BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 23 2009, pages 3174–3180
doi:10.1093/bioinformatics/btp548

Data and text mining

Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion

Shashank Agarwal and Hong Yu*

University of Wisconsin, Milwaukee, Milwaukee WI 53211, USA

Received on May 13, 2009; revised on September 11, 2009; accepted on September 14, 2009

Advance Access publication September 25, 2009

Associate Editor: Limsoon Wong

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 different approaches for automatically classifying sentences in MEDLINE abstracts into the IMRAD categories, few have explored the classification of sentences that appear in full-text biomedical articles. We first evaluated whether sentences in full-text biomedical articles could be reliably annotated into the IMRAD format and then explored different approaches for automatically classifying these sentences into the IMRAD categories. Our results show an overall annotation agreement of 82.14% with a Kappa score of 0.756. The best classification system is a multinomial naïve Bayes classifier trained on manually annotated data that achieved 91.95% accuracy and an average F-score of 91.55%, which is significantly higher than baseline systems. A web version of this system is available online at—http://wood.ims.uwm.edu/full_text_classifier/.

Contact: hongyu@uwm.edu

1 INTRODUCTION

Previous studies have concluded that biomedical texts typically fall into the rhetorical categories of Introduction, Methods, Results and Discussion (IMRAD) [e.g. (Day, 1998; Gabbay and Sutcliffe, 2004; Salanger-Meyer, 1990; Sollaci and Pereira, 2004; Swales, 1990)]. Sollaci and Pereira (2004) concluded that IMRAD has been the only structure adopted by scientific papers since the 1980s.

Although scientific articles are indeed structured under the IMRAD categories at the discourse level, sentences that appear under an IMRAD subheading do not necessarily conform to the expectations of the same IMRAD category. For example, the following is a paragraph from the Results section of a full-text article (Wang *et al*., 2003) in which sentences were manually classified into 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 annotating sentences into the IMRAD categories and computationally classifying sentences into these categories. The motivation for our work is that most text mining systems use sentences as independent units for information extraction [e.g. (Friedman *et al*., 2001)], summarization [e.g. (Yu *et al*., 2009)] and question answering [e.g. (Yu *et al*., 2007)], hence, sentence-level IMRAD classification may benefit such text-mining applications. For example, information extraction tools (e.g. extracting protein–protein interactions) may target evidence-rich results and avoid evidence-lean introductions (Yeh *et al*., 2003). Summarization may use such classification to 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 (Yu *et al*., 2009). Question answering may target different rhetorical categories for answer extraction. For example, definitions can often be extracted from Introductions (Yu *et al*., 2007).

The importance of classifying biomedical text into rhetorical-zone categories has been recognized, and various approaches have been developed to automate the task, although most of these efforts have been directed toward developing approaches for assigning IMRAD categories to sentences that appear in MEDLINE abstracts (McKnight and Srinivasan, 2003; Yu *et al*., 2007).

McKnight and Srinivasan (2003) reported the first automation of this task. They trained supervised machine-learning binary-classifiers on structured abstracts (i.e. the sentences in an abstract were structured by the authors into the IMRAD categories). The authors observed that sentences typically followed the IMRAD order in an abstract and therefore incorporated sentence position as an additional feature of their system. They reported F-scores of 52–79% for assigning each sentence to IMRAD categories. Similarly, Yamamoto and Takagi (2005) built a system to automatically classify abstract sentences into five rhetorical categories—Background,

*To whom correspondence should be addressed.

Purpose, Methods, Results and Conclusions. They reported an *F*-score ranging from 63 to 89.8% for their best classifier over the five categories. .Lin *et al*. (2006) implemented hidden Markov models to attain *F*-scores of 73–89%.

The research efforts described above attempt to classify sentences in abstracts, but to our knowledge, little work has been attempted to predict the IMRAD categories of sentences in full-text biomedical articles. Furthermore, no work has been reported in which manual annotation has been used to assign sentences to their IMRAD categories. Hence, whether a sentence that appears in a full-text biomedical article can be assigned IMRAD categories with a high degree of reliability is still an open question.

In their examination of full-text biomedical articles, Mizuta *et al*. (2006) explored linguistic features and developed richer rhetorical-zone categories, such as problem-setting (i.e. the problem to be solved), insight (i.e. the author's insights), etc. Using 20 annotated full-text articles, supervised machine-learning classifiers (i.e. naïve Bayes and support vector machines) were developed for automation (Mullen *et al*., 2005). Lexical and syntactic features were used along with the location of sentences and zone sequences. Their best performing system incorporated all features and achieved an *F*-score of 70% for all category classification.

