Identifying Discourse Connectives in Biomedical Text

Balaji Polepalli Ramesh, MS[1], Hong Yu, PhD[1]
[1]University of Wisconsin Milwaukee, Milwaukee, WI

Abstract

Discourse connectives are words or phrases that connect or relate two coherent sentences or phrases and indicate the presence of discourse relations. Automatic recognition of discourse connectives may benefit many natural language processing applications. In this pilot study, we report the development of the supervised machine-learning classifiers with conditional random fields (CRFs) for automatically identifying discourse connectives in full-text biomedical articles. Our first classifier was trained on the open-domain 1 million token Penn Discourse Tree Bank (PDTB). We performed cross validation on biomedical articles (approximately 100K word tokens) that we annotated. The results show that the classifier trained on PDTB data attained a 0.55 F1-score for identifying discourse connectives in biomedical text, while the cross-validation results in the biomedical text attained a 0.69 F1-score, a much better performance despite a much smaller training size. Our preliminary analysis suggests the existence of domain-specific features, and we speculate that domain-adaption approaches may further improve performance.

1 Introduction

Discourse connectives are words or phrases connecting or relating two coherent sentences or phrases, and they indicate the presence of discourse relations. For example, observe the following: *in that SF pDCs showed only some evidence of activation in situ,* **but** *once incubated in mixed lymphocyte reactions in the absence of SF they displayed enhanced APC function relative to that of PB pDC.* This example contains a discourse relation in which the connective "but" relates two coherent sentences.

Automatically identifying discourse connectives and their arguments may benefit many natural language processing tasks including text summarization[1] and compression, text generation[2], dialogue understanding[3], scenario-level information extraction, question answering system, sentiment analysis[4], textual entailment[5] and temporal reasoning[6].

Identifying the discourse relations in biomedical domain can help in isolating the text that might indicate the existence of an event or interaction among various entities like drugs, adverse drug events. For example*, In the emergency department, he was given one dose of Solu-Medro 500 mg, however, he was found to have elevated cyclosporin levels at 679,* **so** *this was thought to be the likely cause of his acute renal failure and his cyclosporin was temporarily held*[7]. The above example, the connective "so" reveals a causal relation between cyclosporin (not Solu-Medro) and acute renal failure.

On the other hand, it is a challenging task to automatically identify discourse connectives and their scope. For example, earlier we showed an example of discourse connective "but." The same word "but" is not a discourse connective in the following example: *In the present study, the ameliorating anti-CD4 mAbs W3/25 and OX35 (***but** *not the accelerating mAb, RIB5/2) numerically/significantly increased the DTH to the arthritogen M. tuberculosis.*

Similarly, scope identification is also a hard task. In this study, we report approaches for automatically identifying discourse connectives, the first step towards the development of a full discourse relation parser.

2 Related Work

Discourse parsing has been an active field in the open domain. Soricut and Marcu[8] developed probabilistic models for identifying discourse units based on Rhetorical Structure Theory. Wellner et.al[9], Pitler et.al[10], and Elwell et.al[11] developed discourse parsers trained on the PDTB data.

Discourse studies have taken several directions in the biomedical domain. Light et al.[12] defined biomedical text as *fact*, *speculation,* and *in between*. As shown in Light et al., the sentence "*Pdcd4 may thus constitute a useful molecular target for cancer prevention*" contains fragments expressing a relatively high level of speculation (i.e., "may"), while the sentence "*However, NF-kappaB was increased at 3 h while AP-1 (Jun B and Jun D) and CREB were increased at 15 h*" expresses a fact. Wilbur et al.[13,14] defined five qualitative dimensions (i.e., *focus, polarity, certainty, evidence* and *directionality*) for categorizing the intention of a sentence. As shown in Wilbur et al., the

sentence "*we suppose that an increased LI in breast tissues of this group of patients may help explain the association between BC and thyroid autoimmunity*" only speculates on a possible explanation, while the sentence "*Hyphae-specific genes, HWP1, RBT4 and ECE1, were activated in the elongated filaments caused by the Cdc28p depletion*" provides the evidence.

