A Span-based Dynamic Local Attention Model for Sequential Sentence Classification

Xichen Shang, Qianli Ma*, Zhenxi Lin, Jiangyue Yan, Zipeng Chen
School of Computer Science and Engineering, South China University of Technology
Key Laboratory of Big Data and Intelligent Robot
(South China University of Technology), Ministry of Education
shangxichen@foxmail.com, qianlima@scut.edu.cn*

Abstract

Sequential sentence classification aims to classify each sentence in the document based on the context in which sentences appear. Most existing work addresses this problem using a hierarchical sequence labeling network. However, they ignore considering the latent segment structure of the document, in which contiguous sentences often have coherent semantics. In this paper, we proposed a span-based dynamic local attention model that could explicitly capture the structural information by the proposed supervised dynamic local attention. We further introduce an auxiliary task called span-based classification to explore the span-level representations. Extensive experiments show that our model achieves better or competitive performance against state-of-the-art baselines on two benchmark datasets.

1 Introduction

The goal of Sequential Sentence Classification (SSC) is to classify each sentence in a document based on rhetorical structure profiling process (Jin and Szolovits, 2018), and the rhetorical label of each sentence is related to the surrounding sentences, which is different from the general sentence classification that does not involve context. An example is shown in Figure 1, the document is divided into rhetorical labels such as "background" and "outcome" for five sentences in NICTA dataset. The SSC task is crucial for downstream domains such as information retrieval (Edinger et al., 2017), question answering (Cohen et al., 2018) and so on.

Traditional statistical methods, such as HMM (Lin et al., 2006), CRF (Hirohata et al., 2008; Hassanzadeh et al., 2014), etc., heavily rely on numerous carefully hand-designed features. In contrast, recent methods based on end-to-end neural networks utilize hierarchical sequence encoders followed by the CRF layer to contextualize sentence representations, which achieved promising results. The first neural network approach (Lee and Dernoncourt, 2016) combined RNN with CNN that incorporates preceding sentences to encode the contextual content and further use a CRF layer to optimize the predicted label sequence. Recently, Jin and Szolovits (2018) propose a hierarchical sequential labeling network to make use of the contextual information within surrounding sentences to help classify. Conversely, Cohan et al. (2019) employ BERT (Devlin et al., 2018) to capture contextual dependencies without hierarchical encoding or CRF layer. Yamada et al. (2020) introduce Semi-Markov CRFs (Ye and Ling, 2018) to assign a rhetorical label at span-level rather than single sentence.

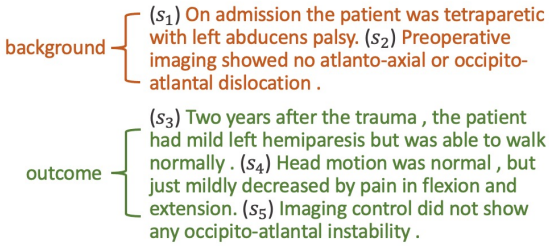

Figure 1: An example of NICTA dataset for SSC task. The text has five sentences and is divided into two segments $\{(s_1, s_2), (s_3, s_4, s_5)\}$ by labels.

Nevertheless, the above-mentioned methods ignore the latent structural information (e.g. segmentation) in the document, which is the grouping of content into topically coherent segments. Intuitively, a segment with several continuous sentences is expected to be more coherent semantics than the text spanning different segments, e.g., the text with two segments in Figure 1. In this paper, we propose a novel span-based dynamic local attention model to explore the latent segment structure in a document for SSC task. First, we introduce

*Corresponding author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 198–203
August 1–6, 2021.

dynamic local attention guided by segmentation supervision signal to focus on the surrounding sentences with coherent semantics, called Supervised Dynamic Local Attention (SDLA). Furthermore, we introduce an auxiliary task called span-based classification, which calculates semantic representations of spans and performs span classification on them to obtain predicted rhetorical labels. The dynamic local attention mechanism and the auxiliary task complement each other to enhance the model capacity to perceive segment structure and improve the performance of SSC task. The results on two benchmark datasets show that our method achieves better or competitive performance than state-of-the-art baselines.

