Context-Aware Sentence Classification in Evidence-Based Medicine

Biaoyan Fang* **Fajri Koto***
School of Computing and Information Systems
The University of Melbourne
{biaoyanf,ffajri}@student.unimelb.edu.au

Abstract

In this paper, we show the effectiveness of before– and after–sentences as additional context for sentence classification in evidence-based medicine. Although pre-trained language models encode contextualized representation, we found that the additional contexts improve sentence classification in terms of ROC (micro) score in the ALTA 2022 shared task. Additionally, averaging the probability of top model predictions boosts the performance, and our results for both public and private test sets officially claim the first rank of the ALTA 2022 shared task.

1 Introduction

Integrating individual clinical expertise and external medicine literature (also known as evidence-based medicine) is the best practice to give care to patients (Sackett et al., 1996; Koto and Fang, 2021). However, obtaining relevant medical literature requires in-depth expertise and can be time-consuming due to the large availability of texts.

A search engine is one of the ways to assist evidence-based medicine, and categorizing sentences in medicine literature based on PICO framework (Kim et al., 2011) can improve the search effectiveness (Amini et al., 2012). PICO mainly consists of four labels: `Population` (P) (i.e. participants in a study); `Intervention` (I); `Comparison` (C) (if appropriate); and `Outcome` (O) (of an Intervention), and can be extended to classes `Background` (B), `Study Design` (S), and `Other` (O) (for sentences that have no relevant content) (Lui, 2012; Kim et al., 2011). The ALTA 2022 shared task uses PIBOSO classes by Kim et al. (2011) and discards `Comparison` (C).

In previous work, Lui (2012) utilized lexical features (e.g. bag-of-words and part-of-speech) and structural features (e.g. sentence length, sentence heading), and fed them to Naive Bayes, SVM, and logistic regression. By stacking the aforementioned features, Lui (2012) demonstrated the effectiveness of logistic regression for this task.

Our work revisits the PIBOSO-based sentence classification task using current state-of-the-art NLP systems (i.e. pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020; Koto et al., 2020)). Similar to Lui (2012), we also use context sentences (i.e. before– and after–sentences) but structure the input to retain the original sequence. Lui (2012) simply used bag-of-words and part-of-speech thus discarding the original sequence information in their features.

We perform context-aware classification using different pre-trained language models including domain-general (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020) and domain-specific models (Gururangan et al., 2020; Gu et al., 2020; Koto et al., 2021; Fang et al., 2021) with two strategies in the classification layer: (1) single embedding, and (2) average pooling. We showcase that both strategies are competitive and significantly better than heuristic *n*-gram features (Lui, 2012). We also show that the ensemble method (Koto and Fang, 2021) by averaging probability prediction of top models improves the ROC (micro) scores, and set our submission in this shared task as the winner.

2 Dataset

The ALTA 2022 shared task adopts the data of Kim et al. (2011). In total, there are 9,244 sentences which are split by the shared-task organizers into 8,216/459/569 for training, public test, and private test sets, respectively. Only labels for training data are available, and for conducting experi-

* equal contribution

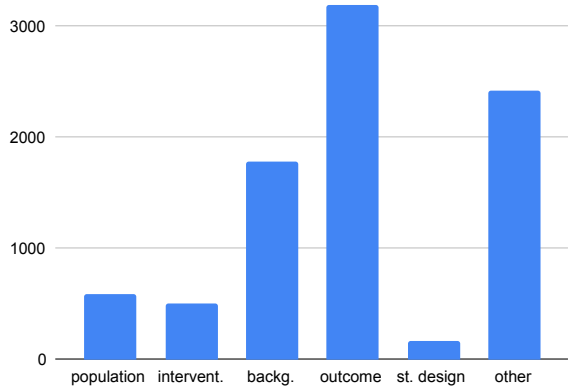

Figure 1: Label distribution of training data.

#document	700
#sentence per document	11.7 ± 6.1
#word per document	210.6 ± 89.1
#word per sentence	17.9 ± 11.2

Table 1: Overall statistics of training data.

ments we randomly sample 768 instances of original training data as the development set and use the remaining for training. The data split ensures that each sentence of a document is in the same set. Please note that we refer `val2022.csv` and `test2022.csv` to the public and private test sets, respectively.

Table 1 shows overall statistics of the training data which consists of 700 documents with 11.7 sentences per document on average. The total number of words per document is 210.6, and each sentence has 17.9 words. There is an imbalanced distribution over the PIBOSO label as described in Figure 1 where `Outcome` is the majority and `Study Design` is the minority class. Please note that this task is a multilabel classification task where one text might consist of more than one label. Further statistics and details regarding the rules of the ALTA 2022 shared task will be described separately by the organizers, and appear alongside this paper.