Other work has identified discourse zones or units. For example, Mullen et al.[15] defined discourse zones of biomedical text including INTRODUCTION, METHOD, RESULT, and CONCLUSION and developed supervised machine-learning approaches to automatically classify a sentence into its rhetorical zone category. Biber and Jones[16] adapted unsupervised TextTiling methods[17] to segment biomedical text into different discourse units on the basis of lexical similarities among the units. Castano et al.[18] built a system for anaphora resolution in biomedical literature. Szarvas et al.[19] created BioScope, a corpus annotated with negative and speculative keywords and their linguistic scope in biomedical text. Agarwal et al.[20, 21] developed a system to automatically identify the negation and hedging cue and scope in biomedical text.

In this paper we focus mainly on automatically identifying discourse connectives. To our knowledge, we are the first group to investigate and annotate corpus-based discourse relations systematically in the biomedical domain and to investigate their applications for biomedical text mining.

3 Materials and Methods

We explored two collections of annotated data for training models to identify discourse connectives.

3.1 PDTB

The Penn Discourse Tree Bank (PDTB) [22] version 2.0 was used as the data set for the experiments. The PDTB set contains 2,159 files (a million tokens) that have been annotated in terms of argument structure, semantics, and attribution of discourse relations and their arguments. The PDTB follows a lexically-grounded approach to discourse structure[23,24], in which a discourse relation is a relation between abstract objects (AOs) mentioned in a text, such as events, states, and propositions[25]. In the PDTB, a discourse relation is considered as strictly binary, with its two AO arguments called Arg1 and Arg2. Relations in PDTB are broadly classified into two categories depending on how the relations are realized. The first type of relation is realized by the presence of an explicit connective. The second type of relation is realized between two adjacent sentences in the absence of an explicit connective. In this study, we focus only on identifying explicit discourse connectives in biomedical text.

3.2 BioDRB

The Biomedical Discourse Relation Bank (BioDRB)[27] is a collection of 24 articles (~100,000 word tokens), a subset of the GENIA corpus[26] that we have annotated for identifying the explicit discourse relations for our experiments. We followed the PDTB annotation guidelines to develop this corpus and reported 85% overall agreement among annotators[28] The subset contains annotated discourse relations and their arguments. In this study we attempt to develop a supervised machine-learning classifier to automatically recognize the discourse connectives; our classifier was trained upon the data we annotated[27].

Figure 1 below shows the plot of number of connectives versus their frequency in the BioDRB corpus. The plot follows a power law distribution: many discourse connectives (e.g., "one day after" and "followed by") occur only once, and a few discourse connectives (e.g., "and" and "however") are very frequent.

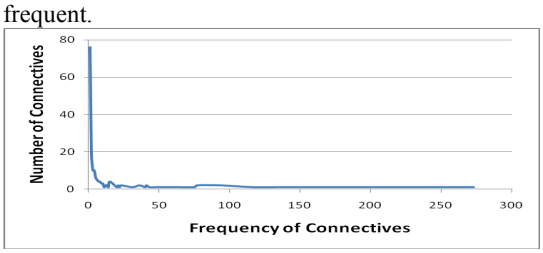

Figure 1: Plot of number of connectives against their frequency.

3.3 Distributional Differences between Discourse Connectives in the PDTB and BioDRB Data

The PDTB and BioDRB corpora contain annotations for 19,458 and 2,642 explicit relations, or 19.5 and 26.4 discourse connectives, respectively, per 1,000 tokens. There are 273 and 178 unique explicit discourse connectives in the PDTB and BioDRB corpuses, respectively.

The data shows that only 44% of the explicit discourse connectives in the BioDRB corpus occur in the PDTB corpus. For example, the discourse connectives that are present in both PDTB and BioDRB include the discourse connectives "and," "however," "also," and "so," all of which are common discourse connectives.

The remaining 56% of the connectives that occur in the BioDRB corpus do not appear in the PDTB

corpus as connectives. Examples include "followed by," "due to," and "in order to", which may suggest domain-specific characters of the BioDRB corpus.

3.4 Supervised Machine Learning and Learning Features

We explored conditional random fields (CRFs), a framework to build probabilistic models to segment and label sequence data for this task. A conditional model specifies the probabilities of possible label sequences given an observation sequence. The model was trained using ABNER, an open-source biomedical named entity recognizer[29]. We applied the default feature set, which comprises the standard bag-of-words, morphology, and n-gram features for this task.

