Handling Variance of Pretrained Language Models in Grading Evidence in the Medical Literature

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

Abstract

In this paper, we investigate the utility of modern pretrained language models for the evidence grading system in the medical literature based on the ALTA 2021 shared task. We benchmark 1) domain-specific models that are optimized for medical literature and 2) domain-generic models with rich latent discourse representation (i.e. ELECTRA, RoBERTa). Our empirical experiments reveal that these modern pretrained language models suffer from high variance, and the ensemble method can improve the model performance. We found that ELECTRA performs best with an accuracy of 53.6% on the test set, outperforming domain-specific models.[1]

1 Background

Evidence-Based Medicine (EBM) is an approach by health practitioners to integrate individual clinical expertise and external evidence from medical literatures in making decisions about the care of patients (Sackett et al., 1996). In practice, understanding the current best evidence from the literature minimizes the unexpected risk of outdated treatments that can be detrimental to patients.

Strength of Recommendation Taxonomy (SORT) (Ebell et al., 2004) is one of the standard scale systems for grading evidence in medical literature and it has been used to assist the EBM approach. SORT groups a medical literature into one of three classes: **A** (*consistent* and *good-quality patient-oriented evidence*), **B** (*inconsistent* or *limited-quality patient-oriented evidence*) and **C** (*other evidence*, such as consensus guidelines, usual practice and opinion). While obtaining these grades on a wide-scale is expensive and requires in-depth medical expertise, previous works (Sarker et al., 2015) have attempted to automate the process by modelling the grading system with *n*-gram language model via SVM (Molla and Sarker, 2011) and ensemble method (Gyawali et al., 2012).

In this work, we focus on investigating the utility of various modern pretrained language models for modelling the evidence grading system in the medical literature. Although transformer (Vaswani et al., 2017) and pretrained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) have achieved impressive performance across various NLP tasks (Wang et al., 2018; Wang et al., 2019) and languages (Koto et al., 2020; Martin et al., 2020), we hypothesize that such evidence grading task is still challenging because of three reasons. First, in-depth medical expertise and knowledge are not always present in the language models. Second, it is very likely that machine learning models suffer from high variance as disagreement in assessing scientific literature is natural, even among the experts. Lastly, obtaining high-quality training data for this task is difficult, and the large transformer-based models potentially suffer from overfitting if the available data is limited.

To address the aforementioned challenges, we use three main strategies. First, we fine-tune domain-specific pretrained models (Gu et al., 2020) that are optimized for medical literature. Previous works (Gururangan et al., 2020; Gu et al., 2020; Alsentzer et al., 2019; Fang et al., 2021; Koto et al., 2021) have shown that such models contain domain-specific knowledge that can boost system performance. Second, we argue that discourse is prominent for this task because each of three SORT classes might have different document structure. For instance, patient-oriented literature and consensus guidelines potentially are written differently in

*equal contribution

[1] Our best result with ELECTRA (large) and ELECTRA (base) put us in the first and second rank on the leaderboard, respectively.

00667 A 10796398 11508437
00668 A 9036306
00669 C 7391096 11204962 7790481 6863528
00670 B 9569395 12069675
00671 B 11083602 10875559 15283004

Figure 1: Sample training data from ALTA 2021 shared task.

	Train	Dev	Test
Evidences	677	178	183
in A	212	48	-
in B	311	80	-
in C	154	50	-
Ave. resources per evidence	2.4	2.5	2.3
Ave. words per abstract	269.9	262.6	274.1
Ave. words per evidence	655.9	653.7	643.9

Table 1: Overall statistics of the ALTA 2021 shared task dataset. Evidence classes in test dataset are withheld by the organizer. "Ave. resources per evidence" means the average number of XML files the evidence has. "Ave. words per abstract" means the average number of words per single abstract. "Ave. words per evidence" means the average number of words per evidence, including journal name, title and abstract.

terms of flow and discourse. In this work, rather than employing a complicated discourse parser (Yu et al., 2018; Koto et al., 2019, 2021), we rely on modern pretrained language models such as ELECTRA (Clark et al., 2020) that contains a rich latent discourse representation (Koto et al., 2021). Lastly, similar to Gyawali et al. (2012), we also perform ensemble learning to tackle the high variance issue of models.

