Sequential Sentence Classification for Longer Documents

Minhwa Lee and **Naveen Jafer Nizar** and **Pranjali Ajay Parse**
College of Information and Computer Sciences
University of Massachusetts Amherst

Abstract

The task of sequential sentence classification (SSC) has been shown to perform well on short-length documents. However, SSC is less successful on long-form documents such as full papers and clinical notes. Our work made four contributions to the SSC task. First, we quantify the factors that explain the performance gap between short and long-form documents. Second, we study the effect of longer context in one of the baselines (Cohan et al., 2019)'s approach to SSC task. Third, we pose the problem of SSC as an intent of discrete classification. Lastly, we show a transfer learning setup with the five datasets of interest, which raises questions of the performance of SSC in such a setup.

1 Introduction

The broad task of document-level understanding of texts benefits from categorization of sentences into their structured rhetoric status. One way to do can be performed at a sentence level, by assigning labels to a sequence of sentences. We call this task **S**equential **S**entence **C**lassification (SSC). Figure 1 describes how SSC tasks would be performed on a short passage.

2 Background

2.1 Motivation

Although SSC has been successfully applied to short passages such as scientific abstracts, its performance on classifying sequence of sentences in longer documents has shown low-quality results. Examples of such documents include scientific papers (Brack et al. (2021), Cohan et al. (2019)), clinical notes (Mullenbach et al., 2021), and legal documents (Grover et al., 2004).

Brack et al. (2021) uses pre-trained language models and transfer learning on different combinations of datasets, sub-domains and shared weights to solve SSC for longer documents. Although the results from Brack et al. (2021) show that its classification results improve across all datasets including full papers and abstracts, comparatively smaller datasets tend to benefit the most from sharing weights from the larger ones.

Also, the common theme across all previous work is the lack of failure analysis, which did not explicitly point out certain causes of errors such as why current pre-trained language models do capture context of sequential sentences and how effectively positions of sentences in a sequence impacts on the language models' performance on classification results.

2.2 Related Work

Hierarchical sequence labeling was a primary approach for the previous work on SSC. For instance, (Jin and Szolovits, 2018) uses contextualized sentence vectors resulted from bidirectional Long-Short-Term-Memory (BiLSTM). Then, the vectors are passed onto a conditional random field (CRF) (Lafferty et al., 2001) layer, which captures interdependency between labels in a short passage.

Ye and Ling (2018) suggests that neural Semi-Markov CRF at the top layer can predict the rhetorical label of spans of sentence representation vectors, by considering all possible spans of various lengths. This approach of forming spans of sentences is effective to continuous, long sentences. The drawback of the approach is, however, that it is computationally expensive to implement.

After the advent of large-scale bi-directional language models such as BERT (Devlin et al., 2019), new approaches to SSC task have been proposed in recent years. Cohan et al. (2019) uses pre-trained weights of SciBERT (Beltagy et al., 2019a) to directly obtain contextualized sentence representations. Then, the sentences separated by the delimiter token [SEP] are finetuned by SciBERT and the

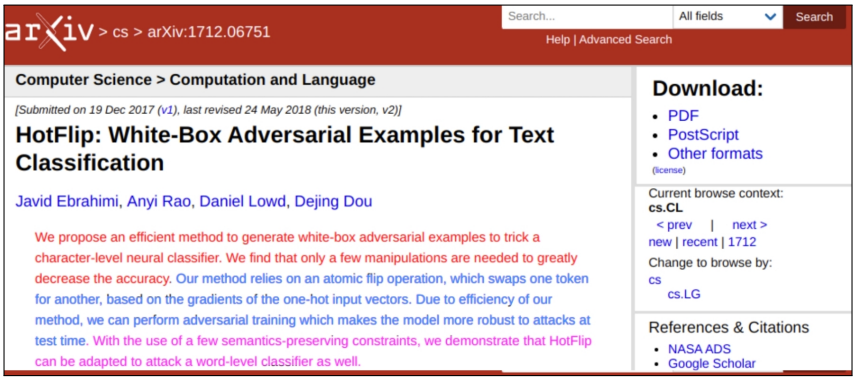


Figure 1: An example of how sequential sentence classification would be performed on an abstract from the CS Abstruct dataset. The lines in red, blue and pink belong to background, method and result respectively.

output of the [SEP] tokens from a multi-layer feed-forward layer is used to predict the rhetorical label of the corresponding sentences in abstracts.

