Identifying Finding Sentences in Conclusion Subsections of Biomedical Abstracts

Yingya Li and Bei Yu(✉)

Syracuse University, Syracuse, NY 13244, USA
{yli48, byu}@syr.com

Abstract. Segmenting scientific abstracts and full-text based on their rhetorical function is an essential task in text classification. Small rhetorical segments can be useful for fine-grained literature search, summarization, and comparison. Current effort has been focusing on segmenting documents into general sections such as introduction, method, and conclusion, and much less on the roles of individual sentences within the segments. For example, not all sentences in the conclusion section are describing research findings. In this work, we developed rule-based and machine learning methods and compared their performance in identifying the finding sentences in conclusion subsections of biomedical abstracts. 1100 conclusion subsections with observational and randomized clinical trials study designs covering five common health topics were sampled from PubMed to develop and evaluate the methods. The rule-based method and the bag-of-words based machine learning method both achieved high accuracy. The better performance by the simple rule-based approach shows that although advanced machine learning approaches could capture the main patterns, human expert may still outperform on such a specialized task.

Keywords: Text classification · Rule-based approach · Machine learning · Biomedicine

1 Introduction

Categorizing sentences by their rhetorical functions is an important task in literature mining. It is particularly useful for the fields that face the challenge of overwhelming volume of publications. For example, identifying the results in empirical studies is a critical step for writing systematic reviews in Evidence-Based Medicine (EBM) [11]. It is also a step toward further analyses, such as identifying potential exaggerations in conclusions [6, 8].

To date many studies have tried to automatically segment sentences either in abstracts or full-texts into sections (e.g., [2, 7, 13, 15, 20]). Nearly all existing studies focus on the general rhetorical level of sentences in the given contexts, classifying abstracts or full text into introduction, method, result and discussion (IMRaD) format. However, simply classifying sentences into the IMRaD structure does not provide adequate granularity for retrieving key information such as study findings, because sentences in each subsection may still serve different rhetorical functions. For instance, sentences in the conclusion subsection can describe studies' findings, limitations, or

N. G. Taylor et al. (Eds.): iConference 2019, LNCS 11420, pp. 679–689, 2019.
https://doi.org/10.1007/978-3-030-15742-5_64

implications for future studies respectively. As shown in the following excerpt, the underscored sentence is the finding of the study, while the second and third sentences represent the implications for future studies.

> *The present meta-analysis suggests that insulin therapy may increase the risk of CRC. More prospective cohort studies with longer follow-up durations are warranted to confirm this association. Furthermore, future studies should report results stratified by gender and race and should adjust the results by more confounders. (PMID 25099257)*

In this work, we focus on the automatic categorization of sentences in conclusion subsections from structured biomedical abstracts. Our goal is to determine whether sentences in conclusions subsections describe study *findings*, as opposed to the *non-finding* ones that describe study implications, limitations, recommendations, and clinical trial registration information.

Rule-based and machine learning methods are the two mainly used approaches in prior sentence categorization studies. Most studies found the machine-learning approaches using features of bag-of-words, semantic relations or structural information of sentence positions work well (e.g., [2, 11, 13, 15]); however, studies also suggest that for texts with controlled vocabularies, rule-based approaches can also be effective [5, 10]. Therefore, we developed both rule-based and machine learning methods and compared their performance in identifying the *finding* sentences in conclusion sections. We used 1000 biomedical abstracts from PubMed as training and 100 abstracts for testing to validate these two methods.

2 Related Work

The task of identifying *finding* sentences in conclusion subsections of abstracts is closely related to automatic section identification and summarization for scientific articles. To realize the automatic process, many studies have aimed at developing schemas and corpora for categorization (e.g., [21, 23, 25]). For example, Teufel and Moens [24] introduced a scheme of Argumentative Zoning (AZ) which classifies sentences in scientific text into categories such as aim, background, own, contrast, and basis on their rhetorical status in scientific discourse. Their experiments suggest that the proposed AZ framework can be used to identify and summarize novel contributions and backgrounds of scientific articles. Liakata et al. [17] proposed two classification schemas to capture the hypotheses, motivations, methods, conclusions etc. based on the rhetorical nature of 265 full papers in physical chemistry and biochemistry. Guo et al. [12] compared the validity of three pre-existing categorization schemas developed from full-text articles on the abstracts of cancer risk assessments. Their results suggest the possibility of applying full-text sentence categorization schema on abstracts.

