Automatically Classifying Sentences in Full-Text Biomedical Articles into Introduction, Methods, Results and Discussion

Shashank Agarwal, MS,[1] Hong Yu, PhD[1,2]

[1]Medical Informatics and [2]Departments of Health Sciences and Computer Science, University of Wisconsin-Milwaukee, Wisconsin

Abstract

Biomedical texts can be typically represented by four rhetorical categories: introduction, methods, results and discussion (IMRAD). Classifying sentences into these categories can benefit many other text-mining tasks. Although many studies have applied approaches to automatically classify sentences in MEDLINE abstracts into the IMRAD categories, few have explored the classification of sentences that appear in full-text biomedical articles. We explored different approaches to automatically classify a sentence in a full-text biomedical article into the IMRAD categories. Our best system is a support vector machine classifier that achieved 81.30% accuracy, which is significantly higher than baseline systems.

1 Introduction

Previous studies have concluded that biomedical texts typically fall into the rhetorical categories of *introduction, methods, results* and *discussion* (IMRAD) (e.g., (1-4)). For example, the following is a paragraph from the results section of a full-text article (5) in which the sentences fall into the IMRAD categories (italic represents *introduction*, underscore represents *methods,* bold represents *results*, and italic-underscore represents *discussion*).

> "*PECAM-1 plays an important role in endothelial cell-cell and cell-matrix interactions, which are essential during vasculogenesis and/or angiogenesis (17, 22).* Here, we examined expression of PECAM-1 mRNA in vascular beds of various human tissues and compared it with expression of PECAM-1 in human endothelial and hematopoietic cells. **A short exposure of the blot probed with GAPDH is shown, because poly(A)$^+$ RNA from the cell lines gives a strong signal within several hours compared with the total RNA from human tissue. Therefore, total RNA from various tissues required a much longer exposure to reveal GAPDH mRNA.** *Human tissue and cell lines expressed multiple RNA bands for PECAM-1, which may represent alternatively spliced PECAM-1 isoforms, the identity of which required further analysis*."

In this study we report our efforts on computationally classifying biomedical texts into the IMRAD categories. Our work may benefit many other text-mining tasks. For example, information extraction (e.g., extracting protein-protein interactions and information relating the interaction network to phenotype) may target evidence-rich *results*, and avoid evidence-lean *introduction*. Summarization may aggregate sentences and provide a summary for each rhetorical category. For example, our work shows that biomedical research scientists prefer to have the IMRAD structure for summarizing the content of a figure (6). Question answering may target on different rhetorical categories for answer extraction. For example, definitions may be extracted from *introduction* (7), and *methods* may be the choice for answering questions such as "how to perform a glucose uptake assay?"

The importance for automatically classifying biomedical text into the rhetorical-zone categories has been recognized and various approaches have been developed to automate the task, although most of the efforts have been made to develop approaches for assigning IMRAD categories to sentences that appear in MEDLINE abstracts (7-8).

McKnight and Srinivasan (8) reported the first automation. They trained supervised machine-learning binary-classifiers on structured abstracts (i.e., the sentences in an abstract have been structured by the authors of the abstract into the IMRAD categories). The trained classifiers were then used to predict the categories of sentences in unstructured abstracts. The authors observed that sentences typically followed the IMRAD order in an abstract, and therefore incorporated sentence positions as additional features. They reported F-scores of 52–79% for assigning each sentence to the IMRAD categories. Lin et al (9) further employed hidden Markov models, which maximized the position feature and improved the binary classification to F-scores of 73–89%. No work has attempted to predict IMRAD categories of sentences in full-text biomedical articles.

Mizuta et al. (10) examined full-text biomedical articles, explored linguistic features, and defined richer rhetorical zone categories that include *problem-setting* (i.e., the problem to be solved; the goal of the present work), *insight* (i.e., the author's insights and findings obtained from experimental results), etc. Using 20 annotated full-text articles, supervised machine-learning classifiers (i.e., naïve Bayes and support vector machines) were developed for the automation (11). The features included lexical, syntactic, location, and zone sequence. Their best performing system, one that incorporated all the features, achieved an F-score of 70% for all category classification.