The recent study uses a hierarchical sequence labeling network to scientific full papers and abstracts (Brack et al., 2021). Word embeddings are contextualized into sentence level using SciBERT and a BiLSTM layer. These sentence-level embeddings are further contextualized with a BiLSTM for the purpose of context enrichment from surrounding sentences, which is thus passed onto a CRF layer to obtain class labels of the input sequence. They also leveraged transfer learning setups to investigate what setups benefit the most from transfer learning and the nature of shared weights that are most optimal among others.

3 Baselines

3.1 Dataset Oveview

- **PubMed (Namata et al., 2012):** a publicly available dataset (based on PubMed) that contains approximately 200k abstracts, totaling 2.4 million sentences.

- **CSAbstruct (Cohan et al., 2019):** A collection of annotated Computer Science abstracts from Semantic Scholar corpus with sentence labels according to their rhetorical roles.

- **DRI (Fisas et al., 2015):** It is made of 40 Computer Graphics papers, which have been obtained from the authors. The labels consist of five top level categories and three subcategories.

- **ART (Liakata et al., 2010):** A set of 225 papers across topics in physical chemistry and biochemistry, with and CSConcepts-style (Core Scientific Concepts) annotations.

- **CLIP (Mullenbach et al., 2021):** A credentialed dataset of 718 annotated clinical action items excerpted from MIMIC-III (Johnson et al., 2016).

3.2 Baselines

We use the following as our baseline for the SSC task, as shown in Figure 2 : (1) The BERT-based model proposed in Cohan et al. (2019) and (2) the hierarchical sequence classification network from Brack et al. (2021).

As explained in Section 2.2, the subfigure 2a directly uses SciBERT's pre-trained weights to create sentence-level embeddings, whose finetuned [SEP] tokens that are also passed onto the linear layer are used to predict the labels of the input sequence. In contrary, the model described in the subfigure 2b uses two-step contextualized embeddings of sentences in the input sequence by using SciBERT and Bi-LSTM, which is then passed to a CRF layer and predict the class labels of the input sequence.

3.3 Baseline Results

In Table 3, we present the results from the two baseline models on each of the five datasets. The two models showed the same F1 score on the test sets of PubMed-20k and CLIP in the same manner, respectively. We observe that compared to other datasets, the baseline model from regular SciBERT (Beltagy et al., 2019a) improves F1 score significantly on classifying the labels of sentences in the test set of DRI.

Characteristics \ Datasets	CSAbstruct	PubMed-20k	DRI	ART	CLIP
Domain	Computer Science	Biomedicine	Computer Graphics	Chemistry	Clinical
Text type	Abstracts	Abstracts	Full papers	Full Papers	Discharge Notes
# documents	2200	20000	40	225	718
# sentences	14708	2.3M	8817	34905	107,494
# avg sentences per document	7	115	215	154	143
# vocabs	293265	1M	16693	68592	18026
# labels	5	5	5	11	8
Train/dev/test split	76/14/10	N/A	70/10/20	70/10/20	74/13/13
OOV Rate	0.295	0.37	0.28	0.46	0.34
Structural Entropy	0.61	0.22	0.65	0.78	0.64

Table 1: Characteristics of the six datasets. Out-of-Vocabulary rate is defined as the number of words in the dev/test sets that are not mentioned in the training set. Each scores has been computed using the NLTK library's tokenizer. See Section 4.1. for further details about structural entropy.

Datasets	**Sentence Labels**
CSAbstruct	Background, Objective, Methods, Results, Other
PubMed-20k	Background, Objective, Methods, Results, Conclusion
DRI	Background, Approach, Challenge, Outcome, Future Work
ART	Background, Object, Method, Conclusion, Results, Goal, Motivation, Hypothesis, Model, Experiment, Observation
CLIP	Appointment, Lab, Procedure, Medication, Imaging, Patient Instruction, Other

Table 2: Annotated Labels of sentences in the five datasets. We observed that labels of sentences in HOLJ dataset is not labeled in sentence level, which is then excluded for the analysis in the table.

Dataset	**Test F1 (Cohan)**	**Test F1 (Brack)**
CSAbstruct	0.80	**0.83**
PubMed-20k	0.88	0.88
ART	**0.50**	0.46
DRI	0.55	**0.70**
CLIP	0.58	0.58

Table 3: Baseline results on the five datasets. Bolded are the higher F1 scores on the test set of each datasets between the two models.