Previous studies typically modeled section identification process as text classification task, which determines a pre-defined label to each individual sentence based on their rhetorical function. In this line of method, classifiers such as Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) are widely used (e.g., [2, 11, 12, 23]). Studies reached different conclusions regarding the performance of the classifiers. For example, Agarwal and Yu [2] trained both MNB and SVM classifiers to identify

sentences of articles sampled from BioMed Central into IMRaD structures. Their results show that MNB performed better than SVM at classifying sentences using bag-of-words with enriched features of words tenses. Gabb et al. [11] compared the performance of models using Naïve Bays (NB) and SVM to automatically identify result sentences from full-text journal articles. Their experiment results indicate that though models built with NB and SVM obtain similar results, classifier trained with SVM using top-100 terms and sentence locations as features tended to have slightly higher F1 scores. In addition to modeling sentence type or section identification as text classification task, some studies modeled the structure of article sections as a sequence labeling problem. Hirohata et al. [13] used n-gram, relative sentence location, and the features from previous and proceeding sentences for text representation to classify the sections of academic abstract into objectives, methods, results, and conclusions using Conditional Random Fields (CRFs). The feature sets of ngram with surrounding sentence features trained with CRFs outperformed the model using SVM. However, these studies used different dataset to develop their methods, thus classification results may not be necessarily comparable to each other.

Other than these machine learning approaches, previous studies also applied rule-based methods to identifying sections or to similar biomedical applications. Friedman et al. [10] built a language processor which relied on semantic grammars to extract clinical information in patient documents and mapped them into controlled vocabulary terms. Chapman et al. [5] developed a simple algorithm for identifying negated sentences in discharged summaries. They implemented a sets of negation phrases with regular expression to detect a large portion of pertinent negatives in the document. Yu et al. [26] relied on a set of pattern-matching rules for mapping an abbreviation in biomedical articles into its full forms. Kilicoglu et al. [14] applied a rule-based approach to automatically recognize self-acknowledged limitations in clinical research publications. The success of these studies indicates that for text with controlled language indicators and patterns, the simple rule-based approach might be an effective one. In recent years, the lexicon-enhanced approach also shows its success in other NLP applications such as sentiment analysis [3, 16], and emotion-detection [4].

Different from previous studies of classifying scientific abstracts and full-text into the general categories, our work intends to identify sentences about study findings from the conclusion subsections in biomedical abstracts. We took advantage of the lexicon-enhanced rule-based NLP approach for sentence type recognition and compared its performance to the commonly used machine learning methods in this work.

3 Method

In this section, we first introduce the process of corpus construction for training and testing (Sect. 3.1). Then we discuss about the rule-based approach (Sect. 3.2), the machine learning classification methods (Sect. 3.3) and evaluation measures used in this study.

3.1 Corpus Construction

To the best of our knowledge, no prior corpus is available for *finding* sentence identification, we then used structured abstracts from original scientific research papers for corpus construction. We focused on the biomedical domain and chose PubMed as the source. The PubMed database has more than 27 million citations for biomedical literature from Medline, life science journals, and online books. Since different study designs may use varied language to describe their research, we applied stratified sampling approach to collect both observational and randomized clinical trials (RCTs) studies from the platform. PubMed's search interface provides one search criterion "Publication Type" that is derived from the MeSH terms for PubType in MEDLINE records (2018MeSH). We applied article's MeSH term to select the RCTs articles (MeSH Unique ID: D016449); used the searching method introduced in [1] to collect case-control, cross-sectional, retrospective, and prospective studies within observational studies; and rescanned the abstracts sections with the keywords to exclude the irrelevant ones. To account for the vocabulary variation among different health issues, we selected five common health topics – nutrition, diabetes, obesity, breast cancer, and cholesterol. The whole downloaded set contains 63498 conclusion subsections from structured abstracts in total. The XML files in PubMed contain occasional parsing errors, so sometimes the conclusion subsections may include paragraphs in the following sections. For quality control purpose, we used the Stanford CoreNLP tool [18] to split the conclusion subsections into sentences, and removed the articles with conclusions longer than four sentences. We then sampled equal number of articles with conclusion length as 1, 2, 3, or 4 sentences from the 63498 set. Our sampled corpus includes 1100 structured abstracts, within which we used 1000 as the training set and 100 as the testing set.

