Enhancing PIO Element Detection in Medical Text Using Contextualized Embedding

Hichem Mezaoui
IMRSV Data Labs
Ottawa, Canada
hichem@imrsv.ai

Aleksandr Gontcharov
IMRSV Data Labs
Ottawa, Canada
aleksandr.gontcharov@imrsv.ai

Isuru Gunasekara
IMRSV Data Labs
Ottawa, Canada
isuru@imrsv.ai

Abstract

In this paper, we investigate a new approach to Population, Intervention and Outcome (PIO) element detection, a common task in Evidence Based Medicine (EBM). The purpose of this study is two-fold: to build a training dataset for PIO element detection with minimum redundancy and ambiguity and to investigate possible options in utilizing state of the art embedding methods for the task of PIO element detection. For the former purpose, we build a new and improved dataset by investigating the shortcomings of previously released datasets. For the latter purpose, we leverage the state of the art text embedding, Bidirectional Encoder Representations from Transformers (BERT), and build a multi-label classifier. We show that choosing a domain specific pre-trained embedding further optimizes the performance of the classifier. Furthermore, we show that the model could be enhanced by using ensemble methods and boosting techniques provided that features are adequately chosen.

1 Introduction

Evidence-based medicine (EBM) is of primary importance in the medical field. Its goal is to present statistical analyses of issues of clinical focus based on retrieving and analyzing numerous papers in the medical literature (Haynes et al., 1997). The PubMed database is one of the most commonly used databases in EBM (Sackett et al., 1996).

Biomedical papers, describing randomized controlled trials in medical intervention, are published at a high rate every year. The volume of these publications makes it very challenging for physicians to find the best medical intervention for a given patient group and condition (Borah et al., 2017). Computational methods and natural language processing (NLP) could be adopted in order to expedite the process of biomedical evidence synthesis. Specifically, NLP tasks applied to well structured documents and queries can help physicians extract appropriate information to identify the best available evidence in the context of medical treatment.

Clinical questions are formed using the PIO framework, where clinical issues are broken down into four components: Population/Problem (P), Intervention (I), Comparator (C), and Outcome (O). We will refer to these categories as PIO elements, by using the common practice of merging the C and I categories. In (Rathbone et al., 2017) a literature screening performed in 10 systematic reviews was studied. It was found that using the PIO framework can significantly improve literature screening efficacy. Therefore, efficient extraction of PIO elements is a key feature of many EBM applications and could be thought of as a multi-label sentence classification problem.

Previous works on PIO element extraction focused on classical NLP methods, such as Naive Bayes (NB), Support Vector Machines (SVM) and Conditional Random Fields (CRF) (Chung, 2009; Boudin et al., 2010). These models are shallow and limited in terms of modeling capacity. Furthermore, most of these classifiers are trained to extract PIO elements one by one which is suboptimal since this approach does not allow the use of shared structure among the individual classifiers.

Deep neural network models have increased in popularity in the field of NLP. They have pushed the state of the art of text representation and information retrieval. More specifically, these techniques enhanced NLP algorithms through the use of contextualized text embeddings at word, sentence, and paragraph levels (Mikolov et al., 2013; Le and Mikolov, 2014; Peters et al., 2017; Devlin et al., 2018; Logeswaran and Lee, 2018; Radford et al., 2018).

More recently, Jin and Szolovits (2018) proposed a bidirectional long short term memory (LSTM) model to simultaneously extract PIO

Proceedings of the BioNLP 2019 workshop, pages 217–222
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

components from PubMed abstracts. To our knowledge, that study was the first in which a deep learning framework was used to extract PIO elements from PubMed abstracts.

In the present paper, we build a dataset of PIO elements by improving the methodology found in (Jin and Szolovits, 2018). Furthermore, we built a multi-label PIO classifier, along with a boosting framework, based on the state of the art text embedding, BERT. This embedding model has been proven to offer a better contextualization compared to a bidirectional LSTM model (Devlin et al., 2018).

2 Datasets

In this study, we introduce PICONET, a multi-label dataset consisting of sequences with labels Population/Problem (P), Intervention (I), and Outcome (O). This dataset was created by collecting structured abstracts from PubMed and carefully choosing abstract headings representative of the desired categories. The present approach is an improvement over a similar approach used in (Jin and Szolovits, 2018).

