An automated method to build a corpus of rhetorically-classified sentences in biomedical texts

Hospice Houngbo
Department of Computer Science
The University of Western Ontario
hhoungbo@uwo.ca

Robert E. Mercer
Department of Computer Science
The University of Western Ontario
mercer@csd.uwo.ca

Abstract

The rhetorical classification of sentences in biomedical texts is an important task in the recognition of the components of a scientific argument. Generating supervised machine learned models to do this recognition requires corpora annotated for the rhetorical categories **Introduction** (or Background), **Method**, **Result**, **Discussion** (or Conclusion). Currently, a few, small annotated corpora exist. We use a straightforward feature of co-referring text using the word "this" to build a *self-annotating* corpus extracted from a large biomedical research paper dataset. The corpus is annotated for all of the rhetorical categories except **Introduction** without involving domain experts. In a 10-fold cross-validation, we report an overall F-score of 97% with Naïve Bayes and 98.7% with SVM, far above those previously reported.

1 Introduction

Sentence classification is an important preprocessing task in the recognition of the components of an argument in scientific text. For instance, sentences that are deemed as conclusions of a research paper can be used to validate or refute an hypothesis presented in background or introduction sentences in that paper. Therefore, in order to understand the argumentation flow in scientific publications, we need to understand how different sentences fit into the complete rhetorical structure of scientific writing.

To perform sentence classification using supervised machine learning techniques requires a large training corpus annotated with the appropriate classification tags. In the biomedical domain, some corpora already exist, but many of these corpora are still limited and cannot be generalized to every context. The task of sentence classification in various rhetorical categories is often performed on *ad hoc* corpora derived from a limited number of papers that don't necessarily represent all of the text in the biomedical domain. For instance, the corpus used by Agarwal and Yu (2009) for the task of sentence classification into the IMRaD categories, is composed of only 1131 sentences.

In this study, we hypothesize that using a simple linguistically-based heuristic, we can build a significantly larger corpus comprising sentences that belong to specific categories of the IMRaD rhetorical structure of the biomedical research text, that will not need domain experts to annotate them, and will represent a wider range of publications in the biomedical literature. We have collected pairs of sequential sentences where the second sentence begins with "This method...", "This result...", "This conclusion...". Our hypothesis is that the first sentence in each pair is a sentence that can be categorized respectively as **Method**, **Result** and **Conclusion** sentences.

We have a number of motivations for this work. First, sentences are the basis for most text mining and extraction systems. The second motivation is that biomedical texts are the reports of scientific investigations and their discourse structures should represent the scientific method that drives these investigations. The third and last motivation is that categorizing sentences into the IMRaD categories can help in the task of extracting knowledge discovery elements from scientific papers.

The contribution of our work is twofold. First, we have used a simple linguistic filter to automatically select thousands of sentences that have a high probability of being correctly categorized in the IMRAD scheme, and second, we have used machine learning techniques to classify sentences in order to validate our hypothesis that this linguistic filter works. The rest of this paper is organized as follows. The next section reviews some related

Proceedings of the First Workshop on Argumentation Mining, pages 19–23,
Baltimore, Maryland USA, June 26, 2014. ©2014 Association for Computational Linguistics

work. In Section 3, a detailed methodology of corpus construction and sentence classification techniques is presented. In Section 4, the results are described.

2 Related Work

The classification of sentences from scientific research papers into different categories has been investigated in previous works. Many schemes have been used and currently no standard classification scheme has been agreed upon. Teufel et al. (1999) use a classification scheme termed Argumentative Zoning (AZ) to model the rhetorical and argumentative aspects of scientific writing in order to easily detect the different claims that are mentioned in a scientific research paper. AZ has been modified for the annotation of biology articles (Yoko et al., 2006) and chemistry articles (Teufel et al., 2009).

Scientific discourse has also been studied in terms of speculation and modality by Kilicoglu and Bergler (2008) and Medlock and Briscoe (2007). Also, Shatkay et al. (2008) and Wilbur et al. (2006) have proposed an annotation scheme that categorizes sentences according to various dimensions such as focus, polarity and certainty. Many annotation units have also be proposed in previous studies. Sentence level annotation is used in Teufel et al. (1999) whereas de Waard et al. (2009) used a multi-dimensional scheme for the annotation of biomedical events (bio-events) in texts.

Liakata et al. (2012) attempt to classify sentences into the Core Scientific Concept (CoreSC) scheme. This classification scheme consists of a number of categories distributed into hierarchical layers. The first layer consists of 11 categories, which describe the main components of a scientific investigation, the second layer consists of properties of those categories (e.g. Novelty, Advantage), and the third layer provides identifiers that link together instances of the same concept.

Some other recent works have focussed on the classification of sentences from biomedical articles into the IMRaD (Introduction, Methods, Research, and, Discussion) categories. Agarwal and Yu (2009) use a corpus of 1131 sentences to classify sentences from biomedical research papers into these categories. In this study, sentence level annotation is used and multinomial Naïve Bayes machine learning has proved to perform better than simple Naïve Bayes. The authors report an overall F-measure score of 91.55% with a mutual information feature selection technique. The present study provides an alternative way to build a larger IMRaD annotated corpus, which combined with existing corpora achieves a better performance.

