Computers in Biology and Medicine 79 (2016) 276–285

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Structuralizing biomedical abstracts with discriminative linguistic features

Sejin Nam[a], Senator Jeong[b], Sang-Kyun Kim[c], Hong-Gee Kim[d], Victoria Ngo[e], Nansu Zong[f,*]

[a] National Center of Excellence in Software, Chungnam National University, South Korea
[b] National Center for Medical Information & Knowledge, Korea National Institute of Health, South Korea
[c] Mibyeong Research Center, Korea Institute of Oriental Medicine, Daejeon, South Korea
[d] Biomedical Knowledge Engineering Laboratory, School of Dentistry, Seoul National University, South Korea
[e] Betty Irene Moore School of Nursing, University of California, Davis, USA
[f] Department of Biomedical Informatics, School of Medicine, University of California, San Diego, USA

ARTICLE INFO

Keywords:
Discriminative linguistic features
Sentence classification
Structured abstract
Biomedical research paper
IMRAD format

ABSTRACT

Objective: Nearly 75% of the abstracts in MEDLINE papers present in an unstructured format. This study aims to automate the reformatting of unstructured abstracts into the Introduction, Methods, Results, and Discussion (IMRAD) format. The quality of this reformatting relies on the features used in sentence classification. Therefore, we explored the most effective linguistic features in MEDLINE papers.
Methods: We constructed a feature set consisting of bag of words, linguistic features, grammatical features, and structural features. In order to evaluate the effectiveness, which is the capability of the sentence classification with the features, three datasets from PubMed Central Open Access Subset were selected and constructed: (1) structured abstract (SA) for training, (2) unstructured RCT abstract (UA-1) and (3) unstructured general abstract (UA-2). F-score and accuracy were used to measure the effectiveness on IMRAD section level and the overall classification.
Results: Adding linguistic features improves the classification of the abstract sentence from 1.2% to 35.8% in terms of accuracy in three abstract datasets. The highest accuracies achieved were 91.7% in SA, 86.3% in UA-1, and 77.9% in UA-2. Linguistic features (dimensions=15) had fewer dimensions than bag-of-words (dimensions= 1541). All representative linguistic features (n-gram and verb phrase, and noun phrase) for each section are identified in our system (available at http://abstract.bike.re.kr).
Conclusion: Linguistic features can be used to effectively classify sentence with low computation burden in MEDLINE abstract.

1. Introduction

Biomedical research paper abstracts can be either structured or unstructured. Unstructured abstracts, which describe studies in a narrative of continuous sentences without a formal heading structure, are recognized as an ineffective format for conveying key research information [1]. Alternatively, structured abstracts help researchers to search, select, read, and extract information about the study through clear instruction [2–8]. A number of initiatives have been proposed to standardize the format and content of structured abstracts [9–12]. The Introduction, Methods, Results, and Discussion (IMRAD) format is predominant in medical scientific writing [9,13]. Despite the gradual rise in percentage of MEDLINE records with structured abstracts increasing from 0.4% in 1989–1991 to 13.1% in 1992–2006, and eventually to 23.0% in 2008 [4], nearly 75% of the abstracts added annually to MEDLINE still present in an unstructured format [14].

We aim to automate the transformation of unstructured abstracts into structured abstracts with the goal of having sentences duly classified into appropriate IMRAD sections. The effectiveness of such sentence classification, which is the capability of a method to produce an expected outcome of classifying unstructured biomedical abstracts into IMRAD sections, depends on both feature selection and classification algorithm. In the biomedical domain, previous studies have prioritized feature selection, often paying comparatively less attention to the linguistic characteristics of the biomedical paper abstracts. In this study, we focused on linguistic features that distinguish specific sections from others in structured abstracts. To identify distinguishing features, we analyzed sentences in a large corpus of structured abstracts and identified a wide array of candidate features characterizing each IMRAD section. Then we investigated which linguistic features, alone or in combinations with other features, achieved the best results in sentence classification. Additionally, as a contribution to

[*] Corresponding author.
E-mail address: nazong@ucsd.edu (N. Zong).

http://dx.doi.org/10.1016/j.compbiomed.2016.10.026
Received 27 May 2016; Received in revised form 28 October 2016; Accepted 31 October 2016

this research, we created a web application (http://abstract.bike.re.kr) that helps users to structure abstracts with linguistic features discovered in this study.

2. Related works

There have been notable research efforts on the automatic classification of sentences in biomedical paper abstracts in the past decade. We reviewed them in terms of the classification features for different data sets.

To target data in general domain, Hirohata et al. [15] used bi-gram, sentence location, and features from previous-next sentences. By employing Conditional Random Fields (CRF), they achieved an accuracy of 95.5% per sentence and 64% per abstract in the cross-validation set for structured abstracts. To extract key sentences from abstracts, Ruch et al. [16] used stemmed n-gram (1≤n≤3), sentence length, and sentence location, and achieved the accuracy of 84.6% for unstructured abstracts with a naïve Bayesian classifier. Guo et al. [17] used previous sentence feature, sentence location, word, bi-gram, verb, verb class, part-of-speech (POS), grammatical relation, subject and object, and voice. They achieved an accuracy of 89% with Support Vector Machine (SVM) classifier. Their best results were achieved when they excluded verbs, but kept all other features.