4 Exploratory Data Analysis

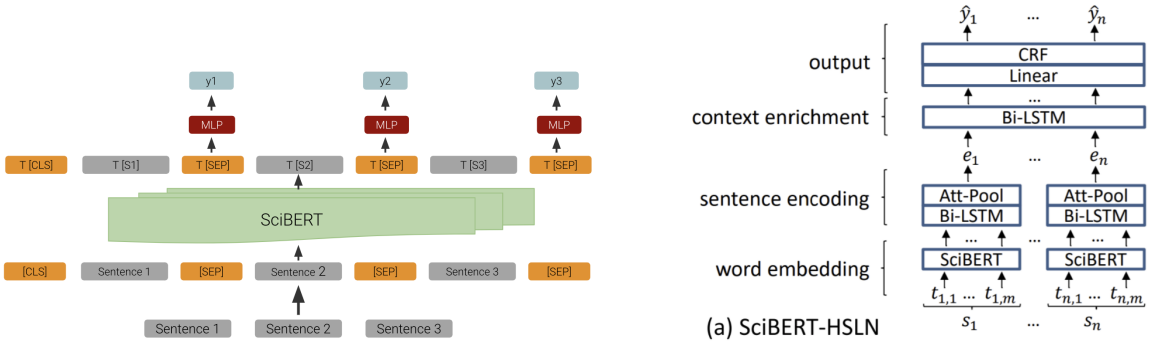
4.1 Label Transitions

First, we carry out a study on the entropy in structure. The corresponding procedures are illustrated with several sankey diagrams as shown in Figure 3. The labels on the left side of each diagram prefix the labels on the right side. We observe the following results from the diagrams:

- For a sentence of label type A, the succeeding sentence in the dataset is often of type A.
- Suppose a sentence labeled with type A (on the left) and labeled with type B (on the right). If A contributes to a large fraction of all connections into B, we infer that B's conditional dependence on A is high. We observe that there frequently occurs such labels that are conditionally dependent on other labels. In the datasets where this occurs frequently, one can infer that the entropy of structure is low.
- PubMed, CSAbstruct and CLIP exhibit low entropy in structure. However, in ART the inter-label transitions do not exhibit strong structure like the others.

A naive way of interpreting results in Figure 3 is that labels of a certain type occur in chunks of the same label type. However, the strong structure in intra-label sequences (A -> A) raises a question of how often a label of type $\notin$ A occurs which is also suffixed and prefixed by sentences of type A in a given dataset. We examine this occurrence of label transitions using several sankey diagrams as presented in Figure 4. The width of the flow indicates how often a label in the left column occurs such that it is prefixed and suffixed by the label in the right column.

We formalize the figures in Figure 4 as $Structural Entropy$ and report it in Table 1. The calculation is done by averaging the normalized entropy on each row of the label transition matrix.



(a) Baseline - (1) (Cohan et al., 2019)

(b) Baseline - (2) (Brack et al., 2021)

Figure 2: Architectures of the two baseline models.

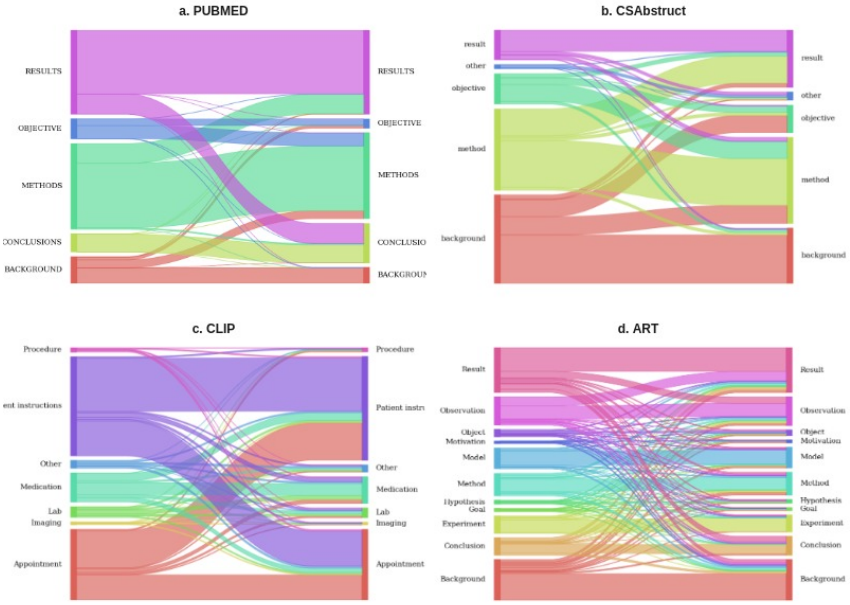


Figure 3: General label transitions for PubMed, CSAbstruct, CLIP and ART datasets

A lower score corresponds to higher signal from structure and vice versa.