2 Proposed Method

In this paper, we propose a Span-based Dynamic Local Attention Model for sequential sentence classification with two novel components: supervised dynamic local attention and auxiliary span-based classification task, respectively. The architecture of our model is shown in Figure 2.

2.1 Sentence Representations

For SSC task, given a sequence of sentences $X = \{x_1, x_2, \cdots, x_N\}$, the model needs to predict the label of each sentence $Y = \{y_1, y_2, \cdots, y_N\}$ based on the context which the sentence appears, where N is the number of sentences. Following the previous work (Yamada et al., 2020), we first feed each sentence into BERT pre-trained with PubMed (Peng et al., 2019) and then extract the encoding corresponding to [CLS] token as sentence encoding $S = \{s_1, s_2, \cdots, s_N\}$ (we implement it using Sentence-BERT (Reimers and Gurevych, 2019)). Then, we employ two bidirectional LSTM layers to produce context-informed sentence representation $h_i^c \in \mathbf{R}^d$ for whole document :

$$H^c = \{h_1^c, h_2^c, \cdots, h_N^c\} \tag{1}$$

2.2 Supervised Dynamic Local Attention

In this section, we introduce dynamic local attention guided by a supervised segmentation signal to learn latent segment structure in a document. Firstly, we generate the sentence-level attention spans for each sentence by training soft masking (Nguyen et al., 2020), using pointing mechanism (Vinyals et al., 2015) to approximate left and right boundary positions of the mask vector. Given the query Q and key K, where $Q = K = H^c$,

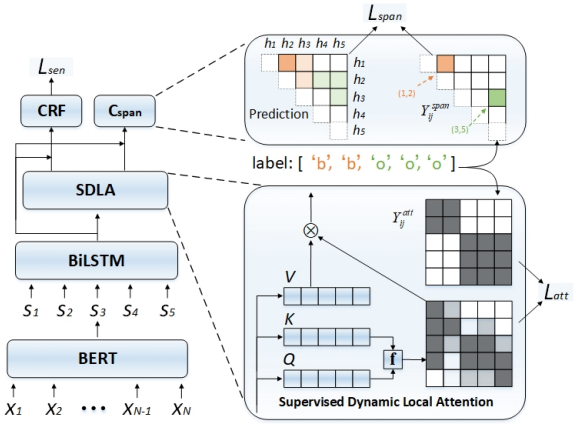

Figure 2: The overview of our model, exemplified by the sample in Figure 1. The labels 'b' and 'o' stand for "background" and "outcome", respectively. C_{span} denotes Auxiliary Span-based Classification Task.

we calculate the left and right boundary matrix $\hat{\phi}_l, \hat{\phi}_r \in R^{N \times N}$ for query Q as follows:

$$\hat{\phi}_l = \mathcal{S}(\frac{Q^T W_L^Q (K W_L^K)^T}{\sqrt{d}} \odot M) \tag{2}$$

$$\hat{\phi}_r = \mathcal{S}(\frac{Q^T W_R^Q (K W_R^K)^T}{\sqrt{d}} \odot M^T) \tag{3}$$

$$M_{ij} = \begin{cases} -\infty, & i < j \\ 1, & i \geq j \end{cases} \tag{4}$$

where $\mathcal{S}$ is the softmax function, $\odot$ is element-wise product, and $W_L^Q, W_L^K, W_R^Q, W_R^Q \in \mathbf{R}^{d \times d}$ are trainable parameters. Eq. (2)-(3) approximate the left and right boundary positions of the mask matrix for the query Q (Each row approximate the mask vector of the entire document corresponding to each sentence in sequence order). Note that we additionally introduce mask matrix M to ensure that the left boundary position l and the right boundary position r generated at position i satisfy this relationship such that $0 \leq l \leq i \leq r \leq N$.

Given the above definitions, the attention span masking matrix M_a can be achieved by compositing the left and right boundary matrix :

$$M_a = (\hat{\phi}_l L_N) \odot (\hat{\phi}_r L_N^T) \tag{5}$$

where $L_N \in \{0, 1\}^{N \times N}$ denotes a unit-value (1) upper-triangular matrix.

