI	67.99

Table 5: Experiment with noisy data.

To test this hypothesis, we performed an experiment. We took the model fine-tuned on *cs.combined* dataset and used it to predict the labels for 150 additional abstracts from arXiv.CS.NI. These abstracts had no corresponding golden labels but we expect the predictions to have accuracy comparable to those of the test subset of *cs.combined*. We augmented the *cs.NI* training subset, so that now it contains  $110 + 150 = 260$  examples (albeit of somewhat questionable integrity). Then, the model that was pre-trained on PubMed was fine-tuned with this augmented corpus. We observed a 2% increase in accuracy on the the original *cs.NI* test subset (Table 5). This vindicates our hypothesis.

### 5.3 Ablation Analysis

The ablation study of the proposed model in the context of *cs.combined* is depicted in Table 6 where we subtract one layer at a time and report the changed accuracy. We observe that the abstract processing step, with the bidirectional LSTM to extract the conceptual information from the sentences, contributes the greatest toward the model’s performance. Embeddings, with their capability to represent the contextual relationship between the tokens, has the second-highest contribution. The CRF in the output generation layer also has a major impact on the model’s performance. This layer takes advantage of the sequential dependency in the output label sequence. This ties into the poor performance on the *cs.NI* dataset, as the sequential information is somewhat lacking in that dataset.

Dataset	Accuracy
Full Model	75.18
– Token Processing	63.56
– Sentence Processing	70.67
– Abstract Processing	43.15
– CRF in Output Layer (replaced with softmax)	66.02

Table 6: Summary of ablation analysis.

## 6 CONCLUSION

In this paper, we proposed a method for automatic discourse classification of sentences in computer science abstracts. We demonstrated that transfer learning with fine-tuning can provide remarkable results even on a sparsely labeled dataset. We observed that due to the difference in presentation style, the nature of the discourse classification of CS abstracts vary across sub-fields. Nevertheless, the results on *cs.combined* demonstrate that the proposed model generalizes fairly well across sub-fields of CS.

## ACKNOWLEDGMENTS

This work is supported by the *National Digital Library of India Project* sponsored by the Ministry of Human Resource Development, Government of India at IIT Kharagpur.

## REFERENCES

- [1] CHAN, J., CHANG, J. C., HOPE, T., SHAHAF, D., AND KITUR, A. SOLVENT: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 31.
- [2] COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
- [3] DERNONCOURT, F. *Sequential short-text classification with neural networks*. PhD thesis, Massachusetts Institute of Technology, 2017.
- [4] DERNONCOURT, F., AND LEE, J. Y. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071* (2017).
- [5] JIN, D., AND SZOLOVITS, P. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. *arXiv preprint arXiv:1808.06161* (2018).
- [6] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [7] PENNINGTON, J., SOCHER, R., AND MANNING, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.
- [8] RIPLE, A. M., MORK, J. G., ROZIER, J. M., AND KNECHT, L. S. Structured abstracts in MEDLINE: Twenty-five years later. *National Library of Medicine* (2012).
- [9] SUTTON, C., MCCALLUM, A., ET AL. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning* 4, 4 (2012), 267–373.
- [10] US NATIONAL LIBRARY OF HEALTH. Structured abstracts. [https://www.nlm.nih.gov/bsd/policy/structured\\_abstracts.html](https://www.nlm.nih.gov/bsd/policy/structured_abstracts.html).