This information can be used to build adversarial examples that follow anti-patterns. This can serve as a model-agnostic adversarial evaluation that is able to test how robust a model is to structures of the labels in the given datasets.

4.2 Label Structure

Relative position (as a proxy for global context) in the document can be an important signal in documents such as scientific papers with specific conventions, more so than the immediate local context. Therefore, as the second step of our analysis, we perform positional analysis on all our scientific full paper, abstracts, and datasets in the miscellaneous domain, as shown in the Figure 5. We identify that most of paper-format datasets (e.g., CSAbstruct, PubMed-20k, ART, DRI) that are irrespective of domain show a certain convention that they follow. For instance, a sequence of scientific sentences from CSAbstruct begins with a background, followed by objective, methodology and result, respectively. Although CLIP belongs to a completely different domain (e.g., Clinical), we observe that it also shows a similar pattern of label structures as like other scientific-domain datasets. These results depict the importance of relative position and structural conventions in longer documents.

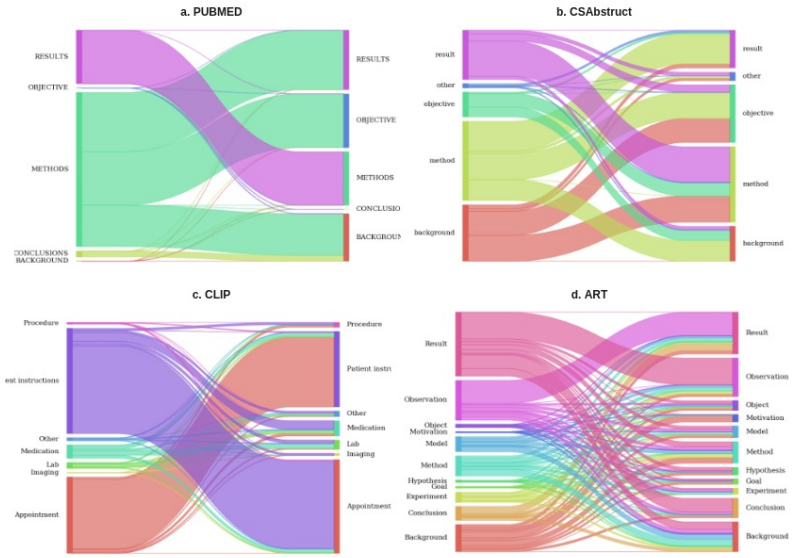


Figure 4: Label sentence transitions for PubMed, CSAbstruct, CLIP and ART datasets, where a label in the left column is prefixed and suffixed by each label in the right column.

4.3 Ablations

To further understand the effect of structure in labels, we generate the following modifications of the original formats of five datasets and run them on the second baseline model (Brack et al., 2021).

- **Shuffled-all** All of train, dev and test sets in each of the five datasets are shuffled internally within each document. By doing so, we would understand how the two baseline models use semantic information in the absence of positional signal of sentences.
- **Shuffled-Test** Only test set is shuffled internally. Thus, we would observe how overly reliant ordering exists in the models that have been trained on any structured dataset.
- **Non-Contextual** No preceding and succeeding context are given to the instances in test set, so we can examine if baseline models can predict labels of an input sequence without any position and context information.
- **Position-Only** A sentence in all three of train, dev, and test sets is replaced with a string that is corresponding to index of the sentence. By doing so, we may understand if this approach can make a prediction without any content signal.

The results of these experiments are presented in Table 4. For PubMed-20k and CSAbstruct, the F1 scores of their position-only ablations are higher than non-contextual ablation. It indicates the fact that for these two datasets, the baseline models rely on signal from label structures more heavily than from the contents of sentences.

