Extracting PICO Elements From RCT Abstracts Using 1-2gram Analysis And Multitask Classification

Xia Yuan  
School of Life Science & Technology, University of Electronic Technology, University of Electronic Science & Technology of China  
bruce_xia6116@163.com

Liao Xiaoli  
School of Life Science & Technology, University of Electronic Technology, University of Electronic Science & Technology of China  
liaoxl@uestc.edu.cn

Li Shilei  
School of Life Science & Technology, University of Electronic Technology, University of Electronic Science & Technology of China  
503303509@qq.com

Shi Qinwen  
School of Life Science & Technology, University of Electronic Technology, University of Electronic Science & Technology of China  
1557679224@qq.com

Li Ke  
School of Life Science & Technology, University of Electronic Technology, University of Electronic Science & Technology of China  
colinlike@163.com

ABSTRACT

The core of evidence-based medicine is to read and analyze numerous papers in the medical literature on a specific clinical problem and summarize the authoritative answers to that problem. Currently, to formulate a clear and focused clinical problem, the popular PICO framework is usually adopted, in which each clinical problem is considered to consist of four parts: patient/problem (P), intervention (I), comparison (C) and outcome (O). In this study, we compared several classification models that are commonly used in traditional machine learning. Next, we developed a multitask classification model based on a soft-margin SVM with a specialized feature engineering method that combines 1-2gram analysis with TF-IDF analysis. Finally, we trained and tested several generic models on an open-source data set from BioNLP 2018. The results show that the proposed multitask SVM classification model based on 1-2gram TF-IDF features exhibits the best performance among the tested models.

CCS Concepts

Computing methodologies → Information extraction

Keywords

PICO extraction; Evidence-based medicine; TF-IDF; 1-2gram; Soft-margin SVM.

1. INTRODUCTION

Evidence-based medicine (EBM) is a major branch of the medical field. Its purpose is to present statistical analyses of issues of clinical focus based on reading, analyzing, and integrating numerous papers in the medical literature [1]. The PubMed database is one of the most commonly used databases in EBM [2][3]. Successful EBM applications, which rely on abundant research-based evidence combined with clinical expertise [4] and systematic reviews, can effectively assist in clinical decision-making. In most cases, the PICO framework is used to develop a well-defined, focused description clinical problem. In this framework, clinical issues are broken down into four components: patient/problem (P), intervention (I), comparison (C), and outcome (O) [4]. In 2014, Abigail M Methley et al. compared the PICO search method with several other search methods and found that the PICO method can significantly improve the efficiency of a literature search [5]. In 2017, John Rathbone et al. studied the literature screening performed in 10 systematic reviews and found that using the PICO framework can significantly improve the efficiency of literature screening [6].

The standard EBM analysis process usually consists of the following steps:

1). Use the PICO framework to describe the clinical issue to be studied and develop a literature search strategy based on the formulated problem.

2). In accordance with the developed literature search strategy, attempt to retrieve all documents that meet the stated requirements.

3). Among the retrieved documents, consider the title, abstract, full text and other information to filter out the articles of interest.

4). Perform a comprehensive analysis of a few of the documents that are ultimately selected and summarize the solutions to and theoretical basis (evidence) for the corresponding clinical problem.

Unfortunately, because the PICO elements are not explicitly identified in the structured abstracts of most medical papers, the retrieval and screening of documents are extremely time-consuming tasks in this era of information explosion. It is often necessary for researchers to thoroughly read each abstract to extract the corresponding PICO information before filtering. Therefore, the ability to automatically extract the PICO elements from the structured abstracts found in PubMed by means of machine learning methods would facilitate the EBM process [7].
In this study, we present a term frequency-inverse document frequency (TF-IDF)-based feature engineering method that incorporates a model based on 1-2gram. We also propose a soft-margin support vector machine (SVM) model based on multitask classification for automatically extracting sentence-level PICO elements from structured abstracts in the biomedical literature. The main contributions of this paper are as follows:

1. First, we analyzed the vocabulary and grammatical features of medical English from the linguistic perspective. We also performed a statistical analysis of word frequency on the PICO sentences in our data set. On this basis, we developed a feature construction method that combines 1-2gram with TF-IDF analysis.

2. By designing 6 sets of controlled experiments, we demonstrated the efficiency of the 1-2gram model. We also performed a performance comparison between the TF-IDF feature engineering method and the word2vec word embedding method.

3. We also compared our model with two classic classification methods used in integration learning, i.e., the random forest (RF) method and XGBoost, using the same open-source data set for training and testing.

