IMPROVED CLAIM DETECTION IN BIOMEDICAL PUBLICATION ABSTRACTS USING TRANSFORMERS

Kelianne Heinz
W266: Natural Language Processing
UC Berkeley School of Information
kelheinz@berkeley.ed

ABSTRACT

Detecting novel claims in scientific publication abstracts is an area of research that would allow for more efficient methods for learning and understanding new developments from the large body of papers that are published each year. In this paper we show that pre-trained domain specific BERT models are able to detect these novel claim sentences without additional abstract context, improving recall and F1 scores over previous work. We show that hidden state embeddings from BERT models can be used as inputs to CRF models, improving precision, recall, and F1 scores over previous work. We examine positional bias-based error patterns, replicating patterns shown in previous work using CRF models and reducing those patterns using position agnostic models. We show that BERT models can have positional bias without explicit positional inputs and explore potential training methods that may impact the degree of positional bias.

INTRODUCTION

Detecting novel claims in scientific publication abstracts is an important field of research. With a huge number of papers published yearly, the ability to extract the key sentences that report state of the art results, propose novel solutions, or describe new findings would allow for automation and streamlining of the analysis of new scientific knowledge.

BACKGROUND

The annotated biomedical abstract claims dataset used for this project was originally developed by Achakulvisut et. al [1].

Their previous work on this dataset used transfer learning in combination with a bi-LSTM CRF model architecture. Like most other work exploring claim extraction from abstracts [2], this work made claim predictions based on the full abstract, emphasizing the importance of understanding the overall structure of the abstract and how different types of sentences typically occur in sequence, a concept known as discourse [9].

Two patterns of errors in model performance were reported, both tied to positional bias. First, the model was prone to false positives for non-claim sentences at the end of abstracts, lowering model precision for end of abstract sentences. Secondly, the model was prone to false negatives for claim sentences in the middle of abstracts, lowering model recall for sentences not at the end of the abstract.

METHODS

DATA

Two datasets were used in this project. The primary dataset was used for the claim detection task. The secondary dataset was used to test transfer learning.

PRIMARY DATASET

The primary dataset used for this project comes from Achakulvisut et. al [1]. It consists of 1,500 biomedical abstracts, expertly annotated by domain specialists with claim/ non-claim labels for each sentence. In total, the dataset contains 11,702 total sentences, 2,276 of which are labeled as claim sentences. For the purposes of comparison to the original work, the same 50/25/25 split of training/validation/testing examples was used, resulting in 750 training examples, and 375 testing and validation examples each. In this dataset, over half of the claim sentences are the last sentence in the abstract.

Secondary Dataset

The secondary dataset used for this project is the PubMed 20k RCT dataset [7], which was used for fine-tune the BERT model weights prior to training with the primary dataset. The dataset consists of 20,000 abstracts from PubMed, though out of consideration for computational limitations a random subset of 50% of the dataset was used in fine-tuning. Each numeric value in the dataset has been replaced with a @ symbol. Sentences in each abstract are labelled with one discourse classification from the classes (Objective, Introduction, Methods, Discussions, Conclusion). In order to match the primary task of binary classification more closely, the labels were converted to a binary indicator of conclusion or non-conclusion sentence.

Task

This work includes two versions of the claim detection task. The first task was to predict the claim or non-claim label for individual sentences. The second task was to predict a sequence of claim and non-claim labels for the sequence of sentences in an abstract.

Evaluation Metrics

The performance of the classification models was evaluated on precision, recall, and F1 score. Accuracy is not reported because the dataset is highly imbalanced towards non-claim sentences. Additionally, the precision and recall for models were calculated conditional on sentence position within each abstract, to evaluate model performance against the specific patterns of errors reported in the literature.

Baseline Model

In addition to comparison against reported literature performance values, models were compared to a baseline model that predicted each abstracts’ last sentence as a claim, and all other positions as non-claims.

BERT Model

BERT architecture was tested to determine if claims could be identified without contextual information from the rest of the abstract. This was also explored as a way to eliminate positional bias by performing single-sentence classification with no explicit positional information or information about the rest of each abstract.