5 Experiments

5.1 Long-range Language Modeling

We observed that structures of sequential sentences highly impact the baselines' prediction results, compared to the context of the sentences. This means that the positional-only signals without any surrounding context and content can achieve fair scores for its prediction performance. Thus, we wanted to observe whether this model can be improved by providing long-range surrounding context. As such, we developed a long-range language model, by using Longformer (Beltagy et al., 2020).

As presented in Figure 6, the model directly uses the longformer layers pretrained with existing SciBERT (Beltagy et al., 2019a). Through longformer's training, the encodings of [SEP] tokens learn contextual information of each sentence in the input sequence, and those tokens are also passed onto the linear and softamx layers to predict the labels of each sequential sentence in the input. Unlike the baseline (1) (Cohan et al., 2019) that handles long sequences of more than 512 tokens by the recursive bisections of sentences, the longformer model is capable of taking an input sequence of 4096 tokens

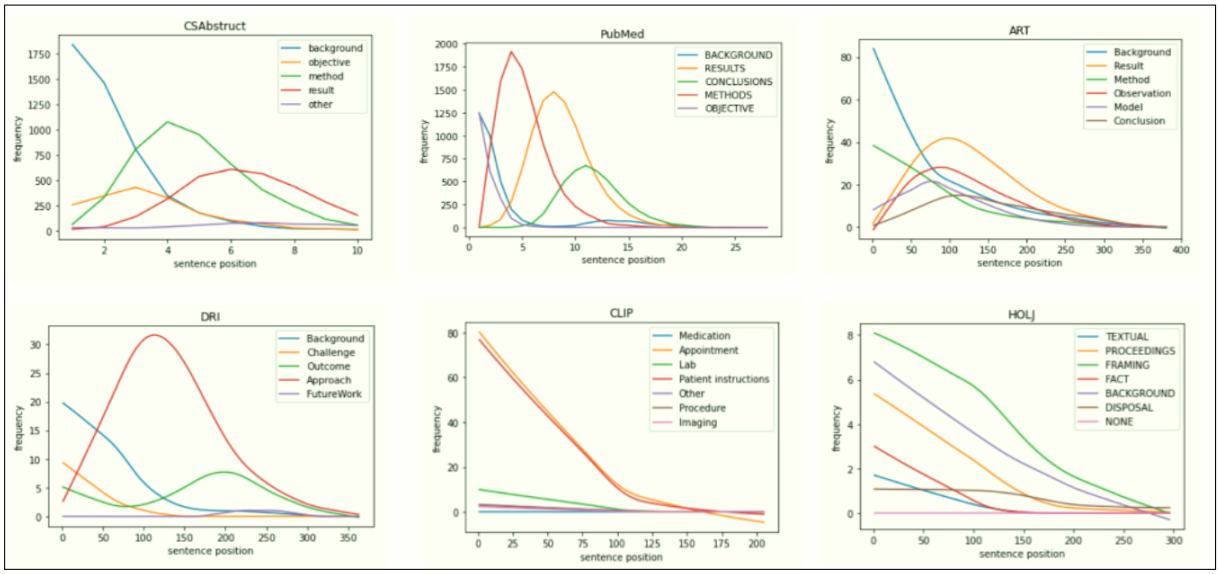


Figure 5: Positional Analysis on different datasets

Dataset \Experiments	Original	Shuffled-All	Shuffled-Test	Non-contextual	Position-Only
CSAbstruct	0.83	0.74	0.49	0.50	0.57
CLIP	0.79	0.64	0.59	0.53	0.44
ART	0.55	0.51	0.48	0.43	0.21
DRI	0.83	0.77	0.66	0.63	0.20
PubMed-20k	0.92	0.92	0.92	0.61	0.69

Table 4: Weighted F1 score for the datasets with the second baseline model from (Brack et al., 2021)

at maximum. Due to the limits on computational resources, we took an input of sequential sentences of 1000 tokens at maximum. Also, we use the batch size of 4 , dropout of 0.1, the Adam optimizer for 20 epochs, and learning rates of $5e^{-6}$,$1e^{-5}$, $2e^{-5}$, or $5e^{-5}$. We chose hyperparameters based on the best performance on the validation set. Finally, we finetuned the model on each of the five datasets.