Our aim was to perform automatic labeling while removing as much ambiguity as possible. We performed a search on April 11, 2019 on PubMed for 363,078 structured abstracts with the following filters: Article Types (Clinical Trial), Species (Humans), and Languages (English). Structured abstract sections from PubMed have labels such as introduction, goals, study design, findings, or discussion; however, the majority of these labels are not useful for P, I, and O extraction since most are general (e.g. *methods*) and do not isolate a specific P, I, O sequence. Therefore, in order to narrow down abstract sections that correspond to the P label, for example, we needed to find a subset of labels such as, but not limited to *population*, *patients*, and *subjects*. We performed a lemmatization of the abstract section labels in order to cluster similar categories such as *subject* and *subjects*. Using this approach, we carefully chose candidate labels for each P, I, and O, and manually looked at a small number of samples for each label to determine if text was representative.

Since our goal was to collect sequences that are uniquely representative of a description of Population, Intervention, and Outcome, we avoided a keyword-based approach such as in (Jin and Szolovits, 2018). For example, using a keyword-based approach would yield a sequence labeled *population and methods* with the label P, but such abstract sections were not purely about the population and contained information about the interventions and study design making them poor candidates for a P label. Thus, we were able to extract portions of abstracts pertaining to P, I, and O categories while minimizing ambiguity and redundancy. Moreover, in the dataset from (Jin and Szolovits, 2018), a section labeled as P that contained more than one sentence would be split into multiple P sentences to be included in the dataset. We avoided this approach and kept the full abstract sections. The full abstracts were kept in conjunction with our belief that keeping the full section retains more feature-rich sequences for each sequence, and that individual sentences from long abstract sections can be poor candidates for the corresponding label.

Category	**Sentences**
I	22818
I O	7
I P	337
O	10994
P	30106
P O	13
NEGATIVE	32053

Table 1: Number of occurrences of each category P, I and O in abstracts.

For sections with labels such as *population and intervention*, we created a mutli-label. We also included negative examples by taking sentences from sections with headings such as *aim*. Furthermore, we cleaned the remaining data with various approaches including, but not limited to, language identification, removal of missing values, cleaning unicode characters, and filtering for sequences between 5 and 200 words, inclusive.

3 BERT-Based Classification Model

3.1 Background

BERT (Bidirectional Encoder Representations from Transformers) is a deep bidirectional attention text embedding model. The idea behind this model is to pre-train a bidirectional representation by jointly conditioning on both left and right contexts in all layers using a transformer (Vaswani et al., 2017; Devlin et al., 2018). Like any other language model, BERT can be pre-trained on different contexts. A contextualized representation is generally optimized for downstream NLP tasks.

Since its release, BERT has been pre-trained on a multitude of corpora. In the following, we describe different BERT embedding versions used for our classification problem. The first version is based on the original BERT release (Devlin et al., 2018). This model is pre-trained on the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia, text passages were extracted while lists were ignored. The second version is BioBERT (Lee et al., 2019), which was trained on biomedical corpora: PubMed (4.5B words) and PMC (13.5B words).

3.2 The Model

The classification model is built on top of the BERT representation by adding a dense layer corresponding to the multi-label classifier with three output neurons corresponding to PIO labels. In order to insure that independent probabilities are assigned to the labels, as a loss function we have chosen the binary cross entropy with logits (BCEWithLogits) defined by

$$E = -\sum_{i=1}^{n}(t_i log(y_i) + (1 - t_i) log(1 - y_i)); \quad (1)$$

where **t** and **y** are the target and output vectors, respectively; **n** is the number of independent targets (n=3). The outputs are computed by applying the logistic function to the weighted sums of the last hidden layer activations, s,

$$y_i = \frac{1}{1 + e^{-s_i}}, \quad (2)$$

$$s_i = \sum_{j=1} h_j w_{ji}. \quad (3)$$

For the original BERT model, we have chosen the smallest uncased model, Bert-Base. The model has 12 attention layers and all texts are converted to lowercase by the tokenizer (Devlin et al., 2018). The architecture of the model is illustrated in Figure 1.

Using this framework, we trained the model using the two pretrained embedding models described in the previous section. It is worth to mention that the embedding is contextualized during the training phase. For both models, the pretrained embedding layer is frozen during the first epoch (the embedding vectors are not updated). After the first epoch, the embedding layer is unfrozen and the vectors are fine-tuned for the classification task during training. The advantage of this approach is that few parameters need to be learned from scratch (Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018).