4. Using the same evaluation indicators, we compared our model with the best two models Naïve Bayes (NB) and long short-time memory (LSTM) from previous studies. Although we used the traditional TF-IDF approach for feature construction, whereas the LSTM model uses word2vec for feature engineering, our model produces better experimental results than the LSTM model does.

2. RELATED WORKS

The first study on the automatic detection and extraction of PICO elements from structured abstracts in the biomedical literature was conducted in 2007 by Demner-Fushman et al. These authors proposed a rule-based pattern matching method for detecting PICO elements in document abstracts by applying corresponding rules formulated by experts.

In recent years, the “C” category has often been incorporated into the “I” category because “comparative” elements can refer to other interventions or to the decision not to participate in clinical randomized controlled trials (RCTs), which should be addressed in the intervention category. In fact, very few abstracts with comparison labels are found in PubMed. Moreover, in most PICO studies, C and I elements are merged into the same category in practice because they are considered to form one semantic group.

Ke-Chun Huang et al. (2011) [17], n-gram analysis has not been adopted. Moreover, the potential advantages of n-gram models have not been mentioned in any previous related biomedical texts.

In this paper, we propose a TF-IDF-based feature engineering method that combines this word-frequency analysis with a 1-2gram analysis to account for the particularities of the language used in medical texts. We compared our novel feature engineering method with that of word2vec using the same data set and the same soft-margin SVM modeling approach. We found that the proposed 1-2gram TF-IDF feature engineering method obviously outperforms word2vec.

3. MATERIALS AND METHODS

To reduce this cumbersome workload, the data used in this study were processed from an open-source data set from BioNLP 2018. The flow chart for this study is shown in Fig 1. The data processing was mainly performed by means of regular matching. For feature engineering, we primarily relied on our 1-2gram TF-IDF model. Finally, we constructed three binary classification models using a standard soft-margin SVM approach for the classification of PICO sentences.

After data preprocessing, three separate data sets were used to train three binary classifier models. Ten-fold cross-validation was applied.

3.1 Data Processing

The original data consist of a total of 24,668 abstracts and 319,968 sentences, of which 21,198 of the abstracts and 27,696 of the sentences have P tags. The specific statistics are shown in Table 1.

| Label | Articles | Sentences |
|-------|----------|-----------|
| P     | 21198    | 27696     |
| I     | 13712    | 24603     |
| O     | 20473    | 32526     |

We designed three separate binary classification models for P, I, and O classification to mitigate the possible effects of the imbalanced data distribution on the model classification results.
and to avoid sentences with fuzzy labels being assigned as negative samples of non-P/I/O labels, which could affect the feature learning capabilities of the model [20]. First, we filtered out all sentences with non-P/I/O tags from the data set. Then, for the P classification model, we replaced all P tags in the data set with a value of 1 and all non-P tags with a value of 0; for the I classification model, we replaced all I tags in the data set with a value of 1 and all non-I tags with a value of 0; and for the O classification model, we replaced all O tags in the data set with a value of 0. We used the stratified sampling method to divide each of the data sets obtained as described above into a training set, a test set and a verification set at a ratio of 8:1:1.

3.2 Feature Extraction

We calculated the TF-IDF for each word in every sentence in the data set and then constructed a dictionary, which we used to vectorize each sentence.

\[
tf_{i,j} = \frac{n_{i,j}}{\sum n_{i,j}}
\]  

The term frequency (TF) refers to the frequency with which a given word appears in a sentence. \(n_{i,j}\) is the number of occurrences of word \(w_i\) in sentence \(s_j\), and the denominator is the sum of the occurrences of all words in sentence \(s_j\).

\[
idf_{i} = \log \frac{|D|}{|\{j: w_i \in s_j\}|}
\]

Where \(|D|\) is the total number of sentences in the data set and \(|\{j: w_i \in s_j\}|\) is the total number of sentences containing the word \(w_i\). If this word is not present in the corpus, the denominator in the above expression will be zero; therefore, we add 1 to the denominator, i.e., \(|\{j: w_i \in s_j\}| + 1\), to prevent division by zero. Finally, the TF-IDF of word \(w_i\) in sentence \(s_j\) is expressed as

\[
tfidf_{i,j} = tf_{i,j} \times idf_{i}
\]

Medical language is a distinct language [21]. In medical English, especially in medical texts, numerous structures, such as nouns, action nouns, and action noun phrases, are used. Nouns and action nouns mostly consist of single specialized words, whereas an action noun phrase is typically composed of two or more words, and a noun phrase is typically an action noun phrase containing a preposition [22]. Therefore, we added a 1-2gram model to our TF-IDF model. A 1-2gram model was chosen because in English, most prepositions are stop words, and most noun phrases and preposition [22]. Therefore, we added a 1-2gram model to our TF-IDF model.