2 Dataset

We conduct our experiments based on the ALTA 2021 shared task[2] which aims to automatically grade evidence in the medical literature. The grading system follows the SORT framework (Ebell et al., 2004) with three classes: **A** (Strong), **B** (Moderate) and **C** (Weak).

As shown in Figure 1 each line in the training data is a single piece of evidence and consists of an ID, a SORT grade, and a list of resource/publication ID(s) from PubMed.[3] Each publication ID is mapped to an XML file containing bibliographic information (e.g. title, author, affiliation, etc.), abstract, and some meta-data such as type and status of the publication.

In Table 1, we present overall statistics of the train, development and test sets. First, nearly 45% of the train and development data are classified as class B. We also found there is no significant difference in terms of the number of resources and words between each subset.

3 Proposed Methods

Figure 2 describes the best model that we submit to ALTA 2021 shared task. We use filtered ensemble method over 3 domain-specific pretrained language models: 1) Biomed BERT (Gu et al., 2020), 2) Biomed RoBERTa (Gururangan et al., 2020) and 3) Biomed RoBERTa that is further pretrained with the training set for 400 epochs, denoted as Task Adaptive Pretraining (TAPT) model; and 3 domain-generic pretrained language models: 1) RoBERTa (Liu et al., 2019), 2) ELECTRA, and 3) ELECTRA (large) (Clark et al., 2020). The selection of RoBERTa and ELECTRA is based on their rich latent discourse representation as reported by Koto et al. (2021).

Given a list of resources or publications $R = \{r_1, r_2, .., r_n\}$ for evidence x, we construct an input sequence as follows. First, each resource r_i consists of journal name j_i, title t_i, and abstract a_i. We form an input sequence x as the concatenation of all texts $j_1 \oplus t_1 \oplus a_1 \oplus ... \oplus j_n \oplus t_n \oplus a_n$. We truncate a resource r_i if the tokens are more than 250, and set the maximum length of the input x to be 512.

To understand the variance of pretrained language models in this task, we fine-tune each model with 100 different random seeds. For ensemble learning, we first select models with accuracy more than hyper-parameter α (values range between 0 and 1) and apply two types of voting mechanism to aggregate the prediction: 1) simple voting based on majority classes, and 2) filtered voting. For the second approach, if the selected n models have an even class distribution, we set class B as the prediction, otherwise normal majority voting is applied. Mathematically, this even prediction is determined based on a threshold β as follows:

$$\frac{1}{3}(|y_A - y_B| + |y_A - y_C| + |y_B - y_C|) \leq \beta$$

where y_A, y_B, y_C are the occurrence of class A, B, and C in n models prediction, respectively (meaning $y_A + y_B + y_C = n$), and $|y_A - y_B|$ indicates the

[2] https://www.alta.asn.au/events/sharedtask2021/index.html

[3] https://pubmed.ncbi.nlm.nih.gov/

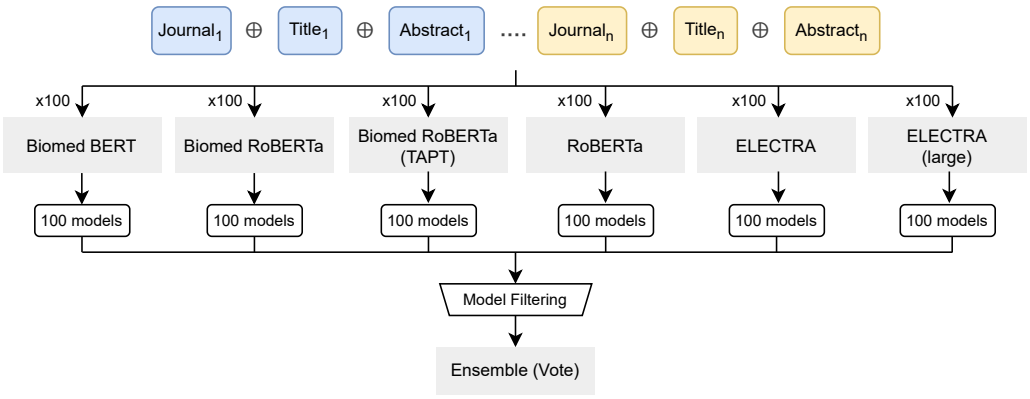

Figure 2: Filtered ensemble model used in this task.

absolute difference of class A and B occurrence. β is a hyper-parameter with values ranging between 0 and n, and $\beta < 0$ means normal majority voting is applied. All parameters (including α and β) are tuned based on the development set.