3 Methodology

In Figure 2, we describe two different approaches for incorporating contextual information: (1) average pooling, and (2) single embedding. Both ways utilize structured input where we added special tokens `<nt>` and `<t>` at the beginning of each non-target (context) and target (main) sentence, respectively. We feed this structured text to pre-trained language models and then process the outputs in two aforementioned ways. Specifically, for average pooling, we first use a masking trick to obtain main sentence embedding and context sentence embedding through averaging. We concatenate the two embeddings (the red and green boxes in Figure 2) prior to the classification layer. For the latter, we merely use the corresponding output of token `<t>` embedding for classification. We argue that the attention mechanism in the transformer (Vaswani et al., 2017) is contextualized to all input tokens, thus encouraging us to test this simpler method.

Our experiments consider domain-general and domain-specific pre-trained language models. It has been shown by previous works (Gururangan et al., 2020; Gu et al., 2020; Koto et al., 2021; Fang et al., 2021, 2022) that domain-general language models are suboptimal for specific domains, and one way to handle this is to use domain-adaptive pre-trained models. In this experiment, we use three domain-general models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), and two domain-specific models: BioBERT from Microsoft (Gu et al., 2020) and DMIS Lab (Lee et al., 2020).[1]

Additionally, we extend the experiments using ensemble learning by averaging probability prediction of top-k models. In similar work, Koto and Fang (2021) has demonstrated the efficacy of ensemble learning in evidence-based medicine-related tasks. The ensemble method is better than a single model since it is capable to enhance model robustness on variance and uncertainty.

4 Experiments

4.1 Settings

As stated in Section 3, we use the huggingface Pytorch framework (Wolf et al., 2020) and select 5 models: 1) BERT,[2] 2) RoBERTa,[3] 3) ELECTRA,[4] 4) BioBERT (Microsoft),[5] and 5) BioBERT (DMIS Lab)[6] for our experiments. Each model is finetuned for 50 epochs with a batch size of 48, a learning rate of 1e-5, and a dropout of 0.5. We consider two settings: (1) without context, i.e. not using any

[1] All models can be accessed in `https://huggingface.co/`

[2] `bert-base-uncased`

[3] `roberta-base`

[4] `google/electra-base-discriminator`

[5] `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`

[6] `dmis-lab/biobert-base-cased-v1.1`

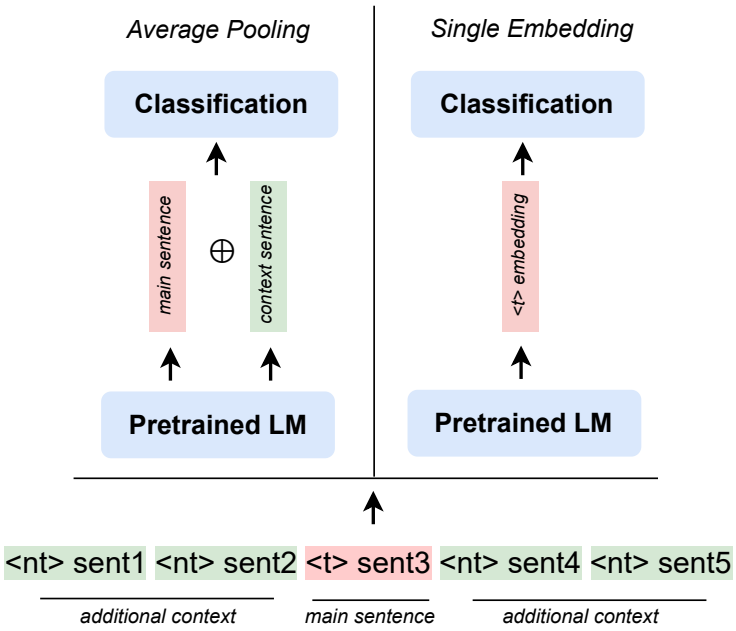

Figure 2: Illustration of our context-aware classification model. <t> and <nt> are special tokens that differentiate target and non-target (context) sentences in the input.

additional context, and (2) with context, i.e. using 4 sentences (2 before– and after–sentences) as the additional context. Models that achieve the best performance on our development set are used.

For evaluation, we report on ROC (micro), following the ALTA 2022 shared task description.

4.2 Results

We tuned our model hyperparameters based on the development set discussed in Section 2, and evaluate them on public and private test sets. Since participants can use up to 100 submissions for the public test set, we use it to pick our best model and predict the private test set. Overall, we found similar results on both public and private test sets, where the context-aware domain-specific model performs best. In this section, we report the results of the private test set. Results for the public test set can be found in Appendix A.

Table 2 shows ROC (micro) scores of all models over the private test set, with and without context. First, consistent with previous works (Devlin et al., 2019; Koto et al., 2020) that pre-trained language models significantly outperform conventional machine learning methods (i.e. Naive Bayes, Logistic Regression, and SVM), with SVM achieves the ROC (micro) score of 91.7 (4 points lower than BERT). Next, we found that the simple single embedding method tends to result in better ROC (micro) scores than the average pooling, with and without contexts. One possible reason is that average pooling on the sentences might introduce unwanted noise, resulting in lower-performance models. The best individual performance is obtained by BioBERT (Microsoft), with 96.6 and 95.6 ROC (micro) scores under single embedding and average pooling approaches, respectively, indicating the benefits of using domain-specific pre-trained language models for this classification task, thus consistent with previous works (Gururangan et al., 2020; Gu et al., 2020; Koto et al., 2021; Fang et al., 2021).