Then we combine self-attention with the attention span masking, enabling the model to focus on semantically related sentences around the target

position and eliminate noisy aggregations :

$$A = \frac{(QW^Q)(KW^K)^T}{\sqrt{d}} \odot M_a \quad (6)$$

$$H^{att} = \mathcal{S}(A)(H^c W^H) \quad (7)$$

where W^Q, W^K, W^H are the trainable parameters.

However, in the absence of a supervised process, the dynamic local attention may fail to focus on the corresponding informative sentences of the target, especially for limited data, so we further introduce the segmentation signal to guide the learning of dynamic local attention to capture coherent semantics more accurately. Specifically, we employ binary cross-entropy loss to describe the differences between attention matrix A and segment signal Y^{att}:

$$\mathcal{L}_{att} = BCE(\sigma(A), Y^{att}) \quad (8)$$

$$Y^{att}_{ij} = \begin{cases} 1, & E_{ij} = 1 \\ 0, & else \end{cases} \quad (9)$$

where σ is sigmoid function. $E_{ij} = 1$ denotes i-th sentence and j-th are in the same segment (e.g. (s_1, s_2) and (s_4, s_5) in Figure 1).

Finally, we concatenate H^c and H^{att} as the contextual representations H and add a CRF layer to classify each sentence.

2.3 Auxiliary Span-based Classification Task

Due to the obvious label consistency of sentences within spans, we introduce an additional auxiliary task called span-based classification to improve the performance at the span-level. To this effect, we consider all possible spans of various lengths and propose a tagging scheme for span-based classification. The scheme uses the same labels as sentence-level to represent the label of a span. Firstly, we represent a span from the i-th sentence to the j-th sentence as a vector h_{ij}, which is concatenated by four-vectors similar to Zhao et al. (2020):

$$h_{ij} = \{h_i; h_j; \hat{h}_{i:j}; \varphi(j - i + 1)\} \quad (10)$$

where $\hat{h}_{i:j}$ is the attention output over the final sentence representation H in the span, and $\varphi(j - i + 1)$ is the feature vector encoding the span size.

We employ a cross-entropy category loss for span-based classification:

$$\mathcal{L}_{span} = CE(\hat{Y}^{span}, Y^{span}) \quad (11)$$

$$Y^{span}_{ij} = \begin{cases} label, & F_{ij} = 1 \\ 0, & else \end{cases} \quad (12)$$

where $\hat{Y}^{span}$ is the output probability at span-level, F_{ij} denotes i-th sentence and j-th sentence (i, j satisfy the relationship $i < j$) are in the same segment and i, j is the beginning and end of the segment respectively (e.g. (s_1, s_2) and (s_3, s_5) in Figure 1).

2.4 Objective Function

The overall objective function includes cross-entropy loss $\mathcal{L}_{sen}$, $\mathcal{L}_{span}$ for sentence and span-based classification and supervised attention loss $\mathcal{L}_{att}$:

$$\mathcal{L} = \mathcal{L}_{sen} + \lambda_{att}\mathcal{L}_{att} + \lambda_{span}\mathcal{L}_{span} \quad (13)$$

where $\lambda_{att}, \lambda_{span}$ are the hyperparameters for balancing the strength of $\mathcal{L}_{att}$ and $\mathcal{L}_{span}$.

3 Experiments

3.1 Experimental Setup

Datasets and Baselines To evaluate the effectiveness of our model, we conduct extensive experiments on two standard benchmark datasets from medical scientific abstracts, i.e. NICTA-PIBOSO (Kim et al., 2011) and PubMed 20k RCT (Dernoncourt and Lee, 2017). The detailed description of both datasets can be found in the appendix. We compare our model with three recent strong neural models, i.e., those of Jin and Szolovits (2018), Cohan et al. (2019), Yamada et al. (2020).