We built three classifiers, all of which were built upon CRF models: *PDTB* was trained and tested on the PDTB corpus; *PDTB-BioDRB* was trained on the PDTB and tested on the BioDRB; and *BioDRB* was trained and tested on the BioDRB data.

4 Results

4.1 Evaluation Methods

We performed a 10-fold cross validation to evaluate *PDTB*. Since the BioDRB corpus incorporates 24 full-text biomedical articles, we performed a 12-fold cross-validation for *BioDRB*. In order to examine whether the training size has an impact on the performance, we performed 10-fold cross-validation experiments on 0.24, 0.48, 0.7 and 1 million tokens of the PDTB corpus.

4.2 Evaluation Metrics

We report recall, precision and F1-score, all of which are commonly used evaluation metrics in natural language processing applications. Recall is the number of correctly predicted discourse connectives divided by the total number annotated discourse connectives in the gold standard. Precision is the number of correctly predicted discourse connectives divided by the total number of predicted discourse connectives. F1-score is the harmonic mean of recall and precision. We also reported accuracy, which is the number of discourse connectives predicted to be correct divided by the total number of discourse connectives present in the corpus.

4.3 Evaluation Results

We report the cross-validation results of *PDTB, PDTB-BioDRB,* and BioDRB on identifying discourse connectives. Our 10-fold cross-validation experiments on 0.24, 0.48, 0.7 and 1 million tokens of the PDTB data showed little difference in performance (t-test was not significant), and we therefore report only *PDTB* and *PDTB-BioDRB*, which were trained on the entire PDTB dataset. Table 1 shows the results of different classifiers for identifying explicit discourse connectives. The difference between any pair is statistically significant (t-test, $p<0.01$).

	PDTB	***PDTB-BioDRB***	***BioDRB***
Precision	0.88±0.02	0.79±0.003	0.79±0.05
Recall	0.81±0.02	0.42±0.006	0.63±0.08
F1-score	0.84±0.01	0.55±0.005	0.69±0.05

Table 1: The performance (average±Std) of different models for identifying discourse connectives.

5 Error Analysis

We performed error analysis on the outputs of *PDTB-BioDRB* and *BioDRB*. We found that many discourse connectives appeared only once in the entire corpus. For example, "in order to" appear as discourse connective only once in the entire corpus and the classifier failed to recognize it as connective, hence had an accuracy of 0. Whereas the "Conversely" app eared as a discourse connective all the time in the corpus, so it had an accuracy of 100%.

Figure 2 below plots the accuracy of discourse connectives as a function of their frequencies.

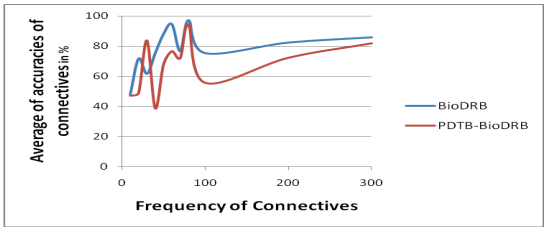

Figure 2: Plot of average of accuracies against the frequency of the connectives.

Figure 2 shows that there is a general trend in that when the frequency of discourse connectives increases, the accuracy of the classifier increases as well. Furthermore, when discourse connectives are frequent (>100), *BioDRB* significantly outperforms *PDTB-BioDRB*.

Connectives in BioDRB	*PDTB-BioDRB*	*BioDRB*
And	91.2%	90.84%
By	71.5%	77.37%
To	83%	89.03%
After	44%	66.26%
However	98.3%	98.3%

Table 2: The top 5 frequent connectives in BioDRB and their accuracy on the *PTDB-BioDRB* and *BioDRB* data.

Table 2 shows the top five frequent connectives that appear in the BioDRB corpus and the corresponding performance by *BioDRB* and *PDTB-BioDRB*. In accordance with Figure 2, *BioDRB* generally out-performed *PDTB-BioDRB*.