To construct a reliable human-annotated dataset to serve as ground truth, we annotated each sentence in the selected corpus as either *category-0* (*non-finding*) or *category-1 (finding).* Table 1 shows examples of the two sentence types. *Category-0* refers to *non-finding* sentences (as shown in Examples 1 and 2); while *category-1* refers to sentences explicitly talking about study *finding* (as shown in Examples 3 and 4). An inter-coder reliability test on a sample of 200 articles with 510 sentences from the training set showed almost perfect agreement of the schema. Specifically, two graduate students with the education background of information studies labelled the sentences extracted from the conclusion subsection of the structured abstracts. Annotators identified the sentence category based on their linguistic indicators. We applied Cohens Kappa k as the inter-coder agreement measure [9]. Kappa values of .61 or above are considered as substantial agreement and .81 or above as almost perfect agreement [19]. The overall k value was .85, indicating the annotation schema for finding sentence identification reached almost perfect inter-coder agreement (Table 2 shows the detailed inter-coder agreement). Disagreements in the annotation were later resolved by the two annotators through discussion.

Annotator 1 annotated the conclusion subsection of the rest 900 articles from both training and testing sets. The final corpus contained 2735 annotated sentences in the conclusion subsections of abstracts from 1100 articles, of which 711 sentences in the conclusion subsections belonged to *category-0*, and 2024 sentences belonged to *category-1*. Table 3 shows the number of sentences per category in the developed corpus.

Table 1. Finding and non-finding sentences from conclusion subsections.

Sentence	Annotation
Example 1: (PMID: 28640840) We propose a novel AI disease-staging system for grading diabetic retinopathy that involves a retinal area not typically visualized on fundoscopy and another AI that directly suggests treatments and determines prognoses	Category-0
Example 2: (PMID: 26504068) This approach may, however, be difficult to implement on a large scale	Category-0
Example 3: (PMID: 28953631) The results of this study showed that TPVBRA combined with bupivacaine and dexmedetomidine can enhance the duration and quality of analgesia without serious adverse events	Category-1
Example 4: (PMID: 28746662) Our current analysis does not support the existence of an association between age at first childbirth and adult-onset diabetes among postmenopausal women, which had been reported previously	Category-1

Table 2. Confusion matrix for the sentence type annotation.

Kappa = .85		A2		
		Category-0	Category-1	All
A1	Category-0	163	11	174
	Category-1	23	313	336
	All	186	324	510

Table 3. Sentence type distribution in annotated corpus.

Dataset	Category-0	Category-1	Total
Training set	659	1841	2500
Testing set	52	183	235
Total	711	2024	2735

3.2 A Rule-Based Approach for Automatic Finding Sentence Identification

We framed the automated identification process as a sentence-level text classification task, and manually identified rules indicative of sentence types. A total of 181 rules were derived from training set, of which 14 were study backgrounds, 13 were study limitations, 59 were implications, 84 were recommendations, and 11 ones were about the clinical trial registration. These rules were generated based on iterative rounds of keywords searching and pattern matching for the linguistic indicators of study backgrounds, limitations, implications, recommendations, and information of clinical trial registration. Similar as the existing rule-based approaches [2, 5], we used regular expression to identify those generated patterns in the original annotated sentences.

We looked for keywords like *"exist in literature"*, *"growing literature"*, *"literature to date"* for introduction of study backgrounds; indicators like *"limited by"*, *"limitations"* for study limitations; phrases such as *"further assessment"*, *"future studies"*, *"future research"*, *"follow-up exploration"* for the implications of current study findings for future explorations; expressions of *"clinicians should"*, *"health policy makers should"*, *"actions should focus on"* for the recommendations of practitioners and experiments alike; and words like *"trial registration"*, *"clinical trial registration"* for the information of clinical trial registration in sentences of conclusion subsections.

All identified rules were grouped into two sets. The first set included 153 short language patterns represented by regular expressions, which were short terms and keywords indicative of *non-finding* sentences (e.g., *"future research"*, *"further investigation"*, *"other studies"*, *"should confirm these findings"*). The second set contained 28 rules that captured longer language patterns describing *non-finding* sentences. For example, one rule in the second set is that if a sentence has phrases of *"is warranted"* or *"are warranted"*; and it does not have conjunctions of *"although"* or *"though"*, it is a *non-finding* sentence. We applied all rules into a rule-based classifier, detecting the category of each input sentence. If a sentence matches any of the rules in the first pattern set, it will be assigned to *category-0* as a *non-finding* sentence; else the classifier will continue to check if the sentence matches any of the other rules in the second set. If the input sentence does not match any of the identified 181 patterns in the first and second pattern sets, it will then be assigned to *category-1*; namely the sentence depicts the study findings. We used macro-averaged precision, recall and F1 scores to evaluate the performance of the proposed rule-based approach on the testing dataset.