Implementation Details We set the size of hidden state to 200 and apply dropout with the probability of 0.5 for BiLSTM. Both hyperparameters λ_{att} and λ_{span} are set to 0.3. The batch size is 30. We use Adam optimizer with learning rate 0.003 and weight decay 0.99 for training. For evaluation, we maximize the score from sentence-level CRF to get the predicted labels of the corresponding se-

Models	Sentence-F1	Span-F1	P_k
NICTA-PIBOSO			
Jin and Szolovits (2018)	82.3	51.1	17.3
Cohan et al. (2019)	83.0	54.3	21.3
Yamada et al. (2020)	84.4	58.7	–
Ours	**86.8**	**62.9**	**12.2**
PubMed 20k RCT			
Jin and Szolovits (2018)	92.8	82.9	5.3
Cohan et al. (2019)	92.9	82.2	5.1
Yamada et al. (2020)	**93.1**	84.3	–
Ours	92.8	**84.5**	**4.1**

Table 1: The results comparison of our model and baselines on two benchmark datasets.

	background	other	intervention	study design	population	outcome
Avg Num. Sent.	2.8	2.6	1.3	1.0	1.1	5.2
Jin and Szolovits (2018)	53.5	34.0	31.7	64.1	70.8	51.4
Cohan et al. (2019)	55.5	41.0	36.9	63.0	69.9	57.4
Yamada et al. (2020)	60.5	**44.8**	34.3	62.4	72.9	64.3
Ours	**60.8**	35.4	**49.0**	**71.4**	**77.6**	**64.4**

Table 2: Average number of sentences in spans and Span-F1 scores for each rhetorical label on NICAT-PIBOSO.

	background	objective	methods	results	conclusions
Avg Num. Sent.	2.6	1.5	4.1	4.2	1.8
Jin and Szolovits (2018)	73.8	73.8	86.7	83.1	90.8
Cohan et al. (2019)	70.6	70.8	86.3	83.9	92.0
Yamada et al. (2020)	**74.7**	73.8	88.5	**85.8**	91.9
Ours	67.1	**74.4**	**89.3**	85.7	**93.2**

Table 3: Average number of sentences in spans and Span-F1 scores for each rhetorical label on PubMed 20k RCT.

quence. Following Yamada et al. (2020), we use Sentence-F1 and Span-F1 as evaluation metrics[1].

3.2 Experimental Results

Tabel 1 report the performance of our approaches against other methods on PubMed 20k RCT and NICTA-PIBOSO, respectively. The results of other methods are obtained from Yamada et al. (2020).

We can observe that our model, whether Sentence-F1 or Span-F1, is significantly better than other methods on NICTA-PIBOS, and we get a result comparable to Yamada et al. (2020) on PubMed 20k RCT. We believe that our model has remarkable performance on NICTA-PIBOS, which has fewer training samples but larger label space, because our model can capture latent segment structure by SDLA component and improve span representations by auxiliary span-based classification.

In addition, table 2 and 3 show the detail results of Span-F1 scores for each rhetorical label. Our model achieves better or similar performance than other baselines, except for "other" on NICAT-PIBOSO and "background" on PubMed 20k RCT. We speculate that the reason is that the sentence semantics corresponding to the "other" label are diverse and not significantly distinguishable from other labels, while the "background" usually appears before the "objective", and the sentence presentations of the two are easily confused.

3.3 Segmentation Performance Evaluation

Specially, if we ignore the rhetorical labels of sentences and only consider the segment boundaries (i.e. binary classification, whether it's a boundary), this can be regarded as text segmentation (Koshorek et al., 2018). We evaluate the segmentation performance of our model using the probabilistic P_k (Beeferman et al., 1999) error score (lower number, the better). The results[2] are shown in the last column of Table 1. Our model consistently outperforms other baselines, suggesting that it also contributes to the text segmentation task.

Ablation Models	Sentence-F1	Span-F1
NICTA-PIBOSO		
Ours	**86.8**	**62.9**
- w/o SDLA	85.1	59.7
- w/o supervised signal	84.9	59.1
- w/o span-based classification	85.6	61.0
PubMed 20k RCT		
Ours	**92.8**	**84.5**
- w/o SDLA	92.3	82.4
- w/o supervised signal	92.6	82.9
- w/o span-based classification	92.6	83.4

Table 4: Ablation study on two datasets.

3.4 Ablation Study

To investigate the effectiveness of the designed components, we conduct an ablation study on the proposed model, and the results are listed in Table 4. With the help of the SDLA component, the performances are improved significantly, and the way we impose the supervised signal to guide the attention proves effective for yielding more true positives. And the auxiliary task of span classification effectively improves Span-F1.

3.5 Attention Visualization and Case Study

As shown in Figure 3, by incorporating supervised signal, the attention focus on a continuous local

[1]Please refer to Yamada et al. (2020) for the detailed calculation way of Sentence-F1 and Span-F1.