4 Experiments

4.1 Set-up

We use the huggingface Pytorch framework (Wolf et al., 2020) for the experiments.[4] In total, there are 6 models: 1) Biomed BERT,[5] 2) Biomed RoBERTa,[6], 3) Biomed RoBERTa (TAPT), 4) RoBERTa,[7] 5) ELECTRA,[8] 6) ELECTRA (large).[9] Each model is fine-tuned for 20 epochs with a batch size of 10, warm-up of 10% of the total steps, learning rate of 5e–5, Adam optimizer with epsilon of 1e–8, and early stopping with patience of 5.

In this work, accuracy is used as the primary evaluation metric, following ALTA 2021 shared task description.

4.2 Results over Development Set

In Table 2, we report the aggregate score (`mean`, `max`, `min`, `std`) of 100 runs of each models. First, we observe that Biomed RoBERTa has the highest average performance of 59.5, but only 0.3 higher than ELECTRA. In fact, Domain-generic models such as RoBERTa and ELECTRA outperform Biomed BERT and Biomed RoBERTa (TAPT), despite their domain/task-adaptive pretraning. We also found that even with 100 different random seeds, all models still have relatively high variance (`std`) with more than 2 points. ELECTRA (large) suffers worst from this issue, compared to the other models.

Model	Accuracy			
	Mean	Max	Min	Std
Biomed BERT	58.7	66.9	52.8	2.9
Biomed RoBERTa	**59.5**	67.4	55.1	2.5
Biomed RoBERTa (TAPT)	58.3	65.7	52.8	2.6
RoBERTa	59.1	64.6	53.9	2.2
ELECTRA	59.2	65.7	44.9	3.6
ELECTRA (large)	53.3	64.6	44.9	6.7

Table 2: Experiment results on development set over 100 different random seeds.

In Table 3, we describe the main experiment results. For baselines, we run unigram and bigram representation with Naive Bayes and Logistic Regression, and found the results are less optimal. For the ensemble method, we perform grid search over $\alpha \in \{0.60, 0.61, 0.62, 0.63, 0.64, 0.65\}$ and $\beta \in \{-1, 0, .., n\}$. n is number of models after filtered by parameter α. Ensemble results presented in Table 3 use the best combinations of α and β.

First, we perform ensemble method with all 500 "base" models from Table 2, and obtain accuracy of 69.7, 2 points higher than the best Biomed RoBERTa model (`max` in Table 2). 8 selected models after filtering with α are 2 Biomed RoBERTa, 2 Biomed RoBERTA (TAPT), 2 Biomed BERT, and 2 ELECTRA. In the next results, we also perform a grid search for each 6 pretrained language models (each initially has 100 models), and found that ELECTRA performs best with an accuracy of 70.2, outperforming all domain-specific models.

Another thing to note is that parameter β or fil-

[4] https://huggingface.co/
[5] microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
[6] allenai/biomed_roberta_base
[7] roberta-base
[8] google/electra-base-discriminator
[9] google/electra-large-discriminator

Model	Hyper-parameters		Filtered models (n)	Acc.
	α	β		
Baseline				
Naive Bayes (unigram+bigram)	–	–	–	46.1
Logistic Regression (unigram+bigram)	–	–	–	51.1
Ensemble method				
All 500 "base" models	0.65	$\{-1, 0, 1\}$	8	69.7
Biomed BERT	0.62	$\{-1, 0, 1, 2, 3\}$	11	68.5
Biomed RoBERTa	0.63	2	7	67.4
Biomed RoBERTa (TAPT)	0.62	4	11	66.3
RoBERTa	0.64	$\{-1, 0, 1\}$	3	67.9
ELECTRA	0.63	$\{-1, 0, 1\}$	6	**70.2**
ELECTRA (large)	0.61	$\{-1, 0, 1, 2, 3, 4, 5\}$	18	67.4

Table 3: Results of baseline vs. ensemble methods on the development set. Parameter α and β are selected based on the grid search.