Other related work includes Shatkay *et al*. (2008) and Wilbur *et al*. (2006), who first annotated 10 000 sentences selected from full-text biomedical articles along five parameters: focus, certainty, evidence, polarity and direction/trend. Sentences were broken into fragments by annotators, and each fragment was annotated by three annotators. They then built a multi-dimensional classifier using support vector machines, in which each sentence was classified along the same parameters. The classifier was trained on those annotated cases for which all three annotators agreed, and it achieved good performance according to its evaluation using 5-fold cross validation.

We previously developed a framework for IMRAD classification (Agarwal and Yu, 2009; Yu *et al*., 2009). This article extends that work to conduct a comprehensive experiment design and data analysis.

2 METHODS

We first examine whether a sentence can be reliably annotated into IMRAD categories, and we then explore text-mining approaches for automation.

2.1 Data

The publicly available BioMed Central full-text corpus was used for this study. We randomly selected 148 articles that explicitly incorporate the IMRAD sections into their structure and then randomly selected five sentences from each of these sections in the articles. This resulted in a total of 2960 sentences ($148 \times 5 \times 4$), which were used for annotating IMRAD categories.

2.2 Annotation, agreement and gold standard

The first author of this article (*AnnotatorAuthor*) developed an annotation guideline and used it to manually annotate 2000 sentences that were randomly selected from the set of 2960 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. We followed this order because the rhetorical structure in biomedical articles follows the IMRAD order, and each rhetorical category linearly follows the other. For example, for logical clarity, sentences in Results frequently introduce relevant background information or methods, while Discussion sentences frequently mention Results. This order of precedence was also recognized by McKnight and Srinivasan (2003) in their study. A confidence value was also assigned to each annotation: 'High' was assigned if the annotator was certain that the sentence belonged to a particular category, 'Medium' if the annotator could only be certain that it belonged to one of two categories, and 'Low' if the annotator could only be certain that it belonged to one of three or more categories.

Table 1. Confidence value assigned by the annotators to the set of 1930 sentences

	AnnotatorAuthor	High	Medium	Low	Total
AnnotatorBiologists	High	1377	193	30	1600
	Medium	227	61	25	313
	Low	11	4	2	17
	Total	1615	258	57	1930

Table 2. Matrix of category assignment by *AnnotatorAuthor* and *AnnotatorBiologists* for sentences annotated with 'High' confidence

	AnnotatorBiologists	I	M	R	D
AnnotatorAuthor	I	389	18	16	78
	M	12	363	21	7
	R	1	14	273	44
	D	13	1	21	106

I: Introduction, M: Methods, R: Results, D: Discussion.

To evaluate the quality of the annotation, we provided five biomedical researchers (*AnnotatorBiologists)* with the annotation guideline and asked them to independently assign IMRAD categories and confidence scores to 400 sentences each. Thus between them, the *AnnotatorBiologists*, who are not the authors of this paper, annotated the same 2000 sentences that were annotated by the *AnnotatorAuthor*, so that every sentence was annotated by two annotators. Since a heuristic-based sentence splitter was used to produce the data, there were cases of training sentences not being split correctly. Annotators were allowed to tag these sentences as artifacts, and these sentences were removed. There were 70 such cases, leaving 1930 sentences for full annotation.

The agreement between the *AnnotatorAuthor* and *AnnotatorBiologists* over the 1930 sentences was 72.54%, with a Kappa score of 0.629. One-thousand three-hundred and seventy-seven sentences were assigned high confidence by both *AnnotatorAuthor* and *AnnotatorBiologists* (Table 1). Annotators agreed on 1131 (82.14%) of these 1377 sentences, with a Kappa score of 0.756. We found that disagreement was mainly caused due to sentence ambiguity. For example, the sentence 'However, physicians in training will likely deal with such marketing influences once in practice' was classified as Introduction by the *AnnotatorAuthor* and Discussion by the *AnnotatorBiologists*. The sentence could have been interpreted as an explanation for a result or as a research question posed by the current study. Such ambiguity is quite common in natural language (Mihalcea, 2003).

The 1131 sentences that both the *AnnotatorAuthor* and *AnnotatorBiologists* annotated with a confidence value of 'High' and were in agreement on with respect to IMRAD categories were used to generate the gold standard to evaluate the different systems discussed in the next section. Of these 1131 sentences, 389 were labeled Introduction, 363 were labeled Methods, 273 were labeled Results and 106 were labeled Discussion (Table 2).

3 AUTOMATIC CLASSIFICATION

We explored rule-based and machine-learning approaches for automatically classifying a sentence into IMRAD categories.