3.3 Soft-margin SVM

In this study, we replaced the traditional multi-classifier model with three binary classification models to complete the classification task. In terms of the selection of binary classification model, we chose the currently mature SVM classifier. Although text categorization is a typical nonlinear classification problem, we ultimately choose to use SVM classifiers for binary classification after comparing several models. As shown by the results of the previous word-frequency analysis, P and I sentences exhibit a certain overlap in their word-frequency distributions (many P and I sentences contain the same words, such as patients, group, and study). Therefore, after vectorization, the vectors of P and I sentences show a certain linear indivisibility in the vector space; consequently, to increase the generalization ability of the model, we used a linear kernel and a soft margin to train a nonlinear SVM. The soft-margin SVM approach represents an improvement over the standard SVM approach. In a standard SVM, the model constraint is

\[
y_i(w^T x_i + b) \geq 1, \quad i = 1, ..., n
\]

where \(x_i\) is the vector representation of sentence \(i\) and \(y_i\) is the label of sentence \(i\), and the objective function is

\[
\min_{\frac{1}{2}||w||^2}
\]

Linear indivisibility means that for some sample points \((x_i, y_i)\), the constraint that the function interval must be greater than or equal to 1 is not satisfied. Therefore, for each sample point, we introduce a slack variable \(\xi_i \geq 0\) such that the function interval plus the slack variable will be greater than or equal to 1; then, the constraint becomes

\[
y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, ..., n
\]

For each slack variable \(\xi_i\), some penalty must be paid; therefore, the objective function becomes

\[
\min_{\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i}
\]

Here, \(C > 0\) is a penalty parameter, which is utilized to control the relative weight between the two terms in the objective function (which serve to “find the hyperplane with the largest margin” and “minimize the deviation of the data points”, respectively). \(\xi\) is a variable that needs to be optimized during the model training process.

4. RESULTS

To prove the optimal performance of the 1-2gram model, we designed 6 sets of control experiments. The n-gram range in the TF-IDF analysis was set to 1, 2, 3, 1-2, 1-3, and 2-3; tests were performed using SVM models trained with these n-gram ranges with all other parameters being the same, and the resulting accuracy rate (acc) and F1 values were recorded as evaluation indicators.

The test results show that the 1-2gram model exhibits the best performance in terms of both the acc and F1 values.

To test the efficiency of the proposed 1-2gram TF-IDF feature engineering method, we compared it with the word2vec method. First, we obtained summary data for 200,000 RCT articles from Ji Young Lee in 2017 [23]. Combined with our existing 24,668 RCT abstracts, we obtained a total of 224,668 abstracts. After a series of processing steps, such as word segmentation and stop word removal, the open-source toolkit gensim was used to train 200-dimensional word vectors. Then, we use the trained word2vec vectors to vectorize all sentences in the data set, performed training and testing using the same soft-margin SVM model, and compared the results with those of the model trained with our 1-2gram TF-IDF features, as presented in Table 2.

| Table 2: Results of the TF-IDF and word2vec comparison experiment |
|---------------------------|--------------------------|--------------------------|
|                           | P elements | I elements | O elements |
| TF-IDF                    |   0.925    |  0.838     |  0.879      |
| word2vec                  |   0.894    |  0.796     |  0.842      |
Note: The experimental results for the TF-IDF method and the word2vec method compared in this study were obtained using the same soft-margin SVM classification model; the P elements column represents the results for a classification model constructed for P sentences, and the I elements column represents the results for a classification model constructed for I sentences.

![ROC curve](image1)

Fig 2. The ROC curves of the TF-IDF and word2vec comparison experiment: (a) results for P elements, (b) results for I elements, and (c) results for O elements.

It is not difficult to see from Table 2 and Fig 2 that the results (in terms of the P, R, and F1 values) obtained with the TF-IDF model are better than those obtained with the word2vec model.

We also compared our method with the RF and XGBoost methods, which are commonly used for SVM modeling and integrated learning, on the same data set. The experimental P, R and F1 values are shown in Table 3. The corresponding ROC curves are given in Fig 3. The experimental results show that the soft-margin SVM model exhibits the best performance on the same data set.

|                  | P elements | I elements | O elements |
|------------------|------------|------------|------------|
|                  | P | R | F1 | P | R | F1 | P | R | F1 |
| SVM              | 0.925 | 0.838 | 0.879 | 0.842 | 0.789 | 0.814 | 0.886 | 0.897 | 0.891 |
| RF               | 0.899 | 0.813 | 0.854 | 0.838 | 0.759 | 0.808 | 0.881 | 0.834 | 0.861 |
| XGBoost          | 0.860 | 0.857 | 0.852 | 0.829 | 0.814 | 0.791 | 0.857 | 0.838 | 0.832 |

Table 3: Comparison of the soft-margin SVM model used in this study with the standard RF and XGBoost models.