3.3 Machine Learning Approaches for Automatic Finding Sentence Identification

We measured the performance of machine learning approaches using variations of bag-of-words representations, and the language indicators from the identified 181 rules as features. For the bag-of-words representations, we chose NB and SVM algorithms with different vectorization methods and enriched features to train the sentence type classifiers, using Scikit-learn python package [22]. NB and SVM are the most popular classification algorithms in current studies of segmenting scientific abstracts and full-text [2, 11, 12]. We used two NB algorithms – multivariate Bernoulli model and the multinomial model. The first one uses word presence and absence as feature value (BNB); while the second one uses word frequency (MNB). For SVM, we combined the SVM (Liblinear) algorithm with three different frequency measures – word presence and absence (SVM-boolean), word frequency (SVM-tf), and word frequency weighted by inverse document frequency (SVM-tfidf).

To further validate the performance of syntactic and semantic structures in classification, we extracted part-of-speech (POS) and dependency parsing from the input sentences using Stanford CoreNLP [18]. The bag-of-words machine-learning approaches then contained the following four feature vectors with different representation methods: (1) simple bag-of-words; (2) bag-of-words with POS tagging; (3) bag-of-words with enhanced dependency parsing; (4) bag-of-words enriched with both POS tagging and enhanced dependency parsing (combining features in (2) and (3) together). For example,

***Original sentence**: Physical activity is also associated with favorable HDL-C. (PMID: 28167327)*
***Bag-of-words**: Physical, activity, is, also, associated, with, favorable, HDL-C*
***Bag-of-words with POS tagging**: Physical-JJ, activity-NN, is-VBZ, also-RB, associated-VBN, with-IN, favorable-JJ, HDL-C-NN*
***Bag-of-words with enhanced dependency parsing**: amod(activity-2, Physical-1) nsubjpass(associated-5, activity-2) auxpass(associated-5, is-3) advmod(associated-5, also-4) root(ROOT-0, associated-5) prep(associated-5, with-6) amod(HDL-C-8, favorable-7) pobj(with-6, HDL-C-8)*

For the machine learning approach based on language indicators from the hand-crafted rules as features, the presence or absence of the identified 181 patterns in the rule-based approach was used in training the classifier. We applied the Decision Tree classifier in Scikit-learn [22] with its default parameter settings as the implementation of the Decision Tree algorithm and compared its performance to the BNB and SVM algorithms using the same representation.

Considering the size of current dataset and the imbalance distribution of *category-0* and *category-1*, we used 10 folds cross-validation for the evaluation of machine learning approaches and reported precision, recall, F1 scores of each category, in addition to the macro-averaged precision, recall, and F1 scores.

4 Result

The majority vote baseline of the test set is .78. Among the three approaches, our experiment result shows that the rule-based method achieved the best performance with an macro-averaged F1 score at .96 level on the test set (as shown in Table 4).

Table 4. Performance of the rule-based model.

Method	Sentence type	Accuracy	Precision	Recall	F1 Score
Rule-based	Category-0	.90	.92	.90	.91
	Category-1	.98	.97	.98	.98
	Macro-averaged	**.96**	**.96**	**.96**	**.96**

The machine learning models based on bag-of-words feature also achieved high performance. The best machine learning model is BNB with unigram and bigram features with a macro-averaged F1 score at .86 level, lower than the .96 by the rule-based model. Tables 5 and 6 list the feature engineering options and results. Table 5 lists the results of unigram experiments. BNB, MNB, and SVM-tfidf have very similar macro-averaged scores across the two sentence type categories, but BNB has slightly higher macro-averaged precision (.84) and recall (.85) values. Table 6 shows that adding bigram features slightly improves the performance of all models except SVM-tf and SVM-tfidf. As shown in Table 6, BNB with unigram and bigram bag-of-words representation has the highest precision (.86), recall (.87) and F1 scores (.86) among all

the machine-learning methods using the same representation. It also has higher accuracy than its counterpart in the bag-of-words unigram experiment and the best performance in bag-of-words unigram representation.

Table 5. Performance of the machine learning models based on unigram features.