3.1 A baseline system

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

3.2 A rule-based system

Rule-based systems have attained success in the biomedical domain [e.g. (Friedman *et al*., 1994; Yu *et al*., 2002)]. We randomly selected eight articles from the TREC Genomics Track text collection (Hersh *et al*., 2006), which contains more than 160 000 full-text biomedical articles. The eight articles contain ∼30 000 words and 1250 sentences. The first author of this article (SA) read each article and then manually identified patterns that were indicative of IMRAD categories. These patterns consisted of individual words or phrases, and sentences were probed for the presence of these patterns using regular expressions. For example, one rule links a sentence to Discussion if the sentence incorporates the words 'our', 'observations' and 'suggests' and the sentence is not associated 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 (*Rule*) that automatically assigns sentences to the appropriate category. Manually identifying rules was a time consuming task; it took the first author two hours per article on average to identify and code all rules in that article.

3.3 Supervised machine-learning systems trained on non-annotated corpus

As stated above, manually creating rules is an expensive process, and developing machine learning approaches with minimum manual effort is an attractive proposition. We explored methods for training supervised machine-learning systems on data that does not require further annotation, and in this respect, our work is inspired by the work of (Yu and Hatzivassiloglou, 2003). We assume that in a full-text, IMRAD-structured article, the majority of sentences in each section will be classified into their 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 incorporating 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. 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-than-random chance of removing noisy training data. We then continued the iteration C_i, $i = 1, 2, \ldots, N$. We found that two iterations gave the best accuracy when tested on the gold standard; hence, we report the performance of *Non2* based on two iterations. *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). Eight thousand randomly selected sentences (2000 for each category) from the structured abstracts in MEDLINE were used to train the classifier.

Non4 was trained on structured full-text sentences instead of abstract sentences. Eight thousand sentences (2000 from each category) from the IMRAD categories were randomly collected from full-text articles in the BioMed Central corpus 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.

3.4 Supervised machine-learning system trained on manually annotated full-text sentences

The non-annotated data is noisy; hence, classifiers trained on this data may not obtain optimal performance. To overcome this disadvantage, we trained 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.

3.5 Machine-learning systems and features

For all supervised classifications, we tested three algorithms: multinomial naïve Bayes, naïve Bayes and support vector machine (SVM). Both naïve Bayes and SVMs are widely used supervised machine-learning algorithms. The probabilistic framework of naïve Bayes follows a multi-variate Bernoulli model in which a sentence is represented by a vector of words; 0 indicates absence, while 1 indicates presence of the word at least once. Multinomial naïve Bayes represents a multinomial distribution of words in a sentence that captures word frequency. The multinomial model has been shown to outperform naïve Bayes in document classification (McCallum and Nigam, 1998). We used the implementation of all three algorithms provided by the open-source Java™-based machine-learning library Weka 3 (Witten and Frank, 2006).

We explored words and *n*-grams features testing them as individual words, combinations of individual words and bigrams, and combinations of individual words, bigrams and trigrams. We observed that citations can be an important feature for distinguishing categories; for example, citations are more frequently used in Introduction than in Results. Hence, we created a new feature to indicate the presence of a citation. All numbers were replaced by a unique symbol—*#NuMBeR*. We did not remove stop words since certain stop words are more likely to be associated with certain IMRAD categories. We also did not remove words that referred to a figure or table, since such references are more likely to occur in sentences indicating the outcome of the study. To observe the contribution of citations, stop words, reference to figures and tables, and replacement of numbers for classification, we ran our best classifier after removing these features.

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 and used the Stanford parser (Klein and Manning, 2003) to identify the presence of these tenses. We also explored the IMRAD categories inherited from a structured full-text article as a feature. This feature was only added in the machine-learning classifier *Man*. On the basis of feature selection, we trained four different *Man* classifiers—*Man-Terms,* which was trained only on term features; *Man-Tense*, which was trained using term features and verb tenses; *Man-IMRAD*, which was trained using term features and the original IMRAD category of the sentence; and *Man-All*, which was trained using term features, verb tenses and the original IMRAD category of the sentence.

We experimented with mutual information and chi-square for feature selection. In experimenting with different top features, we obtained better performance using the top-2500 features.

3.6 Gold standard and evaluation

We created six sets of gold standards. The first gold standard set comprised 1131 sentences that were agreed upon by two annotators and were annotated with 'High' confidence. However, this resulted in the rejection of 799 sentences and a highly unambiguous training set that might not be representative of all sentences in the literature. Thus, we created a second gold standard comprising all the sentences agreed upon by the two annotators. Two additional gold standard sets were created using *AnnotatorAuthor*'s or *AnnotatorBiologist*'s annotations. The remaining two gold standard sets were created using *AnnotatorAuthor*'s or *AnnotatorBiologist*'s 'High' confidence annotations.

We evaluated the supervised machine-learning systems on the first gold standard, which consisted of 1131 sentences. The sentences were randomly divided into 10-folds. Nine folds (1017-8 sentences) were then used for training, and the trained classifier was then tested on the fold that had been held out of the training set (113-4 sentences). All other systems were evaluated 10 times using the same set of holdout sentences as the gold standard. For all systems, we report overall accuracy, recall, precision and F-score for each category and the micro-average of recall, precision and F-score for all systems. The micro-average is the mean when each class is weighted according to its size. 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 sentences predicted in the same category.