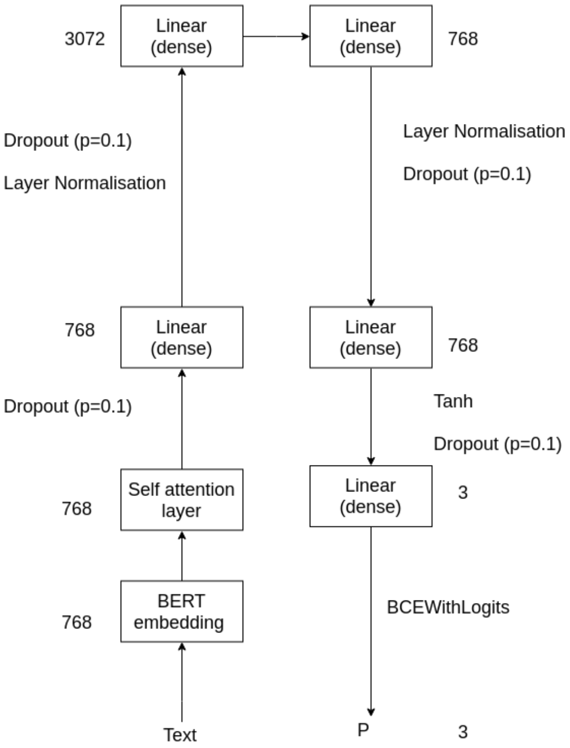


Figure 1: Structure of the classifier.

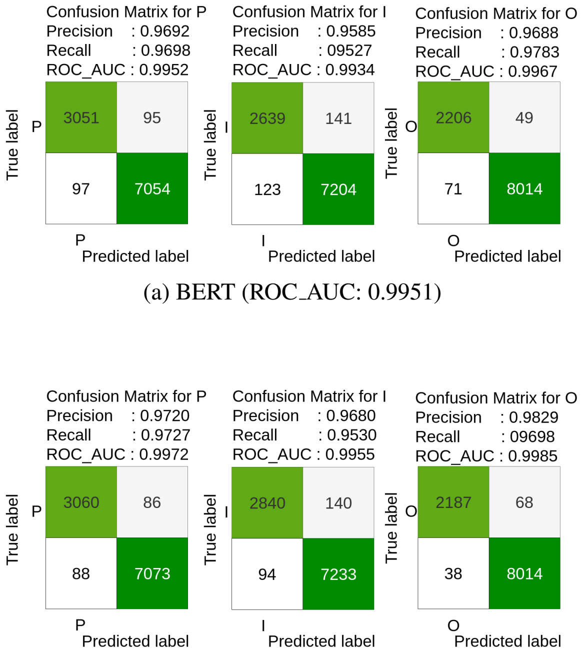
4 Results

4.1 Performance Comparison

In order to quantify the performance of the classification model, we computed the precision and recall scores. On average, it was found that the model leads to better results when trained using the BioBERT embedding. In addition, the performance of the PIO classifier was measured by averaging the three Area Under Receiver Operating Characteristic Curve (ROC_AUC) scores for P, I, and O. The ROC_AUC score of 0.9951 was obtained by the model using the general BERT embedding. This score was improved to 0.9971 when using the BioBERT model pre-trained on medical context. The results are illustrated in Figure 2.

4.2 Model Boosting

We further applied ensemble methods to enhance the model. This approach consists of combin-



(a) BERT (ROC_AUC: 0.9951)

(b) BioBERT (ROC_AUC: 0.9971)

Figure 2: ROC_AUC scores and confusion matrices.