Method	Sentence type	Accuracy	Precision	Recall	F1 score
BNB	Category-0	.79	.76	.79	.77
	Category-1	.91	.93	.91	.92
	Macro-averaged	**.88**	**.84**	**.85**	**.85**
MNB	Category-0	.74	.79	.74	.76
	Category-1	.93	.91	.93	.92
	Macro-averaged	.88	.85	.83	.84
SVM-boolean	Category-0	.76	.73	.76	.74
	Category-1	.91	.91	.90	.91
	Macro-averaged	.86	.82	.83	.82
SVM-tf	Category-0	.75	.73	.75	.74
	Category-1	.90	.91	.90	.91
	Macro-averaged	.86	.82	.83	.82
SVM-tfidf	Category-0	.70	.82	.70	.76
	Category-1	.95	.90	.95	.92
	Macro-averaged	.88	.86	.82	.84

Adding the enriched features introduced in Sect. 3.3 did not further improve the performance of the machine learning models. Among all the models with enriched features, SVM-tfidf using unigram bag-of-words feature with dependency parsing

Table 6. Performance of the machine learning models based on unigram and bigram features.

Method	Sentence type	Accuracy	Precision	Recall	F1 score
BNB	Category-0	.82	.78	.82	.80
	Category-1	.92	.94	.92	.93
	Macro-averaged	**.89**	**.86**	**.87**	**.86**
MNB	Category-0	.75	.80	.75	.77
	Category-1	.93	.91	.93	.92
	Macro-averaged	.89	.86	.84	.85
SVM-boolean	Category-0	.75	.75	.75	.75
	Category-1	.92	.91	.91	.91
	Macro-averaged	.87	.83	.83	.83
SVM-tf	Category-0	.75	.74	.75	.74
	Category-1	.91	.91	.91	.91
	Macro-averaged	.86	.82	.83	.82
SVM-tfidf	Category-0	.72	.84	.72	.78
	Category-1	.95	.91	.95	.93
	Macro-averaged	.89	.87	.84	.85

relations had the best performance. However, it did not outperform the best machine learning model with only bag-of-words feature.

The machine learning models based on the language indicators from the hand-crafted rules did not perform the ruled-based method. Table 7 shows the model using Decision Tree has the best macro-averaged precision (.96), recall (.88) and, F1 scores (.91).

Table 7. Performance of using language indicators from hand-crafted rules as features.

Method	Sentence Type	Accuracy	Precision	Recall	F1 Score
Decision Tree	Category-0	.77	.99	.77	.87
	Category-1	.99	.92	.99	.96
	Macro-averaged	**.94**	**.96**	**.88**	**.91**
BNB	Category-0	.68	.99	.68	.81
	Category-1	.99	.90	.99	.95
	Macro-averaged	.92	.94	.84	.88
SVM	Category-0	.73	.99	.73	.84
	Category-1	.99	.91	.99	.95
	Macro-averaged	.93	.95	.86	.90

5 Discussion

In our experiment, the rule-based method and the bag-of-words based machine learning method both achieved high accuracy, suggesting that the two approaches are effective in identifying *finding* sentences in conclusion subsections of structured abstracts. The high precision, recall and F1 scores of the rule-based approach on the testing set confirm our previous assumption that the rules developed based on linguistic indicators and patterns of sentences in conclusion subsections are more effective to identify *finding* sentences extracted from structured abstracts.

In comparison, machine learning models based on bag-of-words representations and indicators in identified rules as features tend to have higher precision and recall values in classifying *category-1*, but relatively lower values in *category-0*. Feature analyses of the best machine learning method using bag-of-words representation indicate that the classifier has learned some basic sentence type indicators like *"further studies"*, *"further research"*, *"future studies"*, *"larger studies"*, *"associated with"*. However, it was not able to learn linguistic patterns capturing larger language units in *non-finding* sentence as included in the second set of identified rules.

Compared to the rule-based approach, the machine learning models based on linguistic indicators in rules as features are more sophisticated in the process of deciding sentence types. Feature ranking result of the most important features learned by the Decision Tree model shows that rules of the clinical trial registration information, and implications of future studies are the most important ones, thus the model has learned some patterns on *non-finding* sentences. However, this more complicated model did not outperform the simpler rule-based model.

Though the rule-based approach achieved satisfactory results detecting *finding* sentences, error analyses of the misclassified cases suggest room for improvement. The most common error can be attributed to the confounding keywords in finding sentences: a finding sentence can mention both study findings and the implication for future study. The keywords of future study implications will then lead to detection error. On the other hand, current rules can capture the *non-finding* sentences which contain explicit language indicators. However, for *non-finding* sentences lacking the clear cues like indications of study limitations (e.g., *"limitations", "limited by"*) or recommendations (e.g., *"these findings suggest that", "should be introduced to"*), such as "*Such information is crucial to target Web-based support systems to different patient groups*", the rules would not be able to capture them.