Other related work includes Shatkay et al (12, 13). They built a multi-dimensional classifier, where each sentence was classified on five parameters: *focus, certainty, evidence, polarity* and *direction/trend*. The classifier was trained on 10,000 annotated sentences that were selected from full-text biomedical articles, and achieved good performances.

Here, we present our work for automatically classifying sentences appearing in full-text biomedical articles into the IMRAD categories. We have explored rule-based and machine-learning approaches.

2 Methods

We explored rule-based and machine-learning approaches to automatically classify a sentence into the IMRAD categories.

2.1 A Baseline System

As a baseline, we create a simple system (*Baseline*) that assigns a sentence an IMRAD category based on which IMRAD section the sentence occurs in. For example, we assign all sentences in the Introduction section the category *introduction.*

2.2 A Rule-based System

Rule-based systems have shown success in the biomedical domain (e.g., (14, 15)). We randomly selected eight articles from the TREC Genomics Track text collection (16), which contains more than 160,000 full-text biomedical articles. The eight articles contain ~30,000 words and 1,250 sentences. The first author of this paper (SA) read each article and then manually identified patterns that were indicative of the IMRAD categories. For example, one rule links a sentence to *discussion* if the sentence incorporates the words 'our,' 'observations,' and 'suggests' and the sentence does not associate with a citation. A total of 603 rules were identified, of which 410 were *methods* rules, 96 were *results* rules and 97 were *discussion* rules. If a sentence was not identified by any of the *methods*, *results* or *discussion* rules, then that sentence was labeled as *introduction*. We then implemented the rules in a rule-based classifier that automatically assigns sentences to the appropriate category.

2.3 Supervised Machine-Learning Systems Trained on Non-Annotated Corpus

Annotation has always been an expensive process. Therefore, we explored methods for training supervised machine-learning systems on non-annotated data. Our work is inspired by the work of Yu and Hatzivassiloglou (17). We assume that in a structured IMRAD full-text article, the majority of sentences in each section are classified with the respective IMRAD category. For example, even though the sentences under the Introduction section incorporate other categories, we assume that a majority of the sentences are still assigned *introduction*.

We developed four classifiers. The first classifier, *Non1*, was trained on structured sentences from the full-text article that incorporates the test sentence. The IMRAD category of the sentences in the full-text was used as the label of the sentence to build the classifier. Since our training data are noisy, the second classifier, *Non2*, incorporated an iterative classification process that attempts to remove the noisy data from the training set. This classifier was based on the work of Yu and Hatzivassilogou (17). Specifically, for each full-text document, we built the classifier C_1, which was trained on the sentences within the four structured sections. We then applied the same classifier to predict the category of sentences in the training data and then removed those contradictory predictions. We assume that C_1 performs better than random and therefore has a better chance than random to remove noisy training data. We then continued the iteration $C_i,$ $i=1, 2\dots N,$ until the accuracy dropped or stabilized.

Non3 was trained on structured MEDLINE abstracts. We considered an abstract to be structured if it contained the four IMRAD categories or their synonyms (for example, *background* was assigned as *introduction*). 8000 randomly selected sentences (2000 for each category) from the structured abstracts in MEDLINE were aggregated to train the classifier.

Non4 was trained on structured full-text sentences instead of abstract sentences. 8000 sentences (2000 from each category) from the IMRAD categories were randomly collected from full-text articles in the BioMed Central corpus (available at http://www.pubmedcentral.nih.gov/) and used to train the classifier. Unlike *Non1*, *Non4* was trained on sentences from randomly selected articles, whereas

Non1 was trained on sentences from the same article as the test sentences.

2.4 Supervised Machine-Learning System Trained on Manually Annotated Full-Text Sentences

Finally, we manually annotated a set of sentences that appear in full-text biomedical articles and then trained a supervised machine-learning system on the annotated data. We call this classifier *Man*. Feature selection and machine-learning systems are described in the following section. The annotated data will be described in Section 3.2.

2.5 Machine-Learning Systems and Features

For all supervised classifications, we used the support vector machines provided by the open-source Java™-based machine-learning library Weka 3 (http://www.cs.waikato.ac.nz/ml/weka/). The features we explored include words and n-grams. We found that a combination of individual words, bigrams and trigrams led to the best performance. We observed that citations can be an important feature. For example, citations are more frequently introduced in *introduction* than in *results.* We therefore created a new feature to indicate the presence of a citation. All numbers were replaced by a unique symbol.

