Sequential Span Classification with Neural Semi-Markov CRFs for Biomedical Abstracts

Kosuke Yamada[1] **Tsutomu Hirao**[2] **Ryohei Sasano**[1]
Koichi Takeda[1] **Masaaki Nagata**[2]
[1]Graduate School of Informatics, Nagoya University, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
yamada.kosuke@c.mbox.nagoya-u.ac.jp,
{tsutomu.hirao.kp,masaaki.nagata.et}@hco.ntt.co.jp,
{sasano,takedasu}@i.nagoya-u.ac.jp

Abstract

Dividing biomedical abstracts into several segments with rhetorical roles is essential for supporting researchers' information access in the biomedical domain. Conventional methods have regarded the task as a sequence labeling task based on sequential sentence classification, i.e., they assign a rhetorical label to each sentence by considering the context in the abstract. However, these methods have a critical problem: they are prone to mislabel longer continuous sentences with the same rhetorical label. To tackle the problem, we propose sequential span classification that assigns a rhetorical label, not to a single sentence but to a span that consists of continuous sentences. Accordingly, we introduce Neural Semi-Markov Conditional Random Fields to assign the labels to such spans by considering all possible spans of various lengths. Experimental results obtained from PubMed 20k RCT and NICTA-PIBOSO datasets demonstrate that our proposed method achieved the best micro sentence-F_1 score as well as the best micro span-F_1 score.

1 Introduction

Dividing documents into several rhetorical segments is a fundamental task in natural language processing (NLP). For example, abstracts in PubMed, a database of the biomedical literature, can be divided into rhetorical segments such as "Objective", "Methods", "Results", and "Conclusions". Abstracts segmented for each rhetorical role allows us to exploit advanced search. That is, researchers can easily find information by utilizing the structured queries such as "find abstracts that contain 'Covid-19' in 'Objective' and 'Remdesivir' in 'Methods'". Furthermore, the technique can also be used for NLP applications such as academic writing support (Huang and Chen, 2017), scientific trend analysis (Prabhakaran et al., 2016), and question-answering (Guo et al., 2013).

Most previous methods in PubMed have regarded the task as a sequence labeling, namely sequential sentence classification, that assigns rhetorical labels with a `B(egin)/I(nside)` tag set to each sentence while considering the context in the abstract. To this end, some statistical methods with hand-engineered features have been proposed, including Hidden Markov Models (HMMs) (Lin et al., 2006) and Conditional Random Fields (CRFs) (Hirohata et al., 2008; Kim et al., 2011; Hassanzadeh et al., 2014). Recently, with the success of neural network models for NLP tasks, Dernoncourt et al. (2017) and Jin and Szolovits (2018) have employed BiLSTMs to obtain sentence embeddings based on word embeddings and CRFs for assigning labels to the sentences. Cohan et al. (2019) employed a pre-trained language model, SciBERT (Beltagy et al., 2019), which is a variant of BERT (Devlin et al., 2019) trained with scientific papers, to improve the performance of classification without CRFs.

Previous methods cast the segmentation with the labeling as a sentence classification. However, such methods have a critical problem: their performances on longer spans[1] is not so good since they are designed to maximize the prediction of rhetorical roles for a small context.

To tackle the problem, we propose a novel approach, neural sequential span classification, that directly gives the labels for the spans while considering all possible spans of various lengths in the abstract. That is, our method is designed to maximize the performance of classification at the span level rather than the sentence level. Consequently, we introduce Neural Semi-Markov Conditional Random Fields (SCRFs) (Ye and Ling, 2018; Kemos et al.,

[1]In this paper, we call a segment as a "span", which consists of continuous sentences.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 871–877