6 Conclusion

In this work, we focus on detecting *finding* and *non-finding* sentences from the conclusion subsections of structured abstracts. The rule-based method and the bag-of-words based machine learning method both achieved high accuracy. The better performance by the simple rule-based approach shows that although advanced machine learning approaches could capture the main patterns, human expert may still outperform on such a specialized task. For text with controlled linguistic patterns, the rule-based one could be more suitable. Considering the errors caused by the current rules, in future work we will conduct deeper semantic analysis on the generated rules to either introduce more synonyms of identified keywords or to prevent the confounding effects of those patterns for higher precision and recall during the classification. Meanwhile, we will further explore the effectiveness of this rule-based approach for finding sentence recognition in unstructured abstracts.

Acknowledgement. We would like to thank Shiqi Qu who have contributed to the inter-coder agreement checking and corpus construction.

References

1. Search strategies: Study Type public health: Search strategies by study type. http://libguides.adelaide.edu.au/c.php?g=165091p=5799888. Accessed 2 Jan 2018
2. Agarwal, S., Yu, H.: Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. Bioinformatics **25**(23), 3174–3180 (2009)
3. Asghar, M.Z., Khan, A., Ahmad, S., Qasim, M., Khan, I.A.: Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. PLoS ONE **12**(2), e0171649 (2017)
4. Asghar, M.Z., Khan, A., Bibi, A., Kundi, F.M., Ahmad, H.: Sentence-level emotion detection framework using rule-based classification. Cogn. Comput. **9**(6), 868–894 (2017)
5. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Inform. **34**(5), 301–310 (2001)
6. Chiu, K., Grundy, Q., Bero, L.: Spin in published biomedical literature: a methodological systematic review. PLoS Biol. **15**(9), e2002173 (2017)

7. Chung, G.Y.: Sentence retrieval for abstracts of randomized controlled trials. BMC Med. Inf. Decis. Making **9**(1), 10 (2009)
8. Cofield, S.S., Corona, R.V., Allison, D.B.: Use of causal language in observational studies of obesity and nutrition. Obes. Facts **3**(6), 353–356 (2010)
9. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
10. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. J. Am. Med. Inform. Assoc. **1**(2), 161–174 (1994)
11. Gabb, H.A., Lucic, A., Blake, C.: A method to automatically identify the results from journal articles. In: iConference 2015 Proceedings (2015)
12. Guo, Y., Korhonen, A., Liakata, M., Karolinska, I.S., Sun, L., Stenius, U.: Identifying the information structure of scientific abstracts: an investigation of three different schemes. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pp. 99–107. Association for Computational Linguistics (2010)
13. Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M.: Identifying sections in scientific abstracts using conditional random fields. In: Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I (2008)
14. Kilicoglu, H., Rosemblat, G., Malički, M., ter Riet, G.: Automatic recognition of self-acknowledged limitations in clinical research literature. J. Am. Med. Inform. Assoc. **25**(7), 855–861 (2018)
15. Kim, S.N., Martinez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support evidence based medicine. BMC Bioinf. **12**, S5 (2011). BioMed Central
16. Kundi, F.M., Khan, A., Ahmad, S., Asghar, M.Z.: Lexicon-based sentiment analysis in the social web. J. Basic Appl. Sci. Res. **4**(6), 238–248 (2014)
17. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R., et al.: Corpora for the conceptualisation and zoning of scientific papers. In: LREC. Citeseer (2010)
18. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford coreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
19. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica **22**(3), 276–282 (2012)
20. McKnight, L., Srinivasan, P.: Categorization of sentence types in medical abstracts. In: AMIA Annual Symposium Proceedings, vol. 2003, p. 440. American Medical Informatics Association (2003)
21. Mizuta, Y., Korhonen, A., Mullen, T., Collier, N.: Zone analysis in biology articles as a basis for information extraction. Int. J. Med. Inf. **75**(6), 468–487 (2006)
22. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
23. Ruch, P., et al.: Using argumentation to extract key sentences from biomedical abstracts. Int. J. Med. Inf. **76**(2–3), 195–200 (2007)
24. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Comput. Linguist. **28**(4), 409–445 (2002)
25. Teufel, S., Siddharthan, A., Batchelor, C.: Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 3, pp. 1493–1502. Association for Computational Linguistics (2009)
26. Yu, H., Hripcsak, G., Friedman, C.: Mapping abbreviations to full forms in biomedical articles. J. Am. Med. Inform. Assoc. **9**(3), 262–272 (2002)