BERT Architecture

Pre-trained weights for the BERT model were downloaded from `BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext`. Previous work has shown that such a domain specific pre-trained BERT model can improve final model performance [5]. Additionally, using a domain specific BERT model improves tokenization, recognizing technical words that would generally become [UNK] tokens or be decomposed into subword tokens with minimal embedded semantic information by more general tokenizers. The model was initialized using the BertForSequenceClassification class, with pre-trained BERT layers followed by the randomly initialized normalization layer (size=768), dropout layer (p=0.1), dense layer (size=768), dropout layer (p=0.1) and final linear classification layer. Tokenization was done using the BertTokenizer class from the pre-trained domain specific BERT weights with a max length of 256 tokens.

BERT Transfer Learning Fine Tuning

Out of concern for the small sample size of the claim dataset, the value of transfer learning was explored. The BERT model was fine-tuned using sentences from the PubMedRCT dataset that were classified as either conclusion or non-conclusion. Non-conclusion sentences were randomly under sampled to balance the classes. The best fine-tuned model was achieved using an Adam optimizer (lr=1e-5, eps=1e-8) with a linear scheduler with warmup (warmup steps = 100) for 12 epochs with batch size 10. Weights from the epoch with the best validation f1 score were saved for use in transfer learning.

Balancing Classes

Because the target class of claim sentences is underrepresented in the dataset, the effects of randomly under sampling the non-claim class and randomly over sampling the claim class were explored.

BERT Claim Detection Fine Tuning

The BERT architecture was fine-tuned on the annotated claim dataset. The same hyperparameters that were used in the transfer learning fine tuning gave the best results for the claim detection fine tuning. These same hyperparameters were used to train all variations of the claim detection BERT model. Six total permutations of the claim detection BERT model were evaluated. Models were fine-tuned on either unbalanced, balanced by under sampling, or balanced by over sampling training data. Model weights were initialized using the Huggingface pre-trained weights for BERT and randomly initialized weights for the classification head or using the fine-tuned weights from the transfer learning model. For each, the weights from the epoch with the highest validation f1 score were used for final evaluation on the test set data.

BERT + CRF Model

A second architecture including a final CRF layer was explored as a way to investigate the value of providing the model with contextual information about other sentences in each abstract, rather than considering each independently. This was also explored as a way to further investigate the positional bias error pattern, because the CRF layer learns elements about the sequential structure of the abstract.

The conditional random field (CRF) layer takes in time-step information about a series of inputs, learns transition probabilities that determine how likely it is to move from one output label to each other output label in sequence, and then outputs information about the probability of label sequences where each label is dependent on the previous labels' features and label probabilities. The CRF output can be decoded to the most likely sequence of labels for an input series based on the learned transition probabilities [6].

BERT + CRF Architecture

The semantically rich hidden state outputs from the BERT model (size=3072) were used as inputs to a dense layer (n=768), followed by a dropout layer (p=0.1) and second dense layer (n=768) before input to a CRF layer. The BERT weights were not adjusted while training the CRF layer. Instead, sentences were first passed through BERT models and the associated hidden state was extracted for each sentence before being grouped by abstract ID in original sentence position and input into the CRF model. Abstracts were padded to a consistent sequence length of ten sentences, the length of the longest abstracts in the claim dataset. Six total CRF models were trained, one on each set of BERT embeddings from the six fine-tuned BERT permutations.

BERT + CRF Fine Tuning

The CRF architecture was trained using an Adam optimizer with a learning rate of 1e-4 for 10 epochs. Each permutation of the fine-tuned claim detection BERT models was tested as the BERT embedding input, resulting in six total BERT+CRF layer models.

Results and Discussion

BERT Model Results

Table 1 reports the macro averaged precision, recall, and F1 metrics of each BERT model, as well the metrics for the baseline model and the best reported result across all models from the literature for the test dataset [1].

Model		**Test Results (macro avg)**		
Transfer Learning	**Training Data**	**Precision**	**Recall**	**F1**
Literature Best Models		**0.887**	0.730	0.790
Baseline		0.8669	0.7610	0.7983
No	Unbalanced	0.8707	0.8607	0.8655
Yes	Unbalanced	0.8759	0.8545	0.8646
No	Undersampled	0.8625	**0.8868**	**0.8738**
Yes	Undersampled	0.8210	0.8627	0.8390
No	Oversampled	0.8655	0.8561	0.8607
Yes	Oversampled	0.8517	0.8691	0.8600

Table 1 BERT Model Results

The BERT models all outperformed the baseline model and previous best models in recall and F1 score, with the best model improving the recall by .16 points and F1 score by .08 points. The best BERT model underperformed in precision when compared to the literature's best model, though it did improve upon the baseline model.