4 RESULTS

We compared the performance of multinomial naïve Bayes, naïve Bayes and SVM algorithms with feature selection based on mutual information and chi-square using the *Man-All* classifier. We found that the best accuracy was achieved using the multinomial naïve Bayes algorithm with mutual information based feature selection (Table 3). Hence, for subsequent tests, we selected term features by sorting them by their mutual information scores and then trained multinomial naïve Bayes classifiers.

We evaluated the effect of individual words, bigrams and trigrams as text features to train the *Man-All* classifier. Our results indicated that a combination of individual words, bigrams and trigrams gave the best performance (Table 4). Hence, for all machine-learning classification runs, we used this combination.

Table 3. Comparison of the accuracy of multinomial naïve Bayes, naïve Bayes and SVM algorithms when trained on text features selected using mutual information or chi-square

	Multinomial naïve Bayes	Naïve Bayes	SVM
Mutual information	91.95±2.81	85.95±3.44	89.13±2.3
Chi-square	86.03±2.69	86.65±3.07	87.09±2.48

Table 4. Performance of *Man-All* classifier using different combinations of individual words, bigrams and trigrams as term features

	Iw	Iw + *b*	Iw + *b*+*t*
Accuracy	90.10±1.90	91.60±2.14	91.95±2.81
Introduction	F: 91.24±3.38	F: 92.56±3.45	F: 92.65±3.49
Methods	F: 94.42±2.02	F: 95.14±2.14	F: 95.04±3.12
Results	F: 90.63±3.46	F: 91.2±3.01	F: 92.24±3.89
Discussion	F: 60.20±19.1	F: 71.46±13.4	F: 73.77±14.6
Micro-average	R: 90.43	R: 92.01	R: 92.36
	P: 90.00	P: 91.72	P: 91.93
	F: 89.20	F: 91.08	F: 91.55

P: Precision, R: Recall, F: F-score, Iw: Individual words, b: bigrams, t: trigrams.

We report the results of rule-based and machine-learning classifications. Table 5 shows the performance of all classifiers. As mentioned in Section 3.5, the *Man* classifier was trained using a combination of term features with verb tenses in the sentence and the original IMRAD category of the sentence. Table 5 also shows the results of comparing the performance of different features used to train the *Man* classifier. The accuracies obtained by performing 10-fold cross validation using *Man-All* classifier for each of the six gold standards are shown in Table 6.

Our mutual information score showed that the top-10 features were 'citation', 'were', '*#NuMBeR'* (denotes any numeric value), 'is', 'that', 'was', 'has', 'table *#NuMBeR'*, '*#NuMBeR #NuMBeR'* and 'table'. Since terms that are normally removed during normalization of text attained high mutual-information scores, we wanted to study the effect of removing these terms. We compared the performance of the *Man-All* classifier by removing stop words, numbers and references to figures and tables (Table 7). We removed stop words that were obtained from the list of stop words used in the information retrieval library Lucene (Gospodnetic and Hatcher, 2005). The results show that the removal of citation markers, stop words, references to figures and tables and numbers led to a decrease in the performance of the *Man-All* classifier by 1.97, 4.77, 0.71 and 0.17%, respectively.

To evaluate if the number of annotated sentences is sufficient, we performed a 10-fold cross validation of different-sized gold standards. We randomly selected sentences from the original gold standard to create gold standard subsets of sizes ranging from 100 to 1100 sentences in increments of 100. We performed 10-fold cross validation using the *Man-All* classifier, as shown in the case