Note: The acc, P, R and F1 values obtained through 10-fold cross-validation are used to evaluate the models.

![ROC curve](image2)

Fig 3: The ROC curves of the SVM model, RF model and XGBoost model comparison experiment: (a) results for P elements, (b) results for I elements, and (c) results for O elements.
For our classification task, it is not truly meaningful to simply look at the P and R values of the models. We are more interested in evaluating model performance based on the F1 value, which is the harmonic mean of the P and R values. The F1 results indicate that the SVM model significantly outperforms the RF and XGBoost models. The ROC curves shown in Fig 3 also support this finding.

5. DISCUSSION

The data set used throughout this study is an open-source data set from BioNLP 2018 [17]. Because some previous studies have used different data sets and different methods, it is sometimes difficult to compare our findings with previous research results. However, since the problem studied is the same (PICO element extraction from RCT abstracts) and the model evaluation indicators are also the same (P, R, and F1 values), we can nevertheless compare the data presented in some previous research papers with the results of our model. In particular, since the data set used in this project is the same as the data set used in the Di Jin 2018 paper and is similar to the data set used in the Ke-Chun Huang 2011 paper, we will consider the results presented in those papers for comparison, as shown in Table 4.

Table 4: Comparison of model results.

|               | P  | R | F1 | P  | R | F1 | P  | R | F1 |
|---------------|----|----|----|----|----|----|----|----|----|
| NB            | 0.902 | 0.925 | 0.913 | 0.786 | 0.716 | 0.749 | 0.836 | 0.920 | 0.876 |
| LSTM          | 0.885 | 0.828 | 0.836 | 0.749 | 0.815 | 0.781 | 0.845 | 0.832 | 0.838 |
| SVM           | 0.924 | 0.838 | 0.879 | 0.842 | 0.789 | 0.814 | 0.856 | 0.897 | 0.891 |

Note: The LSTM model was described in the Di Jin 2018 paper, the NB model was described in the Ke-Chun Huang 2011 paper, and the SVM model is the model studied in the present paper.

In terms of accuracy (P value), the SVM model shows the best performance for P, I and O elements. In terms of recall (R value), the performance of the NB model is better than that of the SVM model for P and I elements, and the performance of the LSTM model is better than that of the SVM model for I elements. In terms of the F1 value, the SVM model shows the best performance for both I and O elements, but the performance of the SVM model for P elements is not as good as that of the NB model.

The reason for the differences evident in Table 4 is that in the Ke-Chun Huang 2011 study, the authors used the words with the highest frequencies (1-14%) to represent the characteristics of each sentence, whereas we incorporated a 1-2gram model into our TF-IDF analysis to represent sentence characteristics. Using only the highest-frequency words to represent sentence features results in a considerable loss of information. By contrast, the TF-IDF approach not only retains the information contained in all words in a sentence but also highlights the characteristic information provided by low-frequency words that are unique to different sentences.

Regarding the introduction of the 1-2gram model, in the previous method and result sections, we have proven the superiority of the 1-2gram model for application to medical language through comparative experiments. By contrast, the word2vec feature engineering method was used in the Di Jin 2018 study, and the results of this complex feature engineering method are not as good as those of our 1-2gram TF-IDF method, as can be seen by comparing the experimental results on the same data set.

As mentioned in the Di Jin 2018 paper, this method of constructing TF-IDF features based on probability statistics often ignores the correlations between different words and between different sentences. Therefore, its performance in natural language processing is not as good as that of word2vec. However, the TF-IDF method based on 1-2gram performs much better than the word2vec method in terms of sentence feature representation, whether from the statistical results of word frequency of PICO sentences or the experimental results of TF-IDF comparison with word2vec. Because word2vec is better at capturing word similarity, word2vec is more suitable for context-based text understanding tasks. For our task, however, we believe that the
6. CONCLUSION
In this study, we have presented a TF-IDF feature engineering method that incorporates a 1-2gram model. We have also proposed a soft-margin SVM model based on multitask classification for the automatic extraction of sentence-level PICO elements from structured abstracts in the biomedical literature. We tested the performance of our model on an open-source data set and compared the results with those of previous research. We found that our model achieves the best F1 values for I and O elements. Moreover, although the F1 value of our model is slightly inferior to that of an NB model for P elements, our model is superior to the NB model in terms of accuracy (P value). Notably, in this study, we considered only the automatic extraction of PICO information from RCT abstracts. However, the PICO information will, of course, be more completely described in the full text. Therefore, in the future, we will use a deep learning method (self-attention) to construct more complex models for automatically identifying PICO information from the full text of RCT articles.