Biomedical texts frequently report existing knowledge in the present tense and the experimental results in the past tense. We therefore added the presence of these two verb tenses as additional features. We used the Stanford parser (http://nlp.stanford.edu/software/lex-parser.shtml) for identifying the presence of the verb tenses. A final feature we explored is the IMRAD categories inherited in a structured full-text article. This feature was only added in the machine-learning classifier *Man* that was trained on the annotated sentences.

We applied mutual information (18) for feature selection. We experimented with a number of features and found that the top-1000 tended to give a better performance.

3 Evaluation

For each classifier, we report the accuracy (i.e., number of correctly predicted sentences divided by total number of sentences), and F measure, which is the harmonic mean of precision and recall. Here recall is the number of correctly predicted sentences divided by the total number of sentences in the same category, and precision is the number of correctly predicted sentences divided by the total number of predicted sentences in the same category.

3.1 Data

The publicly available BioMed Central full-text corpus was used for this study. We randomly selected 148 articles that incorporate the IMRAD sections in the full-text body and then randomly selected five sentences from each category of these articles. This resulted in a total of 2,960 sentences (148×5×4), from which we further annotated a gold standard set.

3.2 Annotation, Agreement, and Gold Standard

The first author of this paper (Annotator1) developed an annotation guideline and, using the guideline, manually annotated 911 sentences randomly selected from the 2,960 collected sentences into one of the four IMRAD categories. In cases of sentences containing two or more categories, precedence was given to discussion over all other categories, to results over methods and introduction, and to methods over introduction. A confidence value was also assigned to each annotation: 'High' if the annotator was clear that the sentence belonged to a particular category, 'Medium' if the annotator was unsure between two categories, and 'Low' if the annotator was unsure between three or more categories. Of these 911 sentences, 749 sentences were annotated with 'High' confidence. These 749 sentences were used to train the classifier. Of the 749 sentences, 287 were labeled *introduction*, 192 were labeled *methods*, 180 were labeled *results* and 90 were labeled *discussion*.

To evaluate the quality of the annotation, we randomly selected 391 sentences from the 911 sentences. Two biologists (Annotator2 and Annotator3), who are not the authors of this paper, were provided the annotation guideline and independently assigned the IMRAD categories to each of the 391 sentences. Annotator2 annotated 196 sentences, while Annotator3 annotated 195 sentences. Agreement over these 391 sentences was 64.71%. 246 sentences were assigned high confidence by Annotator1 and Annotator2+3 (Table 1). Annotators agreed on 194 (78.86%) of these 246 sentences. Table 2 shows the results of kappa values and overall agreements of the 246 sentences that the annotators assigned high confidence and all 391 sentences regardless of confidence assigned by the annotators. The average kappa value and overall agreement[1] respectively were 0.71 and 89.5% when annotators assigned high confidence and 0.539 and 82.5% when confidence was ignored.

[1] the kappa values and the overall agreements using the calculator at http://www.dmi.columbia.edu/homepages/chuangj/kappa/calculator.htm

The 749 sentences that were annotated with ‘high’ confidence were used as a gold standard for evaluating different systems described in Section 2. For supervised machine-learning system trained on manually annotated full-text sentences, we performed 10-fold cross validation, in which 749 sentences were randomly divided into 10 folds, 9 folds (674-5 sentences) were then used for training. The trained classifier was then tested on the holdout 74-5 sentences. All other systems were evaluated ten times using the same set of the holdout sentences as the gold standard. We report the average recall, precision, and f-score with standard deviation.

4 Results

We report the results of rule-based and machine-learning classifications. Table 3 shows the performance of the classifiers. Table 3 also shows the results of adding two additional feature categories, tense of the verbs and original category of the sentence, to *Man*, as described in the methods.

Our mutual information score showed that the top-10 features were “were,” “citation,” “*NumberNumber*” (denotes any numeric value), “is,” “our,” “that,” “was,” “has,” “been” and “be.”