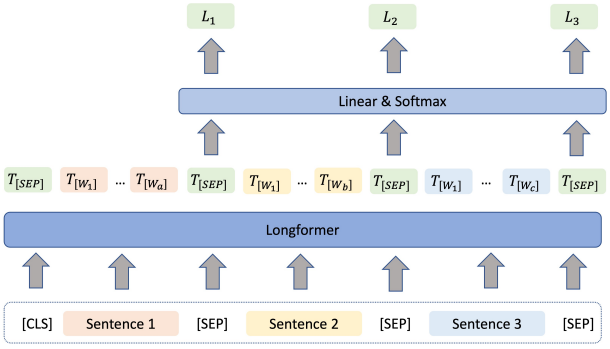


Figure 6: Model architectures with Longformer (Beltagy et al., 2020)

Table 5 presents the F1 scores of the best model on the test set of each of five datasets. The results from all of the five datasets show that our longformer model achieved either comparable or lower F1 scores on test sets than the baseline. Using a long-range language model did not improve the prediction performance, even though the longformer model used more surrounding contexts. Thus, we suggest that signals from long-range surrounding context do not supplement short range structural information when predicting the labels of sequential sentences. These observations motivate our second set of experiments where we treat the problem as a discrete classification problem with no distracting context.

Dataset	Baseline (1)	Our Model
CSAbstruct	0.83	0.80
CLIP	0.79	0.78
ART	0.55	0.46
DRI	0.79	0.73
PubMed-20k	0.93	0.92

Table 5: Comparison of F1 scores between Baseline (1) (Cohan et al., 2019) and our Longformer model on test sets of each five datasets.

5.2 Discrete Classification

We pose the task of SSC as a discrete sentence classification task, for the following reasons:

- To compare against the SSC task in a transfer

learning setup detailed in Section 5.3.

- To understand dataset-specific saturation points of purely content-based signals in making predictions.

We experiment with a SciBERT (Beltagy et al., 2019b) model for classification. The results in Table 6 indicate that the model performs at least as well as the shuffled version of the ablation studies. For the CLIP dataset, the model performs much better than the all-shuffled ablated versions.

Dataset	SSC	SSC shuffled	Disc. CL
CSAbstruct	0.83	0.74	0.75
CLIP	0.79	0.64	0.70
ART	0.56	0.52	0.52
DRI	0.83	0.77	0.79

Table 6: F1 scores comparing the baseline (2) (denoted as "SSC"), the shuffled ablation ("SSC shuffled"), and the discrete classification task ("Disc. CL")

5.2.1 Augmentation

We also experiment with augmmentation to study the extent to which content-based signals can improve performance in a self-supervised augmentation setup. With the two-stage process described below, a T5 model (Raffel et al., 2019) is first trained and inference from the model generates the augmentation set.

First, the training process follows the five-stage process below:

1. Pick an utterance (an instance from the training data) and label pair.
2. Randomly drop each token from the utterance with a probability of 0.4.
3. Reorder the words randomly.
4. Concatenate the label with the tokens.
5. Feed this as input to a T5 model with a learning objective of reproducing the original utterance.

Then, the generation process has been completed as follows:

1. Obtain a mapping between class label and token frequency.
2. Remove stop word tokens from this mapping
3. For each label perform weighted sampling of tokens. The number of tokens sampled follows the same distribution of the examples used during generation.
4. T5 generates examples using the input from the previous step to form the augmentation set.

We report the results from the two experiment setups in which (1) we double the dataset size with augmentations ("**2x**") and (2) we generate examples to balance the number of instances per class ("**Balanced**"). Table 7 shows that the F1 scores slightly increased for three out of the four datasets in the setup where the data size was doubled. The F1 scores dropped for all datasets in the balanced setup. We suspect that the drop in F1 scores happens due to severe imbalance that persists in the test set.

Dataset	Base	2x	Balanced
CSAbstruct	75.4	**76.3**	74.2
CLIP	**70.1**	69.9	67.8
ART	52.1	**52.2**	51.3
DRI	78.5	**79.9**	76.0

Table 7: The F1 scores for Discrete classification across the three following setups: (1) baseline, (2) doubled size of data and (3) balanced dataset setup. Note that for the setup (2) possesses twice as many training samples as in the base setup. Half of them comes from the base model and rest from generation during augmentation.

5.3 Label Unification

We also experiment with a data-driven method of label unification, in order to study the application of transfer learning on a completely new dataset or a few shot setup. With this approach, we combine the datasets {CSAbstruct, ART, DRI, PubMed-20k} into a standard form, based on the semantic similarity of classes across various domains, as shown in Figure 7.