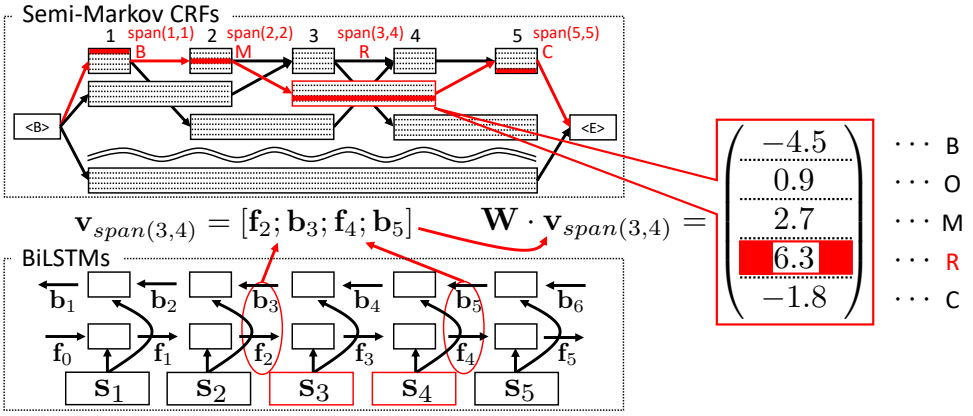

Figure 1: Overview of the neural sequential span classification. In the example, five rhetorical labels, "Background" (B), "Objective" (O), "Methods" (M), "Results" (R), and "Conclusions" (C) can be assigned to spans. The abstract that consists of five sentences is segmented into four spans, span(1,1), span(2,2), span(3,4), and span(5,5), and these spans are labeled as B, M, R, and C, respectively.

2019) to handle spans of the different lengths. To demonstrate the effectiveness of method, we conducted experimental evaluations on two benchmark datasets, PubMed 20k RCT (Dernoncourt and Lee, 2017) and NICTA-PIBOSO (Kim et al., 2011). The results show that our method achieved the best micro sentence-F_1 score of 93.1 and micro span-F_1 score of 84.3 in the PubMed 20k RCT dataset, and the best micro sentence-F_1 score of 84.4 and micro span-F_1 score of 58.7 in the NICTA-PIBOSO dataset.

2 Proposed Method

To perform sequential span classification in an end-to-end manner, we need to represent spans as vectors and handle all possible sequences with various lengths in the abstract. To this end, we introduce BiLSTMs and Semi-Markov CRFs (SCRFs). Figure 1 shows an overview of our method. The BiLSTMs layer generates span vectors from sentence vectors incorporating the context in the abstract, and the SCRFs layer learns the labeling of span sequences by considering all possible sequences of various lengths. The details are described below.

2.1 Span Representation

BiLSTMs have been successfully used to represent spans as vectors in many NLP tasks such as semantic role labeling (Ouchi et al., 2018), syntactic parsing (Stern et al., 2017), and coreference resolution (Lee et al., 2017). BiLSTMs use a forward-LSTM function $\overrightarrow{\text{LSTM}}$ and backward-LSTM function $\overleftarrow{\text{LSTM}}$, where the forward and backward hidden states of the i-th sentence are represented as follows:

$$\mathbf{f}_i = \overrightarrow{\text{LSTM}}(\mathbf{f}_{i-1}, \mathbf{s}_i), \ \mathbf{b}_i = \overleftarrow{\text{LSTM}}(\mathbf{b}_{i+1}, \mathbf{s}_i). \quad (1)$$

Here, $\mathbf{s}_i$ represents the embedding of the i-th sentence. To obtain $\mathbf{s}_i$, we utilize BERT, which has been pre-trained with PubMed (Peng et al., 2019). We insert `[CLS]` tokens at the beginning and `[SEP]` tokens at the end of sentences and then extract vectors corresponding to `[CLS]` tokens in the penultimate layer as sentence vectors. Finally, we represent a span from the i-th sentence to the j-th sentence as a vector, $\mathbf{v}_{span(i,j)}$, which is a concatenation of four vectors as follows:

$$\mathbf{v}_{span(i,j)} = [\mathbf{f}_{i-1}; \mathbf{b}_i; \mathbf{f}_j; \mathbf{b}_{j+1}]. \quad (2)$$