Methods for training supervised machine-learning systems on non-annotated data, were presented in (Yu and Hatzivassiloglou, 2003), which assumed that in a full-text, IMRaD-structured article, the majority of sentences in each section will be classified into their respective IMRaD category. Also, Agarwal and Yu (2009) used the same method to build a baseline classifier that achieved about 77.81% accuracy on their corpus.

3 Methodology

3.1 Constructing a self-annotating corpus from a biomedical dataset

The goal of this study is to show that the classification of sentences from scientific research papers to match the IMRaD rhetorical structure with supervised machine learning can be enhanced using a self-annotating corpus. The first task consists of the curation of a corpus that contains sentences representative of the defined categorization scheme. We have chosen to build the corpus by extracting sentences from a large repository of full-text scientific research papers, a publicly available full-text subset of the PubMed repository.

Since most demonstrative pronouns are co-referential, a sentence that begins with the demonstrative noun phrase “This method...” or “This result...” or “This conclusion...” is co-referential and its antecedents are likely to be found in previous sentences. Torii and Vijay-Shanker (2005) reported that nearly all antecedents of such demonstrative phrases can be found within two sentences. As well, Hunston (2008) reported that interpreting recurring phrases in a large corpus enables us to capture the consistency in meaning as well as the role of specific words in such phrases. So, the recurring semantic sequences “This method...” or “This result...” or “This conclusion...” in the Pubmed corpus can help us to capture valuable information in the context of their usage. A similar technique was used in (Houngbo and Mercer, 2012), to build a corpus for method mention extraction from biomedical research papers.

Our assumption is that a sentence that appears in the co-referential context of the co-referencing phrase “This method…”, will likely talk about a methodology used in a research experiment or analysis. Similarly, a sentence that starts with the expression “This result…” is likely to refer to a result. And, similarly, for sentences that begin with “This conclusion…”. The **Introduction** (Background) rhetorical category does not have a similar co-referential structure. We have chosen to only consider the immediately preceding sentence to the “This” referencing sentence. Some examples are shown below.

Category	# of Sentences	Proportion
Method	3163	31.9%
Result	6288	62.7%
Conclusion	534	5.4%
Total	9985	100%

Table 1: Initial Self-annotated Corpus Statistics

1. *We have developed a DNA microarray-based method for measuring transcript length …* **This method**, called the Virtual Northern, is a complementary approach …

2. *Interestingly, Drice the downstream caspase activated … was not affected by inhibition of Dronc and Dredd.* **This result**, … suggests that some other mechanism activates Drice.

3. *We obtained a long-range PCR product from the latter interval, that appeared to encompass the breakpoint on chromosome 2 …* **This conclusion**, however , was regarded with caution , since …

Table 1 shows the number of sentences per category in this initial self-annotated corpus.

3.1.1 Feature Extraction

We have used the set of features extracted from the Agarwal and Yu (2009) IMRaD corpus. The reason for this choice is to be able to validate our claim against this previous work. Agarwal and Yu (2009) experimented with mutual information and chi-squared for feature selection and obtained their best performance using the top 2500 features comprised of a combination of individual words as well as bigrams and trigrams. A feature that indicates the presence of a citation in a sentence is also used as it can be an important feature for distinguishing some categories; for example, citations are more frequently used in **Introduction** than in **Results**. All numbers were replaced by a unique symbol #NuMBeR. Stop words were not removed since certain stop words are also more likely to be associated with certain IMRaD categories. Words that refer to a figure or table are not removed, since such references are more likely to occur in sentences indicating the outcome of the study. We also used verb tense features as some categories may be associated with the presence of the present tense or the past tense in the sentence. We used the Stanford parser (Klein and Manning, 2003) to identify these tenses.

(a) Classification with Multinomial Naïve Bayes.

Class	Precision	Recall	F-Measure
Method	0.923	0.661	0.77
Result	0.627	0.813	0.708
Conclusion	0.68	0.821	0.744
Average	0.779	0.74	0.744

(b) Classification with Support Vector Machine

Class	Precision	Recall	F-Measure
Method	0.818	0.521	0.636
Result	0.511	0.908	0.654
Conclusion	0.923	0.226	0.364
Average	0.72	0.621	0.604

Table 2: Precision, Recall, F-measure : Classifier trained with the initial self-annotated corpus and tested on a reduced Agarwal and Yu (2009) corpus (Method, Result, Conclusion)

3.1.2 Self-annotation

In our first experiment we trained a model on the initial self-annotated corpus discussed above and tested the model on the Agarwal and Yu (2009) corpus. Table 2 shows F-measures that are below the baseline classifier levels. We suggest that there are two causes: many of the important n-grams in the larger corpus are not present in the 2500 n-gram feature set; and there is noise in the initial self-annotated corpus. To reduce the noise in the initial self-annotated corpus and to maintain the 2500 n-gram feature set we pruned our initial self-annotated corpus using a semi-supervised learning step using an initial model based on the Agarwal and Yu feature set and learned from the Agarwal and Yu corpus. We describe below the semi-supervised method to do this pruning of the initial self-annotated corpus.