As shown in Table 3, of the 44% of the discourse connectives that are present in both the PDTB and BioDRB corpora (BioDRB ∩ PDTB), the two classifiers *PDTB-BioDRB* and *BioDRB* performed almost equally (>94%), while *BioDRB* edged out *PDTB-BioDRB* on recognizing discourse connectives that are present only in the BioDRB corpus (BioDRB ∉ PDTB).

		PDTB-BioDRB	*BioDRB*
BioDRB ∩ PDTB	44%	94.3%	94.9%
BioDRB ∉ PDTB	56%	89.6%	92.7%

Table 3: Accuracies of *PDTB-BioDRB* and *BioDRB*.

We manually analyzed the incorrectly classified instances. Examples 1─3, below, are some examples. The connectives are indicated by **bold**.

Example 1: **One day after** injection, the swelling of the ears was determined with a gauge (Hahn & Kolb, Stuttgart, Germany).

Example 2: The accelerating effect of the mAb RIB5/2 was reproduced in two additional treatment experiments, and this effect was observed **despite** a variable onset of AA in the PBS-treated animals (day 9 to 11); i.e. in all experiments, the onset of AA occurred 2 days earlier than in the controls.

Example 3: that clinical efficacy (and its time course) may depend on the actual immunological constellation **and** that a given anti-CD4 mAb may have beneficial effects only in particular pathologies and/or stages of disease.

Both *PDTB* and *PDTB-BioDRB* failed to recongize the discourse connective "one day after" in example 1. In example 2 *BioDRB* recognized the connective "despite" and *PDTB-BioDRB* failed to recognize it. Example 3 illustrates a case in which *PDTB-BioDRB* recognized the connective "and" and *BioDRB* failed to recognize it.

6 Discussion and Conclusion

In this paper we report the performance of three different classifiers *PDTB*, *PDTB-BioDRB,* and *BioDRB* on detecting discourse connectives in text. *PDTB* was trained on the PDTB corpus, and our results show an F1-score of 0.84. The results of *PDTB* out-performed both *PDTB-BioDRB* (0.55 F1 score) and *BioDRB* (0.69 F1 score) and the results of *BioDRB* outperformed the results of *PDTB-BioDRB*. These results are not surprising because the two domains are very different in that PDTB is a set of news articles and BioDRB consists of biomedical text. The differences can be demonstrated by the fact that 56% of the discourse connectives in the BioDRB corpus do not appear in the PDTB data.

Figure 1 shows that that the distribution of number of discourse connectives as a function of accuracy is similar in both *PDTB-BioDRB* and *BioDRB,* even though *BioDRB* performed higher than *PDTB-BioDRB* overall. The results suggest that the difference in performance may be caused by a few discourse connectives. This hypothesis is validated because our results show that *BioDRB* outperformed *PDTB-BioDRB* when discourse connectives are frequent (>100), although the number of such frequently occurring discourse connectives are few (only 5). As shown in Table 2, *BioDRB* outperformed *PDTB-BioDRB* because of the performance difference in the three most frequently occurring discourse connectives: "by," "to," and "after".

For the common discourse connective "and," *PDTB-BioDRB* performed slightly better than *BioDRB*. An example is shown in Example 3 in which *PDTB-BioDRB* detected the discourse connective "and" and *BioDRB* failed. The results may be in part due to the fact that the training size in the BioDRB corpus is too small.

Example 1 is another case that is due to the small training size. In addition, we found that *PDTB-BioDRB* did not recognize connectives that appeared in the beginning of the sentence (e.g., example 1). The performance of the classifier may be further improved by adding features from the previous sentence.

Our results show that 56% discourse connectives appearing in the BioDRB do not appear in the PDTB, and as a result, *PDTB-BioDRB* failed to detect them. Example 2 is such a case.

The state-of-the-art model for recognizing discourse connectives has a F1 score of 0.94[30], but it relies on the use of rich syntactic features and a parser to attain this score. In our model, we did not include any such features as these features are not annotated in the gold standard, which is the annotated corpus. Furthermore, our results suggest that recognizing discourse connectives in biomedical text might pose challenges that have not been explored in depth.

In the future, we can extend this work to recognize discourse relations that appear in biomedical text, as discourse parsing has many applications in various NLP tasks.