2.2 Neural Semi-Markov CRFs

Neural SCRFs (Ye and Ling, 2018; Kemos et al., 2019) learn parameters to maximize the log-likelihood function, $\sum_{j=1}^{N} \log P(\mathbf{y}_j^* | \mathbf{X}_j)$, where N is the number of training data, $\mathbf{y}_j^*$ is the correctly labeled sequence of spans for the j-th abstract in the training data, and $\mathbf{X}$ is the sequence of sentences in the j-th abstract. The conditional probability, $P(\mathbf{y}|\mathbf{X})$, is obtained by applying the softmax function to the score of a span sequence as follows:

$$P(\mathbf{y}|\mathbf{X}) = \frac{\exp(\text{score}(\mathbf{X}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathbf{Y}} \exp(\text{score}(\mathbf{X}, \mathbf{y}'))}. \quad (3)$$

Here, $\mathbf{Y}$ is the set of all possible labeled span sequences against $\mathbf{X}$. We denote a labeled span sequence as $\mathbf{y}$ and its length as len($\mathbf{y}$). Then, we represent the k-th span as a set of the start sentence index, the end sentence index, and the label,

(i_k, j_k, ℓ_k). The score of the labeled span sequence, $\text{score}(\mathbf{X}, \mathbf{y})$, is defined as follows:

$$\text{score}(\mathbf{X}, \mathbf{y}) = \sum_{k=1}^{\text{len}(\mathbf{y})} e_{(i_k, j_k, \ell_k)} + \sum_{k=0}^{\text{len}(\mathbf{y})} t_{\ell_k, \ell_{k+1}}. \quad (4)$$

The first term of the right-hand side of Equation (4) denotes the sum of the span labeling scores. Here, $e_{(i,j,\ell)}$ is defined as $\mathbf{w}_\ell \cdot \mathbf{v}_{span(i,j)}$. $\mathbf{w}_\ell$ denotes the weight vector for the label ℓ. The second term denotes the score of transition between labels. We assume that no transition occurs between the same rhetorical labels. The weight matrix for the labeling $\mathbf{W}$ and the weight matrix for the transition between labels $\mathbf{T}$ are the parameters, which are optimized by using stochastic gradient descent (SGD). The Viterbi algorithm is utilized to obtain the optimal labeled span sequence.

3 Experiments

3.1 Dataset

We evaluated our method by two standard benchmark datasets, PubMed 20k RCT (Dernoncourt and Lee, 2017) and NICTA-PIBOSO (Kim et al., 2011).

PubMed 20k RCT consists of 200,000 PubMed abstracts on randomized controlled trials annotated with five rhetorical labels, "Background" (B), "Objective" (O), "Methods" (M), "Results" (R), and "Conclusions" (C). PubMed 20k RCT was officially divided into 15,000 documents as the training dataset, 2,500 documents as the development dataset, and 2,500 documents as the test dataset.

NICTA-PIBOSO consists of 1,000 biomedical abstracts with 6 rhetorical labels, "Background", "Other", "Intervention", "Study design", "Population", and "Outcome". Since the dataset is relatively small, we performed 10-fold cross-validation. The ratio of the training dataset, the development dataset, and the test dataset is 8:1:1.

3.2 Compared Methods

To demonstrate the effectiveness of sequential span classification, we compared it with a combination of sequential sentence classification methods, **BiLSTMs+CRFs** as a simple baseline, and two state-of-the-art methods, i.e., those of **Jin and Szolovits (2018)** and **Cohan et al. (2019)**.

As with our method, BiLSTMs+CRFs employ sentence vectors obtained from BERT pre-trained with PubMed (Peng et al., 2019). Thus, the difference between our method and BiLSTMs+CRFs is whether CRFs or SCRFs are used for the sequence labeling.

Jin and Szolovits (2018) is also based on the BiLSTMs+CRFs framework. However, the sentence vectors used as input to the BiLSTMs layers are generated by considering the importance of words by using word-based BiRNN with attention. Cohan et al. (2019) obtain the sentence vectors from SciBERT (Beltagy et al., 2019). Unlike our method, they extract vectors corresponding to tokens `[SEP]`, which are inserted into the sentence boundary, from the top-layer as sentence vectors.