Model	Accuracy	
	Dev	Test
All 500 "base" models	69.7	49.7
ELECTRA	70.2	50.2
ELECTRA (large)	67.4	**53.6**

Table 4: Results of selected model (for shared task submission) on the development and test set.

tered voting mechanism is not significant except for Biomed RoBERTa. From Table 3 we can see that the optimal combinations of α and β for 5 ensemble models have $\beta = -1$, which indicates that the standard majority voting solely can yield the optimal result.

4.3 Results over Test Set

We pick the three best models for ALTA 2021 shared task submission as shown in Table 4. These models are the ensemble methods from Table 3: 1) All 500 "base" models, 2) ELECTRA, and 3) ELECTRA (large). We observe that the gap between development and test set is high, roughly 20 points, which can be due to overfitting problems and small training sets. The best models on the test set are ELECTRA and ELECTRA (large) with the accuracies of 50.2 and 53.6, respectively. Our best result with ELECTRA (large) put us in the first rank on the leaderboard.[10]

[10]The committee limits three submissions for each team. At the end of the competition, ELECTRA result with accuracy 50.2 is picked and put us in the second rank.

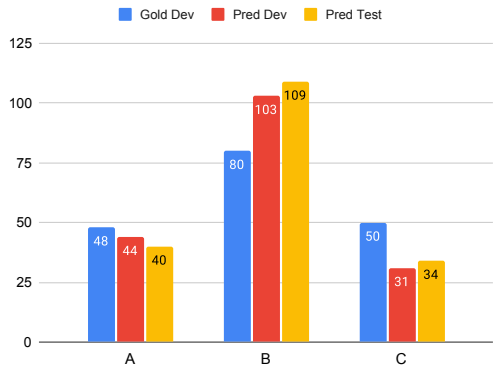

Figure 3: Label distributions on development and test set using ELECTRA (large).

5 Discussions and Conclusion

Figure 3 describes label distributions on development and test sets using our best model, ELECTRA (large). First, we found that the model tends to predict class B on the development, with a disparity of +23 instances with the gold label B. In contrast, the model only classifies 31 instances as class C, despite being there 50 gold labels C. Lastly, our final prediction in the test sets has a ratio of 40:109:34 of class A:B:C, respectively, and the graph in Figure 3 describes a similar shape with the development set prediction.

In conclusion, we have shown in this experiment that grading evidence in the medical literature is a challenging task, and modern pretrained language models suffer from high-variance issues. Interestingly, we found that ELECTRA, the domain-general models outperform domain-specific models through ensemble methods. We argue that this is

because discourse is one of the relevant features for this task. This is in line with Koto et al. (2021) that has shown that the last layer of ELECTRA contains the richest latent discourse representation, compared to BERT, RoBERTa, ALBERT (Lan et al., 2019), GPT2 (Radford et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2019).

Acknowledgments

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

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

Mark H Ebell, Jay Siwek, Barry D Weiss, Steven H Woolf, Jeffrey Susman, Bernard Ewigman, and Marjorie Bowman. 2004. Strength of recommendation taxonomy (sort): a patient-centered approach to grading evidence in the medical literature. *The Journal of the American Board of Family Practice*, 17(1):59–67.

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.

Binod Gyawali, Thamar Solorio, and Yassine Benajiba. 2012. Grading the quality of medical evidence. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 176–184, Montréal, Canada. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2019. Improved document modelling with a neural discourse parser. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 67–76, Sydney, Australia. Australasian Language Technology Association.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian twitter with effective domain-specific vocabulary initialization. *arXiv preprint arXiv:2109.04607*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. 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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer.

2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Diego Molla and Abeed Sarker. 2011. Automatic grading of evidence: the 2011 ALTA shared task. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 4–8, Canberra, Australia.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2015. Automatic evidence quality prediction to support evidence-based decision making. *Artificial intelligence in medicine*, 64(2):89–103.

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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *33rd Annual Conference on Neural Information Processing Systems, NeurIPS 2019*, volume 32, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1097–1102, Brussels, Belgium. Association for Computational Linguistics.