Transfer learning did not consistently impact BERT performance. The models trained on data that had not been corrected for class balance had the highest precision metrics, while the models trained on balanced data, either by under sampling or oversampling, had relatively high recall. The transfer models that were trained on balanced data did have lower precision values than their no-transfer learning counterparts.

BERT Model Positional Results

Model		**Precision Score by Position**		
Transfer Learning	**Training Data**	**Other Positions**	**Last Sentence**	**Delta**
Baseline		0.4496	0.4173	0.03
No	Unbalanced	0.8003	0.7315	**0.07**
Yes	Unbalanced	**0.8042**	0.7366	**0.07**
No	Undersampled	0.7855	**0.7734**	0.01
Yes	Undersampled	0.7335	0.7398	-0.01
No	Oversampled	0.7903	0.7340	0.06
Yes	Oversampled	0.7718	0.7463	0.03

Table 2 BERT Precision Metrics Conditional on Sentence Position

BERT models were also evaluated for precision, recall, and F1 score conditional on whether a sentence was the last in the abstract (last sentence) or located elsewhere in the abstract (other positions), in order to investigate whether the BERT models have similar error patterns to what has been reported in the literature.

The precision score for five of the six models was lower for last sentences when compared to precision at other positions by an average of 0.04 points (Table 2). This indicates a higher false positive rate at the last sentence position versus other positions, replicating the reported error patterns. The difference in performance by position is most pronounced for the models trained on unbalanced data, while models trained on under sampled data had almost no change in performance. Models trained on balanced data with transfer learning had a smaller change in performance than their no-transfer learning counterparts.

Model		**Recall Score by Position**		
Transfer Learning	**Training Data**	**Other Positions**	**Last Sentence**	**Delta**
Baseline		0.5000	0.5000	0.00
No	Unbalanced	0.8228	0.8363	-0.01
Yes	Unbalanced	0.8021	0.8395	**-0.04**
No	Undersampled	**0.8565**	**0.8473**	0.01
Yes	Undersampled	0.8299	0.8071	0.02
No	Oversampled	0.8120	0.8379	-0.03
Yes	Oversampled	0.8324	0.8281	0.00

Table 3 BERT Recall Metrics Conditional on Sentence Position

The change in recall score between last sentence and other positions varies between BERT models (Table 3). Both models trained on unbalanced datasets showed a decrease in recall performance at other positions, while the models trained on under sampled data both showed a slight increase in recall performance at other positions. Models trained on over sampled data varied in their change in performance by position. Overall, recall was slightly lower at the other positions than at the last sentence by an average of 0.01 points. This indicates a higher rate of false positives for other positions, the same error reported in the literature. However, the pattern is not consistent between models, and most differences were less than 0.03 points.

Positional F1 scores can be found in the Appendix (Table 7). All F1 scores were higher for other positions versus last sentence for all models. F1 scores varied as expected based on the changes in precision and recall by position.

BERT + CRF Model Results

Embedding Model		**Test Results (macro avg)**		
Transfer Learning	**Training Data**	**Precision**	**Recall**	**F1**
Literature Best Models		0.887	0.730	0.790
Baseline		0.8669	0.7610	0.7983
No	Unbalanced	0.8763	0.8437	0.8586
Yes	Unbalanced	0.8817	**0.8599**	0.8702
No	Undersampled	0.8902	0.8559	**0.8720**
Yes	Undersampled	0.8852	0.8397	0.8600
No	Oversampled	0.8900	0.8495	0.8677
Yes	Oversampled	**0.9005**	0.8264	0.8567

Table 4 CRF Model Results

Table 4 reports the evaluation metrics for each CRF model on the test dataset, alongside previous literature values and the baseline model values. The BERT model that was used to generate the embeddings fed into the CRF model is identified by the transfer learning and training data columns.

The CRF models all outperformed the baseline model and best literature values in recall and F1 score metrics, and most models outperformed the baseline model and best literature values in precision.

The CRF models all performed better on precision than the corresponding BERT models used for input embedding. Five of the six CRF models performed worse on recall than the corresponding embedding BERT model. Half of the CRF models had better F1 scores and half had worse F1 scores than the corresponding BERT embedding models. For both CRF models using embeddings from BERT models trained on balanced datasets with transfer learning, the decrease in recall performance as compared to the BERT embedding models was larger than that of the no-transfer learning counterparts.