We conduct 13 different experiments with the baseline (1) (Cohan et al., 2019) to examine this approach, where the results are shown in Table 8:

1. **All Unified + All Tested (1 experiment)**: Train the model on all datasets and test it on all combined datasets.
2. **All Unified + Single Testing (4 experiments)**: Train the model on all datasets and test the model on individual datasets.

Experiment	Training Dataset	Testing Dataset	Test F1	Base F1	Vanilla Classification
All Unified + All Tested	D1, D2, D3, D4	D1, D2, D3, D4	0.89	-	-
All Unified + Single Testing	D1, D2, D3, D4	D1	0.78	0.79	-
All Unified + Single Testing	D1, D2, D3, D4	D2	0.61	0.67	-
All Unified + Single Testing	D1, D2, D3, D4	D3	0.59	0.61	-
All Unified + Single Testing	D1, D2, D3, D4	D4	0.86	0.87	-
3 Unified + Single Testing	D2, D3, D4	D1	0.39	0.79	0.55
3 Unified + Single Testing	D1, D3, D4	D2	0.42	0.67	0.52
3 Unified + Single Testing	D1, D2, D4	D3	0.41	0.61	0.58
3 Unified + Single Testing	D1, D2, D3	D4	0.55	0.87	0.59

Table 8: F1 Scores for Label Unification. **D1, D2, D3 and D4** refer to CSAbstruct, ART, DRI and PubMed-20k in this table.

3. **3 Unified + Single Testing (4 experiments)**: Train the model on any three datasets first and test the model on the last fourth dataset that is not trained on.

4. **Vanilla Classification with 3 Unified + Single Testing (4 experiments)**: Train the discrete classifier on any three datasets first and test the same model on the remaining fourth dataset that is not trained on.

As described in Table 8, we observe that in all the "All Unified + Single Testing" experiments, the baseline model in transfer learning setup performs similarly or poorly to the same model trained with only individual datasets. We suggest that there is no improvement in performance when we apply transfer learning on a completely new dataset. Further, the baseline performs poorly in all the "3 Unified + Single Testing" experiments. However, the discrete classifier achieves better results than the baseline, implying that sequential sentence classification does not bode well in either transfer learning or a few-shot setup.

6 Conclusion and Future Work

We demonstrated factors that influence the success of models on the task of sequential sentence classification. These observed factors are highly data-specific and weigh down performance of SSC models on longer documents. Our ablation study attempts to decompose the types of signals that the model of our study utilizes when making predictions of labels of sequential sentences. We observe that position-only signal contributes significantly to datasets with low structural entropy. However, the model is capable of utilizing content-only signals to make predictions under a shuffled-ablation setup despite its poor performance. Our experiments show that developing long-range language models do not improve performance even using signals from long-range surrounding context. The experiments with label unification show that discrete classification can work better than the SSC when performing transfer-learning setup for zero-shot intent classification. This observation aligns with the fact that SSC models rely on positional signals, which does not transfer between datasets.

6.1 Future Work

One feasible future work is to replicate human annotation of existing datasets with no contextual sentences. An aspect of research objectives beyond the scope of our study was the Out-of-Vocabulary (OOV) rate. The augmentation method in this work can address OOV issues if paired with information from a synonym resource.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019a. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. Scibert: A pretrained language model for scientific text.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Domain	Medical	Computer Science	Computer Graphics	BioPhysical and Chem
Dataset	PubMed	CSAbstruct	Dri	ART
	Background	Background	Background	Background
	Objective	Objective	Approach	Object
	Methods	Method	Challenge	Method
	Results	Result	Outcome	Conclusion
	Conclusion	Other	FutureWork	Result
				Goal
				Motivation
				Hypothesis
				Experiment
				Model
				Observation

Figure 7: Examples of Label Unification. Labels coded with same colors corresponds to a new label.

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2021. Sequential sentence classification in research papers using cross-domain multi-task learning. *ArXiv*, abs/2102.06008.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. 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.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Claire Grover, Ben Hachey, and Ian Hughson. 2004. The HOLJ corpus. supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54, Geneva, Switzerland. COLING.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and UMD EDU. 2012. Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8, page 1.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Zhixiu Ye and Zhen-Hua Ling. 2018. Hybrid semi-Markov CRF for neural sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–240, Melbourne, Australia. Association for Computational Linguistics.