7. ACKNOWLEDGMENTS
This project is funded by Sichuan Science and Technology Program Grant No.17PTDJ0078 & 182DYF0569. Thank Deqi Liu, the researcher of Sichuan Provincial Health and Family Planning Policy and Medical Information Research Institute, for his suggestions.

8. REFERENCES
[1] David L. Sackett, W. Scott Richardson, William Rosenberg R. Brian Haynes Churchill Livingstone. "Evidence-based Medicine How to practice and teach EBM". WB Saunders Company. 1997.
[2] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. "Evidence based medicine: what it is and what it isn’t”. Br Med J 1996;312(7023):71–2.
[3] Lin J, Demner-Fushman D. “The role of knowledge in conceptual retrieval”. In: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval – SIGIR '06. Seattle, Washington, USA; 2006. p. 99–106.
[4] W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert SA Hayward. “The well-built clinical question: a key to evidence-based decisions”. ACP journal club, 123(3):A12–A12. 1995.
[5] Abigail M Methley, Stephen Campbell, Carolyn Chew-Graham. “PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews”. BMC Health Services Research (2014) 14:579.
[6] John Rathbone, Loai Albarqouni, Mina Bakhit. “Expediting citation screening using PICObased title-only screening for identifying studies in scoping searches and rapid reviews”. Systematic Reviews (2017) 6:233.
[7] L. M. Md, P. Srinivasan, and P. D. “Categorization of Sentence Types in Medical Abstracts”. In Proceeding of the 2003 Annual Symposium of the American Medical Informatics Association (AMIA), p. 440–444, 2003.
[8] Boser, BE, Guyon, IM, Vapnik, VN.” A training algorithm for optimal margin classifiers”. In: Proceedings of the Annual Workshop on Computational Learning Theory. ACM 2010. 144-152.
[9] Caruana, R." Multitask learning”. Springer US, 1998.p 65-78.
[10] Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. “Utilization of the pico framework to improve searching pubmed for clinical questions”. BMC medical informatics and decision making, 7(1):16. 2007.
[11] Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes." Combining classifiers for robust pico element detection”. BMC medical informatics and decision making, 10(1):29. 2010.
[12] Huang X, Lin J, Demner-Fushman D. ”Evaluation of PICO as a knowledge representation for clinical questions”. AMIA Annu Symp Proc 2006; 2006:359–63.
[13] Demner-Fushman D, Lin J.” Knowledge extraction for clinical question answering: preliminary results”. Assoc Adv Artif Intell 2005:1–9.
[14] Grace Yuet-Chee Chung. " Towards identifying intervention arms in randomized controlled trials: Extracting coordinating constructions”. Journal of Biomedical Informatics 42 (2009) 790–800.
[15] Grace Y Chung.” Sentence retrieval for abstracts of randomized controlled trials”. BMC medical informatics and decision making, 9(1):10. 2009.
[16] Di Jin, Peter Szolovits.” PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks”. Proceedings of the BioNLP 2018 workshop. 2018.
[17] Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. “Pico element detection in medical text without metadata: Are first sentences enough?”. Journal of biomedical informatics, 46(2013):940–946. 2013.
[18] Haibo He, and Edwardo A. Garcia.” Learning from Imbalanced Data”.2009. in IEEE Transactions on Knowledge & Data Engineering, vol. 21, no. , pp. 1263-1284, 2008.
[19] Oded Maimon & Lior Rokach.,” Data Mining and Knowledge Discovery Handbook”. Springer New York Dordrecht Heidelberg London,2010.
[20] Ke-Chun Huang1, Charles Chih-Ho Liu1,Shung-Shiang Yang.,” Classification of PICO Elements by Text Features Systematically Extracted from PubMed Abstracts”. IEEE International Conference on Granular Computing.2011.
[21] Varantola K.,” Special Language and General Language”. ALSEDL2SP Newsletter .1986 .9 (2) :10220
[22] Rosita B. Maglie.” Understanding the Language of Medicine”. Springer Aracne : 2009.p.12-25.
[23] Franck Dernoncourt, Ji Young Lee.” PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts”. IJCNLP 2017.