Table 5. Performance with SD across the 10-folds of all classifiers

	Baseline	*Rule*	*Non1*	*Non2*	*Non3*	*Non4*	*Man-Terms*	*Man-IMRAD*	*Man-Tense*	*Man-All*
A	77.81±4.03	58.18±4.87	74.45±5.22	72.77±5.24	66.94±4.49	73.92±2.39	88.06±3.2	91.34±3.09	88.77±3.19	91.95±2.81
I	R:67.88±8.99	R:86.98±7.01	R:92.38±6.82	R:92.88±8.16	R:51.76±8.62	R:74.04±7.53	R:95.5±2.53	R:97.57±1.95	R:95.24±2.22	R:97.36±2.02
	P:91.79±7.45	P:49.02±6.67	P:62.16±9.22	P:59.08±10.3	P:83.87±6.47	P:79.01±11.5	P:85.87±7.58	P:88.06±7.34	P:86.41±7.22	P:88.59±5.91
	F:77.87±8.11	F:62.42±6.61	F:74.07±8.13	F:71.9±9.65	F:63.53±6.56	F:75.96±8.47	F:90.21±4.23	F:92.38±4.37	F:90.38±3.89	F:92.65±3.49
M	R:85.43±3.7	R:62.13±6.34	R:91.93±6.82	R:91.04±6.03	R:72.48±7.58	R:82.73±4.23	R:93.69±2.23	R:96.11±2.90	R:95.44±2.21	R:96.16±2.39
	P:91.78±2.62	P:80.34±8.80	P:82.14±5.27	P:80.02±6.05	P:84.42±7.65	P:83.33±8.28	P:87.82±5.91	P:94.25±5.04	P:88.00±6.33	P:94.09±5.0
	F:88.44±2.51	F:69.74±5.75	F:86.71±4.71	F:85.07±5.36	F:77.58±5.12	F:82.71±4.23	F:90.49±2.69	F:95.09±3.16	F:91.40±3.30	F:95.04±3.12
R	R:78.56±6.92	R:22.89±9.94	R:74.45±8.36	R:72.36±7.63	R:82.38±4.16	R:67.65±9.02	R:82.84±5.98	R:88.30±7.65	R:84.05±5.30	R:90.81±6.93
	P:78.25±6.25	P:54.51±19.2	P:77.87±7.46	P:77.49±7.73	P:60.46±8.15	P:74.53±8.28	P:91.63±5.83	P:93.43±5.89	P:93.38±4.68	P:94.16±4.15
	F:78.21±5.23	F:31.81±12.9	F:75.95±7.18	F:74.69±6.97	F:69.35±5.12	F:70.56±7.06	F:86.84±4.57	F:90.46±4.39	F:88.31±3.60	F:92.24±3.89
D	R:84.11±8.56	R:29.33±16.1	R:33.96±9.43	R:32.99±9.33	R:69.81±15.8	R:60.18±12.8	R:58.64±20.8	R:64.70±18.5	R:59.24±20.1	R:65.0±18.3
	P:38.77±10.0	P:59.88±27.6	P:82.11±7.66	P:82.74±6.75	P:34.05±8.59	P:38.90±8.61	P:91.59±11.8	P:87.66±14.0	P:91.55±10.4	P:91.05±10.7
	F:52.44±10.5	F:38.25±19.3	F:47.35±10.2	F:46.45±10.2	F:45.07±9.71	F:46.51±8.66	F:68.3±16.1	F:72.36±15.4	F:69.03±15.0	F:73.77±14.6
ma	R: 77.61	R: 58.13	R: 82.43	R: 81.72	R: 67.49	R: 73.99	R: 88.41	R: 91.78	R: 89.23	R: 92.36
	P: 83.55	P: 61.42	P: 74.23	P: 72.46	P: 73.73	P: 75.56	P: 88.42	P: 91.30	P: 89.09	P: 91.93
	F: 78.96	F: 55.12	F: 76.08	F: 74.42	F: 67.71	F: 74.06	F: 87.43	F: 90.91	F: 88.21	F: 91.55

A: Accuracy, I: Introduction, M: Methods, R: Results, D: Discussion, ma: Micro-average, F: *F*-score, R: Recall, P: Precision.

Table 6. Performance of the *Man-All* classifier on different training sets obtained when annotations of *AnnotatorAuthor* or *AnnotatorBiologists* are used as the gold standard and/or confidence value is ignored

Gold standard	Only 'High' confidence sentences	All sentences
Sentences agreed by *AnnotatorAuthor* and *AnnotatorBiologists*	91.95 (77.81)	88.29 (74.29)
Annotations of *AnnotatorAuthor*	88.31 (69.12)	82.95 (65.08)
Annotations of *AnnotatorBiologists*	81.04 (68.5)	78.24 (64.82)

The value in the parentheses is the performance of *Baseline* for that training set.

Table 7. Performance of the *Man-All* classifier when stop words, citations, numbers and reference to figures and tables were removed from term features

	No features removed	Remove sw	Remove citations	Remove numbers	Remove f&t
A	91.95±2.81	87.18±1.74	90.98±2.01	91.78±2.41	91.42±2.53
da		4.77	1.97	0.17	0.53
I	F:92.65±3.49	F:88.3±4.61	F:91.77±3.14	F:93.38±2.57	F:92.78±2.95
M	F:95.04±3.12	F:93.04±3.02	F:95.02±2.90	F:95.17±1.99	F:94.73±3.21
R	F:92.24±3.89	F:89.07±4.62	F:90.14±4.43	F:90.71±3.59	F:90.12±3.51
D	F:73.77±14.6	F:33.28±16.7	F:71.65±13.8	F:72.97±14.7	F:74.17±13.5
ma	R: 92.36	R: 85.66	R: 91.39	R: 92.13	R: 91.79
	P: 91.93	P: 86.15	P: 91.20	P: 91.89	P: 91.43
	F: 91.55	F: 84.10	F: 90.54	F: 91.40	F: 91.02

A: Accuracy, da: Decrease in accuracy, I: Introduction, M: Methods, R: Results, D: Discussion, ma: Micro-average, F: *F*-score, R: Recall, P: Precision, sw: stop words f&t: reference to figures and tables.

of the original gold standard. The accuracies of classifiers using different-sized gold standards are shown in Figure 1.