3.3 Model Parameters

We used the batch size of 30, the hidden layer size of 50, 100, or 200, and the learning rate of 0.005, 0.01, 0.02, or 0.05 as hyperparameters. The parameters of all methods are optimized with the training dataset,[2] and the hyperparameters are tuned with the development dataset.[3]

3.4 Evaluation Measures

As evaluation measures, we employ the micro sentence-F_1 score, a de-fact standard evaluation measure to measure the performance of the labeling at the sentence level and the micro span-F_1 score to measure the performance of the labeling at the span level.[4] Sentence-F_1 is defined as a harmonic mean of sentence-precision and sentence-recall based on a perfect match of sentence-by-sentence labels (e.g., "Background", "Method"). However, we believe that sentence-F_1 is not suitable for measuring the performance of the segmentation. For example, when an abstract consists of five sentences with the gold label sequence, 'B-M-M-R-C' and a prediction, 'B-M-R-R-C' are given, sentence-precision and sentence-recall are 4/5 and sentence-F_1 is also 4/5. While the result seems that the prediction performs well, the segmentation of "Method" and "Results" are failed. Thus, we introduced span-F_1 that is defined as a harmonic mean of precision and recall based on a perfect match of span-by-span labels. Span-F_1 of the above example is lower than sentence-F_1; the score is 2/4.

[2] See the Supplemental Materials about the number of parameters, training time, and epochs.

[3] The best model is the hidden layer size of 100 and the learning rate of 0.01.

[4] The hyperparameters are tuned to maximize the sentence-F1 score.

	Sentence-F_1	Span-F_1
Proposed	**93.1**	**84.3**
BiLSTMs+CRFs	91.8	81.2
Jin and Szolovits	92.8	82.9
Cohan et al.	92.9	82.2

Table 1: Micro sentence-F_1 and span-F_1 scores obtained from PubMed 20k RCT.

	Sentence-F_1	Span-F_1
Proposed	**84.4**	**58.7**
BiLSTMs+CRFs	84.1	57.7
Jin and Szolovits	82.3	51.1
Cohan et al.	83.0	54.3

Table 2: Micro sentence-F_1 and span-F_1 scores obtained from NICTA-PIBOSO.

3.5 Results

Tables 1 and 2 show the results for the micro-averaged sentence-F_1 scores and span-F_1 scores against PubMed 20k RCT and NICTA-PIBOSO, respectively.[5] The results of Jin and Szolovits (2018) and Cohan et al. (2019) are obtained by running their codes.[6]

Our method achieved the best scores for both evaluation measures in both datasets. Remarkable differences between our method and the other methods could be observed in span-F_1. In particular, the significant gain of our method's score against BiLSTMs+CRFs, which employs the same sentence vectors as our method, implies that sequential span classification performs better than sequential sentence classification. We observe both sentence- and span-F_1 scores in NICTA-PIBOSO are lower than those in PubMed 20k RCT. We believe that the results are caused by the small number of training data and the large number of rhetorical label sequence types in the training data.[7]

We perform significant tests using the permutation test with Bonferroni correction at significance level=0.05. There were significant differences between our method and BiLSTMs+CRFs, Jin and Szolovits (2018), and Cohan et al. (2019) in span-F_1 of PubMed, between our method and Jin and Szolovits (2018), and Cohan et al. (2019) in span-F_1 of NICTA-PIBOSO. There were no significant differences between our method and baselines in sentence-F_1 scores on both datasets. As we mentioned before, we believe that span-F_1 is more suitable than sentence-F_1 for measuring the performance of the segmentation. Thus, the results demonstrate the effectiveness of our method.

To evaluate the effectiveness of our method in detail, we examined span-F_1 scores for each rhetorical label. The results are shown in Tables 3 and 4. In PubMed 20k RCT, our method achieved the best scores on four rhetorical labels and a comparable score for "Conclusions". In NICTA-PIBOSO, our method achieved the best scores on four rhetorical labels. These results also indicate the effectiveness of sequential span classification. In particular, significant improvements were confirmed for "Background", "Methods", and "Results" in Pubmed 20k RCT and "Background", "Other", and "Outcome" in NICTA-PIBOSO, which contain a larger number of sentences than the other rhetorical labels. This is a significant advantage of sequential span classification over sequential sentence classifications.