[2]Since Yamada et al. (2020) don't release their codes, we are unable to evaluate its P_k performance. The P_k results of other models are obtained by running their codes.

Sentence	Gold	Base	Ours
Tizanidine hydrochloride , an alpha (2) - adrenergic receptor agonist , is a widely used medication for the treatment of muscle spasticity .	B	B	B
Clinical studies have supported its use in the management of spasticity caused by multiple sclerosis (MS) , acquired brain injury or spinal cord injury .	B	B	B
It has also been shown to be clinically effective in the management of pain syndromes , such as : myofascial pain , lower back pain and trigeminal neuralgia .	B	B	B
This review summarizes the recent findings on the clinical application of tizanidine .	O	B	O
Our objective was to review and summarize the medical literature regarding the evidence for the usefulness of tizanidine in the management of spasticity and in pain syndromes such as myofascial pain .	O	B	O
We reviewed the current medical and pharmacology literature through various internet literature searches .	O	B	O
This information was then synthesized and presented in paragraph and table form .	O	O	O

Table 5: Examples of label predictions for NICTA-PIBOSO abstract by BERT+BiLSTM+CRF (Base) and our proposed method (Ours). B and O denote background and other labels respectively.

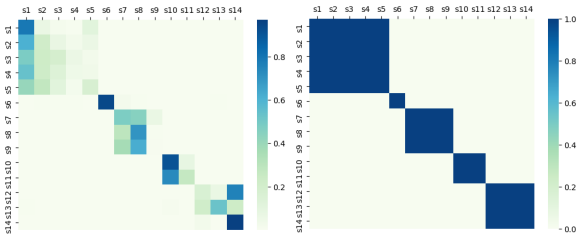

Figure 3: Visualization of attention weights (left) and supervised signal (right). The deeper color means the higher weight.

span around the gold span. The visualization results not only verifies the effectiveness of the supervised signal, but also reveals the interpretability of our proposed SDLA.

Table 5 shows the results of **Base** and **Ours** method for an abstract obtained from NICTA-PIBOSO. Our model correctly identified the boundary between the spans labeled by background (B) and other (O), which shows our model benefit from capturing latent segment structure identifying the more indistinguishable segmentation boundaries.

4 Conclusion

In this paper, we propose a novel model for SSC task, which includes a supervised dynamic local attention to explore the latent segment structure of the document, and an auxiliary task to improve the performance at span-level representations. We demonstrate the effectiveness of our model on two datasets and find that our model also performs well in the text segmentation scenario. In future work, we will consider joint learning sequential sentence classification and text segmentation.

Acknowledgments

We thank the anonymous reviewers for their helpful feedbacks. The work described in this paper was partially funded by the National Natural Science Foundation of China (Grant No. 61502174, and 61872148), the Natural Science Foundation of Guangdong Province (Grant No. 2017A030313355, 2019A1515010768 and 2021A1515011496), the Guangzhou Science and Technology Planning Project (Grant No. 201704030051, and 201902010020), the Key R&D Program of Guangdong Province (No. 2018B010107002) and the Fundamental Research Funds for the Central Universities.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

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.

Daniel Cohen, Liu Yang, and W Bruce Croft. 2018. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1165–1168.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed

200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tracy Edinger, Dina Demner-Fushman, Aaron M Cohen, Steven Bedrick, and William Hersh. 2017. Evaluation of clinical text segmentation to facilitate cohort retrieval. In *AMIA Annual Symposium Proceedings*, volume 2017, page 660. American Medical Informatics Association.

Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. 2014. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of biomedical informatics*, 49:159–170.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

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.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Jimmy Lin, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the hlt-naacl bionlp workshop on linking natural language and biology*, pages 65–72.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2020. Differentiable window for dynamic local attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6589–6599, Online. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2692–2700.

Kosuke Yamada, Tsutomu Hirao, Ryohei Sasano, Koichi Takeda, and Masaaki Nagata. 2020. Sequential span classification with neural semi-Markov CRFs for biomedical abstracts. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 871–877, Online. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2018. Hybrid semi-markov crf for neural sequence labeling. *arXiv preprint arXiv:1805.03838*.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, et al. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248.