Acknowledgement: We thank Dr.Rashmi Prasad for fruitful discussion and acknowledge the support of a

University of Wisconsin-Milwaukee Graduate School grant to Dr.Hong Yu.

References

1. Marcu, D. Improving summarization through rhetorical parsing tuning. In The Sixth Workshop on Very Large Corpora 1998; 206–215.
2. Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. Artificial Intelligence 1993, 63(1-2), 341–385.
3. H. Hernault, P. Piwek, H. Prendinger, and M. Ishizuka. Generating dialogues for virtual agents using nested textual coherence relations. In Proceedings of the 8th International Conference on Intelligent Virtual Agents 2008;139–145.
4. R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007.
5. B. MacCartney and C. D. Manning. Natural logic for textual inference. In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing 2007;193–200.
6. I. Mani, M. Verhagen, B. Wellner, et al. Machine learning of temporal relations. ACL 2006;753–760.
7. Uzuner Ö, Szolovits P, Kohane I.: i2b2 Workshop on Natural Language Processing challenges for clinical records. AMIA Annu Symp Proc. 2006.
8. Soricut R and Marcu D. Sentence level discourse parsing using syntactic and lexical information. In Proc of HLT-NAACL 2003;149-156.
9. Wellner B, Pustejovsky J, Havasi C, Rumshisky A and Sauri R. Classification of discourse coherence relations: an exploratory study using multiple knowledge sources. The 7th SIGdial Workshop on Discourse and Dialogue 2006; 117-125.
10. Pitler E, Louis A and Nenkova A. Automatic sense prediction for implicit discourse relations in text. In Proc of 47th annual meeting of the ACL and the 4th IJCNLP of the AFNLP 2009;683-691.
11. Elwell R and Baldridge J. Discourse connective argument identification with connective specific rankers. The 2008 IEEE International Conference on Semantic Computing 2008;198-205.
12. Light M, Qiu X, Srinivasan P: The language of bioscience: fact, speculations, and statements in between. In BioLink: Linking Biological Literature, Ontologies and Databases., 2004:17-24.
13. Wilbur WJ, Rzhetsky A, Shatkay H: New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 2006;7:356.
14. 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.
15. 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;7:52-58.
16. Biber D, Jones J: Merging corpus linguistic and discourse analytic research goals: Discourse units in biology. Corpus Linguistics and Linguistic Theory 2005.
17. Hearst M: TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics 1997;23:33-64.
18. Castano J, Zhang J, Pustejovsky: Anaphora resolution in biomedical literature. In International Symposium on Reference Resolution. 2002.
19. Szarvas G, Vincze V, Farkas R, Csirik J: Bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of Current Trends in Biomedical Natural Language Processing 2008.
20. Agarwal S, Yu H. Biomedical negation scope detection with Conditional Random Fields. JAMIA. Manuscript under review. 2010
21. Agarwal S, Yu H. Detecting Hedge Cues and their Scope in Biomedical Literature with Conditional Random Fields. JBI. Manuscript under review. 2010
22. Miltsakaki E, Prasad R, Joshi A and Webber B, The Penn Discourse Treebank. www.seas.upenn.edu/~pdtb/.
23. Webber B, Joshi A: Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In Discourse Relations and Discourse Markers: Proceedings of the Conference. Edited by Stede M, Wanner L, Hovy E, Somerset. Association for Computational Linguistics 1998:86–92.
24. Webber B, Joshi A, Stone M, Knott A: Anaphora and Discourse Structure. Computational Linguistics 2003, 29(4):545–587.
25. Asher N: Reference to Abstract Objects. Dordrecht: Kluwer 1993.
26. Kim J, Ohta T, Tateisi Y, Tsujii J: GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics 2003, 19 Suppl 1:i180–182.
27. R Prasad , S Mcroy, N Frid, H Yu, and A Joshi. BioDRB: The Biomedical Discourse Relation Bank 2010. *In preparation*.
28. Yu H, Frid N , McRoy S, et al. A pilot annotation to investigate discourse connectivity in biomedical text. BioNLP, 2008.
29. Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005; 21:3191-3192.
30. E Pitler and A Nenkova. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers 2009;13-16.