Figure 2 shows the results of BiLSTMs+CRFs and our proposed method for an abstract obtained from PubMed 20k RCT. In the abstract, "Results" consists of six sentences. BiLSTMs+CRFs failed the labeling of the last sentence in "Results." As a result, that failed the labeling of the two spans, "Results" and "Conclusions". On the other hand, our method successfully labeled all spans.

4 Conclusions

In this paper, we proposed the neural sequential span classification that directly assigns rhetorical labels to each span in a biomedical abstract by considering all possible spans of various lengths. To perform this classification technique, we introduced neural Semi-Markov CRFs. Evaluation results obtained from PubMed 20k RCT and NICTA-PIBOSO datasets show that our method outperformed state-of-the-art sequential sentence classification methods. In other words, our method achieved the best scores for both micro sentence- and span-F_1 scores. In particular, we found a remarkable improvement in the span-F_1 score. Furthermore, the classification accuracy for long spans, that is, rhetorical labels containing a larger num-

[5] See the Supplemental Materials about validation performance.

[6] Their codes are available at https://github.com/jind11/HSLN-Joint-Sentence-Classification and https://github.com/allenai/sequential_sentence_classification, respectively.

[7] The number of correct rhetorical label sequences of PubMed 20k RCT and NICTA-PIBOSO are 45 and 168, respectively.

	Background	Objective	Methods	Results	Conclusions
# of sentences	2.6	1.5	4.1	4.2	1.8
Proposed	**74.7**	**73.8**	**88.5**	**85.8**	91.9
BiLSTMs+CRFs	70.2	68.6	85.8	83.1	90.1
Jin and Szolovits	73.8	**73.8**	86.7	83.1	90.8
Cohan et al.	70.6	70.8	86.3	83.9	**92.0**

Table 3: Average number of sentences in spans and span-F_1 scores for each rhetorical label in PubMed 20k RCT.

	Background	Other	Intervention	Study design	Population	Outcome
# of sentences	2.8	2.6	1.3	1.0	1.1	5.2
Proposed	**60.5**	**44.8**	34.3	62.4	**72.9**	**64.3**
BiLSTMs+CRFs	57.7	43.5	**38.1**	**64.7**	72.6	63.5
Jin and Szolovits	53.5	34.0	31.7	64.1	70.8	51.4
Cohan et al.	55.5	41.0	36.9	63.0	69.9	57.4

Table 4: Average number of sentences in spans and span-F_1 scores for each rhetorical label in NICTA-PIBOSO.

Sentence	**Gold**	**Base**	**Prop.**
Compare the effect of financial incentives on response to a cancer survivors' postal questionnaire.	O	O	O
Prostate cancer survivors in Ireland, 1.5-18 years after diagnosis, were randomized to the (1) "lottery" arm [a 1 lottery scratch card sent with the questionnaire (n=2,413)] or (2) "prize" arm [entry into a draw on return of a completed questionnaire (n=2,407)].	M	M	M
Impact of interventions on response overall and by survival period ("short term": <5 years after diagnosis; "long term": 5 years after diagnosis) was compared as was cost-effectiveness.	M	M	M
Adjusted response rate was 54.4%.	R	R	R
Response was higher among younger men (P<0.001) and those with earlier stage disease (P=0.002).	R	R	R
A modest 2.6% higher response rate was observed in the lottery compared with the prize arm [multivariate relative risk (RR)=1.06; 95% confidence interval (CI): 1.00, 1.11].	R	R	R
When stratified by survival period , higher response in the lottery arm was only observed among long-term survivors (multivariate RR=1.10; 95% CI: 1.02, 1.19; short-term survivors: RR=1.01; 95% CI: 0.94, 1.09).	R	R	R
Costs per completed questionnaire were 4.54 and 3.57 for the lottery and prize arms, respectively.	R	R	R
Compared with the prize arm, cost per additional questionnaire returned in the lottery arm was 25.65.	R	C	R
Although more expensive, to optimize response to postal questionnaires among cancer survivors, researchers might consider inclusion of a lottery scratch card.	C	C	C

Figure 2: Examples of label predictions for PubMed 20k RCT abstract by BiLSTMS+CRFs (**Base**) and our proposed method (**Prop.**). The PMID of the abstract is 25704725.

ber of sentences, e.g., "Methods", "Results", "Outcome" was improved by our method.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. 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'19)*, pages 3613–3618.