To further test the robustness of *Man* system, we split the sentences into 5- and 10-folds after sorting them by publication date. We performed a 5- and 10-fold cross validation by training the *Man-All* classifier on this data and obtained an accuracy of 91.78±1.87 and 92.13±2.67%, respectively. *Man-All*'s accuracy on randomly distributed data was 91.95±2.81%.

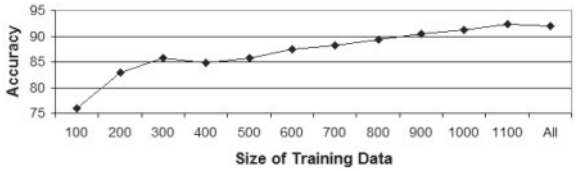

Fig. 1. Accuracy of the *Man-All* classifier for different sized gold standard subsets.

5 DISCUSSION

Our inter-annotator agreement results show that when sentences tagged with high confidence were compared, an agreement of 82.14% was observed, with a Kappa score of 0.756. This indicates good agreement between the annotators (Fleiss, 1981). By ignoring the confidence value, overall agreement and Kappa score dropped to 72.54% and 0.629, respectively, which indicates acceptable agreement (Fleiss, 1981). This indicates that the use of confidence values while annotating improves agreement and quality of annotation. Table 1 indicates that 71.35% sentences were assigned with high confidence by both annotators. Although selecting only those sentences that were agreed upon and assigned with 'high' confidence by both annotators decreases the size of the gold standard substantially, we believe that the improvement in the quality of annotation due to the removal of ambiguous sentences offsets the disadvantage of fewer training data. For example, ambiguous sentences, such as '*For all of these comparisons the strength of the correlations will be weakened by the different time frames used; CPG six months, troublesomeness four weeks, GHQ 12 and EQ 5D today'*, were annotated with 'low' confidence by both annotators. However, a 10-fold cross-validation evaluation using sentences annotated with any confidence value by one annotator (*AnnotatorAuthor*) showed 82.95% accuracy (Table 6). Our results indicate that if all ambiguous cases were also used for learning the classifier, it would achieve an

accuracy of 82.95%. Hence, at worst, it is possible to develop a system performing with 82.95% accuracy. In Figure 1, it can be seen that the difference in the performance of the *Man-All* classifier at the gold standards of 1000, 1100 and all 1131 sentences is small (they are within 1.0% of each other). This suggests that the current gold standard, which includes 1131 'high' confidence sentences, is sufficient for obtaining a classifier that performs well.

The top features identified by mutual information showed the importance of citation markers, numbers, stop words and reference to figures and tables. Consistent with the mutual information results, we observed the *F*-score for Discussion decreased by ∼40% on removing stop words (Table 7). Our results indicate that stop words are important for identifying Discussion sentences. For example, the sentence '*A possible limitation of our study may be the relatively low percentage of duodenal biopsies performed in patients with primary biliary cirrhosis tested positive for at least one antibody class'* was incorrectly predicted as Introduction when stop words were removed, whereas it was correctly predicted as Discussion when stop words were not removed. This is because the stop words 'our' and 'may' are used by the classifier to identify a sentence as a Discussion sentence.

Our results show that the baseline classifier (*Baseline*) achieved a competitive performance of 77.81% accuracy, which suggests that many of the sentences in full-text articles are indeed structured. However, it also suggests that ∼22% of sentences do not belong to the category they appear in. Hence, it is not surprising that the supervised machine-learning system trained on uncategorized sentences (*Non1*) achieved an accuracy of 74.45%. This performance is poorer than that of the baseline classifier and might be due to noise in the training data and/or a smaller training dataset. Similarly, the iterative classifiers (*Non2*) that attempt to remove noisy data also performed worse (72.77%) than *Baseline*. The results suggest that the iterative classifier might have removed sentences that were misclassified, leading to a decrease in training size and hence a decrease in performance. Results of iterative machine-learning classifications support previous work in opinion/fact classification (Yu and Hatzivassiloglou, 2003).

The rule-based classifier (*Rule*) was expected to perform with high precision; however, this was not the case. The precision for Methods, Results and Discussion rules was between 54 and 81%, which indicates that the rules were not exclusive.