Table 1: Confidence value assigned by the annotators to the set of 391 sentences

		Annotator2 + Annotator3			
		High	Medium	Low	Total
Annotator1 (SA)	High	246	72	5	323
	Medium	38	18	8	64
	Low	4	0	0	4
	Total	288	90	13	391

Table 2: Annotator1 vs. Annotator2+3’s agreement on annotating sentences into the IMRAD categories.

	High Confidence Sentences		All Sentences	
	Kappa	OA(%)	Kappa	OA(%)
Introduction	0.688	88.2	0.514	80.1
Methods	0.862	94.3	0.704	89.0
Results	0.756	90.7	0.58	85.2
Discussion	0.532	84.6	0.358	75.7

OA: Overall Agreement

5 Discussion

Table 2 shows the kappa and the overall agreement by IMRAD category for sentences annotated with high confidence. The unweighted kappa score average was 0.71, which indicates good agreement between the annotators (19). The lowest and highest agreements were seen in Discussion and Methods, respectively, with kappa values of 0.532 and 0.862, respectively. The results indicate the challenge for a consistent sentence annotation in the Discussion category. Consistent with the confidence of annotation, our results show a decreased agreement when the confidence is not “High.” When confidence was ignored, average kappa score was 0.539, which is still in the range of an acceptable agreement (19).

Our results show that the baseline classifier achieved a competitive performance of 69.29% accuracy, which suggests that much of the sentences in full-text articles are indeed structured. It is not surprising that the supervised machine-learning system that is trained on the uncategorized sentences (*Non1*) achieved a similar performance (69.03%). We found an enhanced performance, although only slightly, in the iterative classifiers (*Non2*) that attempt to remove noisy data. The results show that the classifier indeed performed better than randomly, and was able to remove noise cases from the training data. On the other hand, multiple-classifiers may remove only those “easy” cases. Furthermore, because the classifiers remove sentences from the training data, the sentence removal led to decreases in training size, which may lead to a performance decrease in machine-learning classification. Results of iterative machine-learning classifications support our previous work in opinion/fact classification (17).

The rule-based classifier (*Rule-based*) was expected to perform with high precision; however, this was not the case. The precision for methods, results and discussion rules was between 52% and 68%. This could indicate the rules were not exclusive, and hence, as shown in our results, corpus based approaches present better options.

Although machine-learning classifiers trained on the structured abstracts (*Non3*) are widely considered as one of the best systems, our results show that these systems performed the worst (58.88%), a 10.4% decrease over the baseline system that considers a sentence based on which IMRAD section the sentence occurs in. The poor performance may be caused by the fact that sentences in full-text articles may be composed differently from sentences in abstracts. Our results strongly demonstrated that a full-text–specific classifier is needed.

Our results show that the classifier trained on the annotated sentences from randomly selected full-text articles (*Non4)* performed with 60.08% accuracy, which is much lower than a similar classifier *Non4* which was trained on sentences in the same article. The results show that the classifier performed better when trained on sentences in the same article than those across. This local effectiveness needs to be further investigated. Cohesion and semantics may play a role for IMRAD categorization.

Table 3: Performance (%) with standard-deviation across the 10-folds of all classifiers.

	Base-line	*Rule based*	*Non1*	*Non2*	*Non3*	*Non4*	*Man*			
							Words	Words + tense	Words + IMRAD	Words+Tense+IMRAD
A	69.29±3.54	55.40±8.80	69.03±3.86	69.43±3.41	58.88±5.95	60.08 ±4.36	75.83±5.08	76.10±4.48	81.04±4.82	81.30±4.67
I	69.9±5.76	63.4±10.8	69.7±5.77	69.7±5.77	61.4±9.65	66.6±4.04	80.6±6.31	82.2±6.69	83.5±4.99	84.3±5.13
M	81.2±6.73	59.7±11.3	80.8±5.72	81.4±5.49	70.8±6.21	66.2±7.45	76.3±7.02	76.2±7.79	83.9±8.96	84.1±8.12
R	72.2±7.26	32.0±8.43	71.3±8.46	71.9±8.02	54.5±11.8	54.5±12.4	69.7±8.78	68.3±7.63	77.6±10.2	77.2±11.2
D	46.3±12.3	37.5±18.2	46.6±13.3	46.7±13.2	39.4±13.7	42.6±12.9	59.7±21.8	59.4±20.0	58.4±24.9	61.5±14.8
WA	70.5	51.8	70.2	70.5	59.5	60.7	74.4	74.6	79.2	79.8

A: Accuracy, I: Introduction f-score, M: Methods f-score, R: Results f-score, D: Discussion f-score, WA: Weighted average of f-score.