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'19)*, pages 3691–3697.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP'17)*, pages 308–313.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural networks for joint sentence classification in medical paper abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 694–700.

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 (NAACL-HLT'19)*, pages 4171–4186.

Yufan Guo, Ilona Silins, Ulla Stenius, and Anna Korhonen. 2013. Active learning-based information structure analysis of full scientific articles and two applications for biomedical literature review. *Bioinformatics*, 29(11):1440–1447.

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 (IJCNLP'08)*, pages 381—388.

Hen-Hsen Huang and Hsin-Hsi Chen. 2017. DISA: A scientific writing advisor with deep information structure analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'17)*, pages 5229–5231.

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 (EMNLP'18)*, pages 3100–3109.

Apostolos Kemos, Heike Adel, and Hinrich Schütze. 2019. Neural semi-Markov conditional random fields for robust character-based part-of-speech tagging. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*, pages 2736–2743.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics*, 12(2):S5.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, pages 188–197.

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 (LNLBioNLP'06)*, pages 65–72.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*, pages 1630–1642.

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. In *Proceedings of the 18th BioNLP Workshop and Shared Task (BioNLP'19)*, pages 58–65.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16)*, pages 1170–1180.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*, pages 818–827.

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 (ACL'18)*, pages 235–240.

A Number of Parameters

Table 5 shows the number of parameters that are optimized in the training phase. The methods of Jin and Szolovits and Cohan et al. use word vectors as input, while the proposed method and BiLSTMs+CRFs use sentence vectors as input. This is the reason why the number of parameters in the proposed method and BiLSTMs+CRFs is much smaller than that in the other two methods. If we regard the parameters of BERT as a part of the parameters of our proposed methods and BiLSTM+CRFs, the number of parameters in the proposed method, BiLSTMs+CRFs, and Cohan et al. is almost the same.

	# of parameters
Proposed	328,872
BiLSTMs+CRFs	329,069
Jin and Szolovits	10,663,048
Cohan et al.	110,058,391

Table 5: Number of parameters in each method.

B Training Time and Epochs

Table 6 shows the training time and the number of epochs for PubMed 20k RCT. Table 7 shows the average of training time and the average number

of epochs in 10-fold cross-validation for NICTA-PIBOSO. We trained all models on a single Nvidia GeForce GTX 1080 Ti GPU.

	training time	epochs
Proposed	3.24×10^5	60
BiLSTMs+CRFs	5.40×10^3	30
Jin and Szolovits	1.35×10^5	90
Cohan et al.	1.84×10^5	2

Table 6: Training time (seconds) and the number of epochs in the PubMed 20k RCT development dataset.

	training time	epochs
Proposed	4.15×10^4	98.7
BiLSTMs+CRFs	1.94×10^2	19.4
Jin and Szolovits	2.14×10^3	11.9
Cohan et al.	4.92×10^2	4.1

Table 7: Training time (seconds) and the number of epochs in the NICTA-PIBOSO development dataset.

C Validation Performance

Tables 8 and 9 show the validation performance on PubMed 20k RCT and NICTA-PIBOSO development datasets, respectively.

	sentence-F_1	span-F_1
Proposed	93.2	83.5
BiLSTMs+CRFs	92.3	82.0
Jin and Szolovits	93.2	83.6
Cohan et al.	93.1	82.9

Table 8: Validation performance on the PubMed 20k RCT development dataset.

	sentence-F_1	span-F_1
Proposed	85.7	62.1
BiLSTMs+CRFs	85.8	62.5
Jin and Szolovits	82.4	53.3
Cohan et al.	84.3	57.2

Table 9: Validation performance on the NICTA-PIBOSO development dataset.