Randomized Controlled Trials (RCT) prevailed as the main target of sentence classification efforts. McKnight and Srinivasan [18] used sentence location and bag-of-words (BOW) as the main features for classifying sentences in abstracts of RCTs. Their SVM classifier achieved accuracies of 85.5% and 74.2% per section in structured and unstructured abstracts, respectively. Yamamoto and Takagi [19] used TF/IDF value, presence of auxiliary verb, verb tense, term vector, and chi-squared values of term, section collocation, and subject (noun)-predicate (verb) pair in each section. However, the features did not significantly outperform the McKnight's system, with an accuracy of 88% for structured abstracts and 58.2% for unstructured abstracts. Xu et al. [20] used unprocessed BOW and sentence order in RCT abstracts. Applying Maximum Entropy and Hidden Markov Model (HMM) augmentation, they achieved an accuracy of 92.8% for sentences in structured abstracts. Lin et al. [21] captured discourse transitions from section to section with content structure model, and section heading. Applying HMM to the RCT abstracts, they achieved an accuracy of 82.1% for structured abstracts and 79.4% for unstructured abstracts. Chung exploited unigram, POS, sentence location, and features from previous-next sentences in order to extract key sentences from the RCT abstracts. Applying CRF, Chung [22] achieved the per-sentence accuracy of 94.2% in structured abstracts, and 87.6% in unstructured abstracts.

In summary, the commonly used features in previous studies are BOW and sentence location. Although they contribute to recognizable effectiveness, previous attempts hardly reflect characteristics of the biomedical domain, and only apply the approach of computer science. Linguistic studies suggest that verbs and n-grams (i.e., lexical bundles, multi-word patterns, or clusters) may be important classification features in that they help to fulfill communicative purpose of each section; different verbs and n-grams predominantly occur at different discourse levels [23–28].

Cortes [27] analyzed the relationship of the n-grams to the moves in the introduction section of research article. Some n-grams made of more than five words (e.g., 'the purpose of this study was') triggered the communicative function of a move (describes the objective of study). Williams [24] also classified verbs into seven functional categories and analyzed their distribution by IMRAD section. Observation verbs were either mainly active (show, present, follow) or almost exclusively passive (find, observe, see, demonstrate). About 70% of Relations verbs (follow, compare, relate) were located in the results and discussion section. Hanania and Akhtar [23] analyzed the voice, tense, aspect, and modality of finite verbs in five rhetorical sections (Introduction, Review, Methods, Results, Discussion) of three natural science texts. Passive and past verbs were predominant in Methods; the present tense was frequently used in Introduction; the modal verbs were highest in Discussion and lowest in Methods; perfective and progressive verbs were less utilized uniformly.

Our study could be differentiated from previous findings in that we extract distinguishing linguistic features, such as n-gram, verb phrase, and noun-phrase, from each IMRAD section of large-scale biomedical abstracts, and utilize the features for efficient classification. We propose the new feature construction method using the linguistic features and identify which feature and combinations produce the best results in sentence classification.

3. Methods

To identify which feature(s) performs best in sentence classification in terms of effectiveness, sets with all features were prepared. The universe included Bag-of-words (BOW), linguistic features, grammatical features, and structural features. Fig. 1 illustrates the workflow in this study.

3.1. Data preparation

We used PubMed Central (PMC) Open Access Subset for data mining [29], which contains 536,682 articles (as of Nov 23, 2012). The number of articles having structured abstracts was 160,150 that accounted for 29.73% of the corpus. From the structured abstracts, research-type article abstracts (n=152,083) were collected as training data by selecting the articles with article-type "research-article". The remaining unstructured research-type abstracts (n=221,261) were used as the test dataset.

3.1.1. Sentence splitting and POS tagging

Sentences were split and POS tagged with the LingPipe of the MedPost [30] Parser, which is one of the popular natural language processing tools written in Java [31,32]. The total number of sentences in the dataset was 1,694,998 with a mean of 11.15 and a standard deviation of 3.70 per abstract. Each sentence was appended to the corresponding section heading.

3.1.2. Normalization of section headings

The structured abstract corpus has various section headings (n=1628). These variations include plurals (e.g., 'Conclusion', 'Conclusions'), modifiers (e.g., 'Conclusion', 'Major Conclusion'), different word sequences (e.g., 'Conclusions and Significance', 'Significance and Conclusion'), and combined section headings (e.g., 'Method and Result', 'Result and Discussion'). Different section headings were merged into 1001 headings using Open Refine [33], a tool for cleaning and transforming messy data, and grouped into the representative section headings. Afterward, the top 50 most frequent headings (98.97% of all merged headings) were normalized to the IMRAD sections based on the NLM mapping list [34]. The mapping list assigns section headings of the PMC article abstracts to one of the five standard headings (i.e., Objective, Background, Methods, Results, and Conclusion). In our study, 'Objective' and 'Background' were combined into 'Introduction', and 'Conclusion' was renamed as 'Discussion'.

3.2. Feature selection

The features were categorized into linguistic features, BOW,