The top features identified by mutual information showed the importance of citation markers, numbers and stop words. Accordingly, our results show that the word tense feature improved +0.27% (from 75.83% to 76.10%). Because of the strong performance of the baseline system, it is not surprising to see an improvement in performance (+5.21%) when the inherited IMRAD categories were added as the learning feature. We found that the best performance was to integrate both features. This resulted in an accuracy of 81.30%, which is 12% higher than the baseline system and 22.4% higher than the machine-learning system trained on structured abstracts.

Even though the annotated data are small—we had a total of 749 annotated sentences that were used for IMRAD categorization—we achieved a competitive performance system that is likely applicable to text-mining applications. We speculate that the systems can be further enhanced when more data are annotated and used for supervised machine learning.

6 Conclusion

In this study, we have explored several systems for automatically classifying a sentence that appears in a full-text article into the corresponding IMRAD category. An important finding in our work is that the IMRAD classifier that was trained on sentences in abstract does not perform well on sentences that appear in full-text. The best-performing system was a support vector machine classifier that was trained on manually annotated sentences that appear in full-text. The system achieved an accuracy of 81.30%, a performance that is 22.42% higher than the machine-learning system trained on sentences in abstract.

Acknowledgements: We acknowledge the support of 1R21RR024933-01A1, 5R01LM009836-02, and the University of Wisconsin-Milwaukee's RGI in 2007-2008, all to Hong Yu. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the NIH's views.

References

1. Day R. How to Write & Publish a Scientific Paper.: Cambridge University Press, Cambridge; 1998.
2. Gabbay I, Sutcliffe R. A qualitative comparison of scientific and journalistic texts from the perspective of extracting definitions. ACL Workshop on Question Answering in Retricted Domains; 2004; 2004.
3. Salanger-Meyer F. Discoursal movements in medical English abstracts and their linguistic exponents: A genre analysis study. INTERFACE: Journal of Applied Linguistics. 1990;4(2):107-24.
4. Swales J. Genre Analysis: English in Academic and Research Settings: Cambridge University Press, Cambridge, England; 1990.
5. Wang Y, Su X, Sorenson CM, Sheibani N. Tissue-specific distributions of alternatively spliced human PECAM-1 isoforms. Am J Physiol Heart Circ Physiol. 2003 Mar;284(3):H1008-17.
6. Yu H, Agarwal S, Johnston M, Cohen A. Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. J Biomed Discov Collab. 2009 Jan 6;4(1):1.
7. Yu H, Lee M, Kaufman D, Ely J, Osheroff JA, Hripcsak G, et al. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. J Biomed Inform. 2007 Jun;40(3):236-51.
8. McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts. AMIA Annu Symp Proc. 2003:440-4.
9. Lin J, Karakos D, Demner-Fushman D, Khudanpur S. Generative content models for structural analysis of medical abstracts. HLT-NAACL BioNLP; 2006; New York City; 2006.
10. Mizuta Y, Korhonen A, Mullen T, Collier N. Zone analysis in biology articles as a basis for information extraction. Int J Med Inform. 2006 Jun;75(6):468-87.
11. Mullen T, Mizuta Y, Collier N. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. ACM SIGKDD Explorations Newsletter. 2005;7(1):52-8.
12. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ. Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics. 2008 Sep 15;24(18):2086-93.
13. Wilbur WJ, Rzhetsky A, Shatkay H. New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics. 2006;7(1):356.
14. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994;1(2):161-74.
15. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc. 2002 May-Jun;9(3):262-72.
16. Hersh W, Cohen A, Roberts P, Rekapalli H. TREC 2006 Genomics Track overview. TREC Genomics Track conference; 2006; 2006.
17. Yu H, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of Empirical Methods in Natural Language Processing (EMNLP); 2003; Sapporo, Japan; 2003.
18. Yang Y, Pedersen J. A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference (ICML'97); 1997; 1997.
19. Fleiss J. Statistical methods for rates and proportions. New York: John Wiley & Sons (eds.); 1981.