Although a machine-learning classifier trained on structured abstracts (*Non3*) is widely considered to be among the best systems, our results show that these systems did not perform well (66.94%). We noticed that this represented a 10.87% decrease in accuracy over a baseline system that considers a sentence based on the IMRAD section in which it occurs. The poor performance may be caused by the fact that sentences in full-text articles are composed differently than sentences in abstracts. For example, abstracts rarely contain such features as citations and references to figures and tables that were shown to have an effect on the performance of machine-learning classifiers (Table 7). As shown in our results, when we removed citation and reference to figures and tables as features from the *Man* classifier, accuracy decreased by 1.97% points and 0.53%, respectively. However, this still does not justify the difference in performance between classifiers trained on abstract sentences and those trained on manually annotated sentences. On further inspection, we found that abstract sentences are often noisy, similar to full-text sentences. Of the 100 sentences randomly selected from structured abstracts that were analyzed by the first author of this article, it was observed that 27% did not belong to the category they appear in (results not shown). Our results strongly demonstrate that a classifier specific to full texts is needed and that high quality annotated data is a must.

Our results also show that the classifier trained on annotated sentences from structured, full-text articles that were randomly selected (*Non4*) performed with an accuracy of 73.92%. This is lower than the performance of *Baseline*. However, this was not completely unexpected, since the performance of *Baseline* indicates that ∼22% of the sentences do not belong to the category they appear in, which results in *Non4* being trained on extremely noisy data. Also, it was noted that the performance of *Non4* is similar to classifier *Non1,* which was trained on sentences in the same article. This similarity in performance suggests that the presence of noise in the training data was responsible for low performance. It also justifies the need for a manually annotated corpus for classifying sentences into IMRAD categories.

Our results showed that multinomial naïve Bayes performed better than SVM at classifying sentences. This divergence needs to be investigated further. When studying the effect of non-text features on the *Man* classifier, adding the tense of verbs feature (*Man-Tense*) to the classifier based on term features only (*Man-Terms*) improved accuracy by 0.71% (from 88.06 to 88.77%). Because of the strong performance of the baseline system, it is not surprising to see an improvement in performance (+3.28%) when the inherited IMRAD categories were added as the learning feature (*Man-IMRAD*). We found that the best performance was produced by integrating both features (*Man-All*). This resulted in an accuracy of 91.95%, which is 14.14% points higher than the baseline system. Also, our classifier is robust as the performance of *Man-All* on time-distributed and randomly-distributed data was not statistically significant. The high accuracy of our system could help in developing many other classification applications, for instance, citation classification, which we intend to explore in the future.

Table 8 shows a sample of two sentences that were incorrectly classified by *Man-All* for analysis. The first sentence was misclassified as Methods due to the presence of the feature 'analyzed', while the feature 'observed' and the presence of a number were used to classify the second sentence as Results.

Our work relates to the work of (Shatkay *et al*., 2008), particularly with respect to their introduction of a 'Focus' dimension, which contains the categories 'Scientific', 'Generic' and 'Methodology'.

Table 8. Sample of sentences from the gold standard, the category assigned by annotators and the *Man-All* classifier

Incorrectly classified sentences	Gold standard	Assigned
When we analyzed the ROC curve just using the 796 cases with localized cancers, we found a similar area of 0.64 (SE 0.01)	Results	Methods
It is indeed the longest intron observed in FrRUNT , but it is nevertheless very short, spanning just 1372 bp	Introduction	Results

They reported an accuracy of 92% for this three-category classification problem, which compares well with the accuracy of our classifier. However, instead of classifying entire sentences, they broke the sentences into fragments and annotated the fragments, which might help reduce ambiguous cases. However, since fragment boundaries cannot be easily identified automatically, we did not explore a fragment-based classification strategy.

Similar to our study, (Mullen *et al*., 2005) also classify sentences in full-text articles; however, our study differs from their work in three important ways: First, our IMRAD categories represent a coarse-level representation of four categories, while Mullen *et al*. presented a fine-level categorization comprising 10 categories. We propose that their categories 'Background' and 'Problem-setting' map to Introduction, 'Method' maps to Methods, 'Result' maps to Results, while 'Insight' and 'Implication' map to Discussion. Four of their categories, 'Else', 'Connection', 'Difference' and 'Outline' do not map to any of the IMRAD categories. We believe that the IMRAD representation naturally represents the overall rhetorical structure of full-text biomedical articles and that it can be applied to specific text mining tasks [e.g. figure summarization (Yu *et al*., 2009)]. More importantly, we believe that our coarse-level categorization leads to stronger annotation agreement. Note that the data reported in Mullen *et al*. (2005) was annotated by only one annotator, and hence, there was no report of inter-annotator agreement. Finally, because of the coarse nature of our representation, our system yields higher classification accuracy (91.55% *F*-score) than Mullen *et al*.'s system (70% *F*-score).