Positional F1 scores can be found in the Appendix (Table 8). All F1 scores were higher for other positions versus last sentence for all models. These scores varied as expected based on the changes in precision and recall by position.

BERT + CRF Positional Results

The CRF models were also evaluated for precision, recall, and F1 score conditional on whether a sentence was the last in the abstract (last sentence) or located elsewhere in the abstract (other positions).

Embedding Model		Precision Score by Position		
Transfer Learning	**Training Data**	**Other Positions**	**Last Sentence**	**Delta**
Baseline		0.4496	0.4173	0.03
No	Unbalanced	0.8040	0.7300	0.07
Yes	Unbalanced	0.8138	0.7403	0.07
No	Undersampled	0.8190	**0.7623**	0.06
Yes	Undersampled	0.8215	0.7173	0.10
No	Oversampled	0.8250	0.7374	0.09
Yes	Oversampled	**0.8474**	0.7116	**0.14**

Table 5 CRF Precision Metrics Conditional on Sentence Position

The precision score for all CRF models was lower for last sentences versus precision at other positions by an average of 0.09 points (Table 5). This is a greater difference in performance than was observed with the BERT only models (0.04-point change), indicating the CRF models have a more pronounced change in false positive rate for last sentence versus other positions than the BERT models. The CRF models using embeddings from BERT models trained on unbalanced data behaved similarly between transfer learning and no transfer learning. CRF models using embeddings from BERT models trained on balanced data with transfer learning showed a larger change between last sentence and other position precision performance than that of the no-transfer learning counterparts.

Embedding Model		Recall Score by Position		
Transfer Learning	**Training Data**	**Other positions**	**Last Sentence**	**Delta**
Baseline		0.500	0.5000	0.00
No	Unbalanced	0.7914	0.8509	-0.06
Yes	Unbalanced	**0.8073**	0.8362	-0.03
No	Undersampled	0.7901	**0.8717**	**-0.08**
Yes	Undersampled	0.7832	0.8316	-0.05
No	Oversampled	0.7910	0.8508	-0.06
Yes	Oversampled	0.7609	0.8365	**-0.08**

Table 6 CRF Recall Metrics Conditional on Sentence Position

The recall score for all CRF models was higher for the last sentence compared to the recall at other positions by an average of 0.06 points (Table 6). This indicates that the CRF models have an increase in false negatives in other positions as compared to last sentences, replicating the error pattern reported in the literature. The 0.06-point difference is a larger and more consistent impact than the difference observed in the BERT models, which had an average decrease of 0.01 points (Table 3). There is no obvious pattern between the positional recall performance of the BERT model used for the CRF inputs, and the positional performance of the CRF model.

Discussion

BERT models and CRF models with BERT embeddings as inputs all performed well on the claim detection task.

BERT models still exhibited some position dependent performance, with lower precision scores on last sentences than other sentences. Even though the BERT models were not being trained on any explicit positional information, there are likely some grammatical or structural differences for sentences located at different positions in the abstract leading the BERT model to learn positional indicators that bias predictions towards false positives on last sentences. There may also be some difference in structure between last sentence claims, and claims at other positions in the abstract. For example, the following claim comes from the middle of an abstract: *"Here we show that the TIFY8 ZIM domain is functional and mediated interaction with PEAPOD proteins and NINJA"*, which was correctly predicted by five of six BERT models and one of six CRF models. This claim is the last sentence of the abstract: *"Therefore, social cues reflecting population density were sufficient to elicit increased offspring growth through an adaptive hormone-mediated maternal effect"*, which was correctly identified by all models. Key words like "here" and "therefore" give the model information about the sentence position, and also highlight a potential difference in claim structure by position.

Transfer learning from the secondary dataset did not prove to be consistently impactful. This may be due to how specific to the domain of biomedical abstracts the Huggingface pre-trained weights were, resulting in the secondary dataset adding little additional value to the BERT weights.

BERT models trained on unbalanced datasets generally appeared to have more positional bias, with those BERT models showing the largest change in precision score by position and consistent changes in recall score by position. This may be because these BERT models saw a far larger proportion of other position sentence types (i.e., introduction sentences, background), leading to more training on positional indicators that the model then weighed when predicting the sentence label.