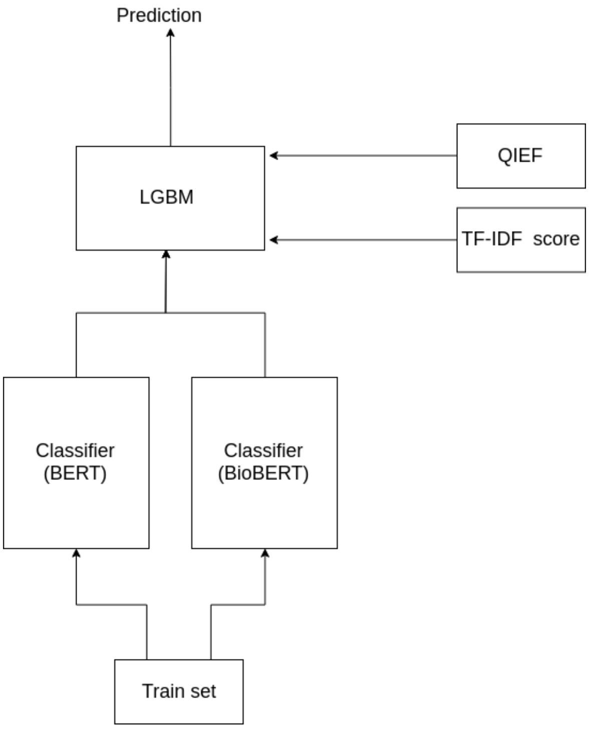


Figure 3: An illustration of the LGBM framework: : combining the two base models and the TF-IDF and QIEF features.

ing predictions from base classifiers with features of the input data to increase the accuracy of the model (Merz, 1999).

We investigate an important family of ensemble methods known as boosting, and more specifically a Light Gradient Boosting Machine (LGBM) algorithm, which consists of an implementation of fast gradient boosting on decision trees. In this study, we use a library implemented by Microsoft (Ke et al., 2017). In our model, we learn a linear combination of the prediction given by the base classifiers and the input text features to predict the labels. As features, we consider the average term frequency-inverse document frequency (TF-IDF) score for each instance and the frequency of occurrence of quantitative information elements (QIEF) (e.g. percentage, population, dose of medicine). Finally, the output of the binary cross entropy with logits layer (predicted probabilities for the three classes) and the feature information are fed to the LGBM.

Model	**ROC_AUC**	**F1**
BERT	0.9951	0.9666
BioBERT	0.9971	0.9697
BERT + TF-IDF + QIEF	0.9981	0.9784
BioBERT + TF-IDF + QIEF	0.9996	0.9793
BERT + BioBERT + TF-IDF + QIEF	0.9998	0.9866

Table 2: Performance of the classifiers in terms of ROC_AUC and F1 scores.

We train the base classifier using the original training dataset, using 60% of the whole data as training dataset, and use a five-fold cross-validation framework to train the LGBM on the remaining 40% of the data to avoid any information leakage. We train the LGBM on four folds and test on the excluded one and repeat the process for all five folds.

The results of the LGBM classifier for the different boosting frameworks and the scores from the base classifiers are illustrated in Table 2. The highest average ROC_AUC score of 0.9998 is obtained in the case of combining the two base learners along with the TF-IDF and QIEF features.

5 Discussion and Conclusion

In this paper, we presented an improved methodology to extract PIO elements, with reduced ambiguity, from abstracts of medical papers. The proposed technique was used to build a dataset of PIO elements that we call PICONET. We further proposed a model of PIO elements classification using state of the art BERT embedding. It has been shown that using the contextualized BioBERT embedding improved the accuracy of the classifier. This result reinforces the idea of the importance of embedding contextualization in subsequent classification tasks in this specific context.

In order to enhance the accuracy of the model, we investigated an ensemble method based on the LGBM algorithm. We trained the LGBM model, with the above models as base learners, to optimize the classification by learning a linear combination of the predicted probabilities, for the three classes, with the TF-IDF and QIEF scores. The results indicate that these text features were adequate for boosting the contextualized classification model. We compared the performance of the classifier when using the features with one of the base learners and the case where we combine the base learners along with the features. We obtained the best performance in the latter case.

The present work resulted in the creation of a PIO elements dataset, PICONET, and a classification tool. These constitute an important component of our system of automatic mining of medical abstracts. We intend to extend the dataset to full medical articles. The model will be modified to take into account the higher complexity of full text data and more efficient features for model boosting will be investigated.

References

Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545.

Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1):29.

Grace Y Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC medical informatics and decision making*, 9(1):10.

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

R Brian Haynes, David L Sackett, W Scott Richardson, William Rosenberg, and G Ross Langley. 1997. Evidence-based medicine: How to practice & teach ebm. *Canadian Medical Association. Journal*, 157(6):788.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Di Jin and Peter Szolovits. 2018. Pico element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Christopher J Merz. 1999. Using correspondence analysis to combine classifiers. *Machine Learning*, 36(1-2):33–58.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

John Rathbone, Loai Albarqouni, Mina Bakhit, Elaine Beller, Oyungerel Byambasuren, Tammy Hoffmann, Anna Mae Scott, and Paul Glasziou. 2017. Expediting citation screening using pico-based title-only screening for identifying studies in scoping searches and rapid reviews. *Systematic reviews*, 6(1):233.

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.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.