Also, as stated in Section 3, we explore the importance of additional context, i.e. before– and after–sentences for this task. Table 2 shows consistent improvements of pre-trained language models when incorporating additional context, with a maximum gain achieved by BERT (with average pooling) with 3 absolute ROC (micro) scores. We argue that the improvements might come from a better understanding of the target sentence when additional contexts are provided.

Furthermore, inspired by Koto and Fang (2021), we experimented with the ensemble method to improve model robustness and mitigate the performance variance. Specifically, we ensemble top-k ($k = 3, 4, 5$) models under each setting. Results in Table 2 show that ensemble methods achieve strong performance across different settings, outperforming single pre-trained language models. For better

Model	Without Context		With Context	
	Single Emb.	Ave. Pooling	Single Emb.	Ave. Pooling
Baselines				
Naive Bayes	85.9		–	
Logistic Regression	84.2		–	
SVM	91.7		–	
Pre-trained language models				
BERT	95.4	94.1	97.0	96.7
RoBERTa	96.1	94.9	97.6	97.9
ELECTRA	96.3	95.2	97.6	97.5
BioBERT (Microsoft)	96.6	95.6	97.7	96.2
BioBERT (DMIS Lab)	96.2	95.5	97.3	95.8
Ensemble – averaging Top-k models				
Ensemble (Top-3)	97.0	96.9	98.0	98.1
Ensemble (Top-4)	97.0	96.9	98.0	98.0
Ensemble (Top-5)	97.0	96.7	98.0	98.3
Ensemble – further averaging the Ensemble (Top-k) models of Single Embed. and Ave. Pooling				
Combine of Ensemble (Top-3)	97.3		**98.7**	
Combine of Ensemble (Top-4)	97.3		98.6	
Combine of Ensemble (Top-5)	97.2		98.5	

Table 2: ROC (micro) scores over private test set.

utilization of contextual information, we further average the ensemble top-k models from single embedding and average pooling approaches, achieving further improvements across ensemble top-k models. The best performance, 98.7 ROC (micro) score, is achieved when averaging two ensemble top-3 models and used as our final result.

5 Conclusion

In this paper, we propose a context-aware multi-label sentence classifier in evidence-based medicine. We show the effectiveness of using the additional context, i.e. before– and after–sentences in pre-trained language models, by considering two incorporation approaches, single embedding, and average pooling, which capture different perspectives of additional context. The utilization of the ensemble method further shows the benefits of combining single embedding and average pooling models, achieving the best performance in the ALTA 2022 shared task.

Acknowledgments

In this work, Biaoyan Fang is supported by a graduate research scholarship from the Melbourne School of Engineering, while Fajri Koto is supported by the Australia Awards Scholarship (AAS), funded by the Department of Foreign Affairs and Trade (DFAT), Australia.

References

Iman Amini, David Martinez, and Diego Molla. 2012. Overview of the ALTA 2012 shared task. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 124–129, Dunedin, New Zealand.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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.

Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3481–3495,

Dublin, Ireland. Association for Computational Linguistics.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1362–1375, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. 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.

Fajri Koto and Biaoyan Fang. 2021. Handling variance of pretrained language models in grading evidence in the medical literature. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 218–223, Online. Australasian Language Technology Association.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Lui. 2012. Feature stacking for sentence classification in evidence-based medicine. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 134–138, Dunedin, New Zealand.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Results on Public Test Set

Table 3 shows results across all models on the public test set.

Model	Without Context		With Context	
	Single Emb.	**Ave. Pooling**	**Single Emb.**	**Ave. Pooling**
Baselines				
Naive Bayes		90.9		–
Logistic Regression		86.2		–
SVM		91.5		–
Pre-trained language models				
BERT	96.2	94.6	97.2	97.2
RoBERTa	96.2	95.1	97.5	97.4
ELECTRA	96.1	95.4	97.1	97.8
BioBERT (Microsoft)	96.8	96.1	97.3	97.3
BioBERT (DMIS Lab)	96.0	94.5	97.1	96.5
Ensemble – averaging Top-k models				
Ensemble (Top-3)	96.7	96.7	97.6	98.2
Ensemble (Top-4)	96.8	96.5	97.7	98.1
Ensemble (Top-5)	96.8	96.7	97.8	98.1
Ensemble – further averaging the Ensemble (Top-k) models of Single Embed. and Ave. Pooling				
Combine of Ensemble (Top-3)		97.0		**98.4**
Combine of Ensemble (Top-4)		97.0		98.3
Combine of Ensemble (Top-5)		97.1		98.3

Table 3: ROC (micro) scores over public test set.