The CRF models outperformed BERT models in precision, with generally lower recall and F1 scores. One explanation may be that by predicting the most probable sequence of labels using information from the full abstract, CRF models were able to learn about the structure and composition of abstracts as a whole, including how many claims are likely to be in an abstract and where they are likely to be located. This may have reduced the overall number of false positives, and therefore resulted in a higher precision score. Contrastingly, BERT models evaluate each sentence independently so the improved recall and F1 scores may be from the reduction of false negatives on claims located in unusual positions or located within an abstract with a high density of claims.

CRF models all exhibited position dependent performance, even when the BERT embeddings used for the CRF model did not exhibit positional bias. This may indicate that the CRF layer will inherently learn positional bias because it is training on sequences of labels that do correlate with position.

Conclusions

BERT models were able to achieve superior results to literature values in recall and F1 scores, without a substantial decrease to precision, on single sentence classification with no additional context provided by the rest of the abstract. The addition of a CRF model on BERT embeddings improved precision over literature values, though recall and F1 scores were frequently lower than that of the corresponding BERT model.

The specific patterns of errors reported in the literature (false positives in the last sentence, false negatives at other positions) were reproduced with the CRF models. The BERT models demonstrated a reduction of these errors when compared to the CRF models, though the BERT models did have somewhat differential performance conditional on position, most substantially impacting precision.

Limitations & Future Work

These results are specific to biomedical domain abstracts and may not generalize to other domains or other types of text. Putting attention on the BERT embeddings instead of a CRF layer may improve overall results, without adding as much positional bias as the CRF layer.

References

[1] Achakulvisut, T., Bhagavatula, C., Acuna, D. E., & Kording, K. P. (2019). Claim extraction in biomedical publications using deep discourse model and transfer learning. *CoRR, abs/1907.00962*. https://doi.org/https://doi.org/10.48550/arXiv.1907.00962

[2] Li, X., Burns, G., & Peng, N. (2021). Scientific discourse tagging for evidence extraction. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. https://doi.org/10.18653/v1/2021.eacl-main.218

[3] Wührl, A., & Klinger, R. (2021). Claim detection in biomedical Twitter posts. *Proceedings of the 20th Workshop on Biomedical Language Processing*. https://doi.org/10.18653/v1/2021.bionlp-1.15

[4] Deka, P., Jurek-Loughrey, A., & P., D. (2022). Improved methods to aid unsupervised evidence-based fact checking for online Heath News. *Journal of Data Intelligence*, *3*(4), 474–504. https://doi.org/10.26421/jdi3.4-5

[5] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23. https://doi.org/10.1145/3458754

[6] Chawla, R. (2018, April 26). *Overview of conditional random fields*. Medium. https://medium.com/ml2vec/overview-of-conditional-random-fields-68a2a20fa541

[7] Dernoncourt, F., & Lee, J. Y. (2017, October 17). *PubMed 200K RCT: A dataset for Sequential Sentence Classification in Medical Abstracts*. arXiv.org. https://arxiv.org/abs/1710.06071

[8] Li, S. (2020, August 2). *Multi class text classification with deep learning using bert*. Medium. https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613

[9] D'Souza, J. (2022, July 7). *Language discourse in the context of natural language processing - A quick look - tib-blog.* TIB. https://blogs.tib.eu/wp/tib/2022/07/07/language-discourse-in-the-context-of-natural-language-processing-a-quick-look/

Appendix

BERT Model		**F1 Score by position**		
Transfer Learning	**Training Data**	**Other positions**	**Last Sentence**	**delta**
Baseline		0.4734	0.4549	0.02
No	Unbalanced	0.8108	0.7602	**0.05**
Yes	Unbalanced	0.8032	0.7659	0.04
No	Undersampled	**0.8154**	**0.8012**	0.01
Yes	Undersampled	0.7691	0.7643	0.00
No	Oversampled	0.8006	0.7630	0.04
Yes	Oversampled	0.7977	0.7742	0.02

Table 7 BERT Model F1 Scores Conditional on Sentence Position

Embedding Model		**F1 Score by position**		
Transfer Learning	**Training Data**	**Other positions**	**Last Sentence**	**delta**
Baseline		0.4734	0.4549	0.02
No	Unbalanced	0.7975	0.7578	0.04
Yes	Unbalanced	**0.8105**	0.7696	0.04
No	Undersampled	0.8036	**0.7953**	0.01
Yes	Undersampled	0.8007	0.7429	**0.06**
No	Oversampled	0.8067	0.7670	0.04
Yes	Oversampled	0.7961	0.7335	**0.06**

Table 8 CRF Model F1 Scores Conditional on Sentence Position