6 CONCLUSION

In this study, we have explored several systems for automatically classifying a sentence that appears in a full-text article into its corresponding IMRAD category. An important finding in our work is that the IMRAD classifier that was trained on sentences in abstracts does not perform well on sentences appearing in full text. The best-performing system was a multinomial naïve Bayes classifier trained on manually annotated sentences that appear in full text. The system achieved an accuracy of 91.95%, a performance that is ∼14% points higher than the baseline system. A web version of our classifier is available online at: http://wood.ims.uwm.edu/full_text_classifier/.

ACKNOWLEDGEMENTS

The authors thank Dr Lamont Antieau for proofreading this manuscript.

Funding: National Institutes of Health grants 1r21rr024933-01a1, 5r01lm009836-02 and the University of Wisconsin-Milwaukee's RGI in 2007–2008 (to H.Y.).

Conflict of Interest: none declared.

REFERENCES

Agarwal,S. and Yu,H. (2009) Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussions. *2009 AMIA Summit on Translational Bioinformatics*. The American Medical Informatics Association, San Francisco, CA, USA, pp. 6–10.

Day,R. (1998) *How to Write & Publish a Scientific Paper*. Cambridge University Press, Cambridge.

Fleiss,J. (1981) *Statistical Methods for Rates and Proportions*. John Wiley and Sons, Inc., New York.

Friedman,C. *et al*. (1994) A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.*, **1**, 161–174.

Friedman,C. *et al*. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17** (Suppl. 1), S74–S82.

Gabbay,I. and Sutcliffe,R. (2004) A qualitative comparison of scientific and journalistic texts from the perspective of extracting definitions. *ACL Workshop on Question Answering in Restricted Domains*. Association for Computational Linguistics, Barcelona, pp. 16–22.

Gospodnetic,O. and Hatcher,E. (2005) *Lucene in Action*. Manning Publications, Greenwich, CT.

Hersh,W. *et al*. (2006) TREC 2006 genomics track overview. *TREC Genomics Track Conference*. National Institute of Standards and Technology (NIST), Gaithersburg, MD, pp. 52–78.

Klein,D. and Manning,C.D. (2003) *Accurate Unlexicalized Parsing*. Association for Computational Linguistics, Morristown, NJ, USA, pp. 423–430.

Lin,J. *et al*. (2006) Generative content models for structural analysis of medical abstracts. *HLT-NAACL BioNLP*, The Association for Computational Linguistics, New York City.

McCallum,A. and Nigam,K. (1998) A comparison of event models for naive Bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization*. The AAAI Press, Madison, Wisconsin, pp. 41–48.

McKnight,L. and Srinivasan,P. (2003) Categorization of sentence types in medical abstracts. *AMIA Annu. Symp. Proc.*, The American Medical Informatics Association, Washington DC, pp. 440–444.

Mihalcea,R. (2003) The role of non-ambiguous words in natural language disambiguation. *Proceedings of the Fourth RANLP*. Current Issues in Linguistic Theory (CILT), Borovets, Bulgaria, pp. 387–396.

Mizuta,Y. *et al*. (2006) Zone analysis in biology articles as a basis for information extraction. *Int. J. Med. Inform.*, **75**, 468–487.

Mullen,T. *et al*. (2005) A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *ACM SIGKDD Explor. Newslett.*, **7**, 52–58.

Salanger-Meyer,F. (1990) Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study. *INTERFACE: J. Appl. Linguist.*, **4**, 107–124.

Shatkay,H. *et al*. (2008) Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, **24**, 2086.

Sollaci,L.B. and Pereira,M.G. (2004) The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J. Med. Library Assoc.*, **92**, 364.

Swales,J. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, Cambridge, UK.

Wang,Y. *et al*. (2003) Tissue-specific distributions of alternatively spliced human PECAM-1 isoforms. *Am. J. Physiol. Heart Circ. Physiol.*, **284**, H1008–H1017.

Wilbur,W.J. *et al*. (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.

Witten,I.H. and Frank,E. (2006) *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier/Morgan Kaufmann, Amsterdam.

Yamamoto,Y. and Takagi,T. (2005) A sentence classification system for multi biomedical literature summarization. *Proceedings of the 21st International Conference on Data Engineering Workshops*. IEEE Computer Society.

Yeh,A.S. *et al*. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19** (Suppl. 1), i331–i339.

Yu,H. *et al*. (2002) Mapping abbreviations to full forms in biomedical articles. *J. Am. Med. Inform. Assoc.*, **9,** 262–272.

Yu,H. and Hatzivassiloglou,V. (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*. Sapporo, Japan.

Yu,H., *et al*. (2007) Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J. Biomed. Inform.*, **40**, 236–251.

Yu,H. *et al*. (2009) Are figure legends sufficient? Evaluating the contribution of associated text to biomedical figure comprehension. *J. Biomed. Disc. Collabor.*, **4**, 1.