Our method for categorizing sentences into the IMRaD categories does not work for the **Introduction** category, so from the Agarwal and Yu (2009) IMRaD corpus, we have extracted instances belonging to the **Method**, **Result** and **Conclusion** categories and have used this corpus to build a model with a supervised multinomial Naïve Bayes method. This model is then used to classify sentences in the initial self-annotated corpus. When the model matches the initial self-annotated corpus category with a confidence level greater than 98%, this instance is added to what we will now call the model-validated self-annotated corpus. The composition of this model-validated corpus is presented in Table 3.

Category	# of Sentences	Proportion
Method	878	23.6%
Result	2399	64.5%
Conclusion	443	11.9%
Total	3719	100%

Table 3: Model-validated Self-annotated Corpus Statistics

3.2 Automatic text classification

For all supervised learning, we have used two popular supervised machine-learning algorithms, multinomial Naïve Bayes (NB) and Support Vector Machine (SVM), provided by the open-source Java-based machine-learning library Weka 3.7 (Witten and Frank, 2005).

4 Results and Discussion

In the first classification task a classifier is trained with the model-validated self-annotated corpus using 10-fold cross-validation. The model achieves an F-measure score of 97% with NB and 98.7% with SVM. See Table 4. The average F-measure that Agarwal and Yu (2009) report for their 10-fold cross-validation (which includes **Introduction**) is 91.55. The category F-measures that Agarwal and Yu (2009) report for their 10-fold cross-validation with the features that we use are: **Method**: 91.4 (95.04) (their best scores, in parentheses, require inclusion of the IMRaD section as a feature), **Result**: 88.3 (92.24), and **Conclusion**: 69.03 (73.77).

In the last classification task, a classifier is trained with the model-validated self-annotated corpus and tested on the Agarwal and Yu (2009) corpus. The F-measures in Table 5 are a substantial improvement over those in Table 2.

(a) Classification with Multinomial Naïve Bayes.

Class	Precision	Recall	F-Measure
Method	0.981	0.957	0.969
Result	0.966	0.992	0.979
Conclusion	0.98	0.885	0.93
Average	0.971	0.971	0.971

(b) Classification with Support Vector Machine

Class	Precision	Recall	F-Measure
Method	0.986	0.984	0.985
Result	0.988	0.995	0.992
Conclusion	0.986	0.95	0.968
Average	0.987	0.987	0.987

Table 4: Precision, Recall, F-measure : Classifier trained with the model-validated self-annotated corpus (Method, Result, Conclusion) using 10-fold cross-validation

(a) Classification with Multinomial Naïve Bayes.

Class	Precision	Recall	F-Measure
Method	0.937	0.806	0.866
Result	0.763	0.873	0.814
Conclusion	0.836	0.911	0.872
Average	0.858	0.847	0.848

(b) Classification with Support Vector Machine

Class	Precision	Recall	F-Measure
Method	0.893	0.824	0.857
Result	0.763	0.85	0.804
Conclusion	0.835	0.811	0.823
Average	0.837	0.832	0.833

Table 5: Precision, Recall, F-measure : Classifier trained with the model-validated self-annotated corpus and tested on a reduced Agarwal and Yu (2009) corpus (Method, Result, Conclusion)

Sentence classification is important in determining the different components of argumentation. We have suggested a method to annotate sentences from scientific research papers into their IMRaD categories, excluding **Introduction**. Our results show that it is possible to extract a large self-annotated corpus automatically from a large repository of scientific research papers that generates very good supervised machine learned models.

Acknowledgments

This work was partially funded through a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to R. Mercer.

References

Shashank Agarwal and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Anita de Waard, Paul Buitelaar, and Thomas Eigner. 2009. Identifying the epistemic value of discourse segments in biology texts. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 351–354. Association for Computational Linguistics.

Hospice Houngbo and Robert E. Mercer. 2012. Method mention extraction from scientific research papers. In *Proceedings of COLING 2012*, pages 1211–1222, Mumbai, India.

Susan Hunston. 2008. Starting with the small words. Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics*, 13:271–295.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: A linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11):S10.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 423–430. Association for Computational Linguistics.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Ben Medlock and Ted Briscoe. 2007. Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 992–999. Association for Computational Linguistics.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. 2008. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1493–1502. Association for Computational Linguistics.

Manabu Torii and K. Vijay-Shanker. 2005. Anaphora resolution of demonstrative noun phrases in Medline abstracts. In *Proceedings of PACLING 2005*, pages 332–339.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition.

Mizuta Yoko, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75:468–487.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136. Association for Computational Linguistics.