

BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text

Sindhuja Gopalan

AU-KBC Research Centre
MIT Campus of Anna University
Chromepet, Chennai, India
sindhujagopalan@au-kbc.org

Sobha Lalitha Devi

AU-KBC Research Centre
MIT Campus of Anna University
Chromepet, Chennai, India
sobha@au-kbc.org

Abstract

This paper describes a Natural language processing system developed for automatic identification of explicit connectives, its sense and arguments. Prior work has shown that the difference in usage of connectives across corpora affects the cross domain connective identification task negatively. Hence the development of domain specific discourse parser has become indispensable. Here, we present a corpus annotated with discourse relations on Medline abstracts. Kappa score is calculated to check the annotation quality of our corpus. The previous works on discourse analysis in bio-medical data have concentrated only on the identification of connectives and hence we have developed an end-end parser for connective and argument identification using Conditional Random Fields algorithm. The type and sub-type of the connective sense is also identified. The results obtained are encouraging.

1 Introduction

Due to advancements in bio-medical field, a large number of bio-medical literatures are available. It is crucial to extract knowledge from these literatures, to prevent the loss of important findings required for progression in bio-medical field. For extracting the information from the text, the text needs to be analyzed linguistically and it is absolutely essential to add such linguistic information to the text for future research. Natural language processing (NLP) methods are being widely used to analyse bio-medical text by performing tasks like automatic summarization, translation, named entity recognition, discourse analysis, speech recognition, etc.,. Discourse analysis is one such fundamental topic in NLP domain that makes a text linguistically rich. Discourse analysis is the study of the relation between phrase, clauses or sentences in a text. The basic units of discourse relations are discourse markers and their arguments. Discourse markers are words or phrases that establish a relation between two discourse units thereby connecting two events.

Example [1]

- a) *Embryonic Stem cells have a high mitotic index and form colonies. So, experiments can be completed rapidly and easily.*
- b) Clinical lung cancers containing a higher abundance of *ALDH* and *CD44* co expressing cells were associated with lower recurrence free survival.

In the above Example 1 (a) “so” is the discourse connective that indicates a relation between two discourse units. These discourse units are labeled as arguments. It is not easy to identify the discourse connectives, as all connectives are not discourse markers. In some cases it acts as conjunction simply unifying two words. Consider the Example 1 (b), where “and” acts as a conjunction connecting two bio-medical named entities “ALDH” and “CD44 co expressing cells”.

The intended goal of our present work is to study the “discourse relation” in Medline abstracts and develop a system to automatically extract these relations. Cohen et al., (2010) has examined the structural and linguistic aspects of abstracts and bodies of full text articles. Their work shows that full parsing

This work is licensed under a Creative Commons Attribution 4.0 International License. License details:
<http://creativecommons.org/licenses/by/4.0/>

of article bodies is more difficult than abstracts as article bodies has longer sentences. They also assessed the incidence of conjunctions in abstracts and article bodies. The difference between abstracts and bodies was not statistically significant with only slightly more conjunction in article bodies than in abstracts. Hence, we used abstracts to develop the system instead of whole articles. Since abstracts give an overview of a work, the sentences in the abstract need to be coherent. Hence it is well connected by connectives. The occurrence and type of discourse connectives in the bio-medical domain vary from other domain.

Example [2]

Furthermore, juglone blocked the adipogenic medium-induced activation of PPAR, C/EBP, C/EBP, and ERK pathways, which was rescued by Ad-PIN1 infection. In summary, the present study shows for the first time that PIN1 acts as a significant modulator of odontogenic and adipogenic differentiation of HDPSCs, and may have clinical implications for regenerative dentistry.

In the above Example 2, the discourse marker “In summary” occurs commonly in bio-medical abstracts, whereas in general domain it is used to a lesser extent. There are many such connectives that occur in bio-medical domain and may or may not occur in other domains. Also, work done by Ramesh et al., (2012) show that the classifier trained on open domain performs poorly on bio-medical domain. Hence there is a demand to spring up a discourse parser for bio-medical domain.

Our work makes a substantial contribution to the development of the discourse annotated corpus for bio-medical domain. At this period, there is no end-end discourse parser available for automatically identifying the connectives, sense and its arguments in bio-medical domain; hence we have developed a discourse parser for explicit relation identification from bio-medical corpus. This system can further be used in NLP tasks like machine comprehension, extraction of semantic relations, co-reference resolution, etc. In the next section, works related to discourse analysis are detailed. Section 3 describes the annotation task. Section 4 explains the method used. Results are demonstrated and discussed in section 5. We conclude and outline our future work in section 6.

2 Related Works

In this section, we describe the annotation works and various systems developed for automatic identification of the discourse relations.

2.1 Annotation

Large scale annotated corpora for discourse analysis is developed in recent years that play a major role in natural language research. Lopatkova et al., (2009) in their project has assigned structure of clauses to Czech sentences from the Prague dependency tree, but as a new stratum of syntactic annotation. The PDTB adds low level discourse structure and semantics using connective specific semantic role labels (Prasad et al., 2008a). Dinesh et al., (2005) in their paper has disclosed substantial divergences among the syntactic structure and discourse structure in terms of the arguments of connectives. Miltsakaki et al., (2005) has presented a set of manual sense annotation studies for three connectives since, while and when. The work on discourse connectives also extends to other languages. The discourse relation has been studied in various languages like Turkish (Zeyrek and Webber, 2008), French (Roze et al., 2010), Arabic (Al-Saif et al., 2010), Hindi (Prasad et al., 2008b), Tamil (Rachakonda and Sharma, 2011; Menaka et al., 2011) and cross lingual variation is analyzed in Indian languages like Malayalam, Hindi and Tamil (Devi et al., 2014). Discourse annotation work has also been widened to bio-medical domain by Mihaila et al., (2013) and Tateisi et al., (2000)

The first work on the annotation of bio-medical discourse connectives and its arguments was done by Prasad et al., (2011). They have developed a Biomedical Discourse Relation Bank (BioDRB) in which they have annotated explicit and implicit discourse relation in 24 open access full text bio-medical articles from GENIA corpus containing 4911 sentences. They have adapted the annotation guidelines from PDTB, but have introduced new conventions and modification for sense classification. The related works show that BioDRB is the only discourse relation tagged corpus available in the bio-medical field. Motivated by the need to construct a corpus to be used for the text mining in the bio-medical field, we

developed a discourse annotated corpus using Medline abstracts.

2.2 Systems Developed

Existing systems for connective and argument identification are developed based on the PDTB. Elwell and Baldrige (2008) shows that using models for specific connectives and the types of connectives and interpolating them with a general model improves performance. Lin et al., (2012) has developed a full discourse parser using the parsing algorithm in PDTB style. This parser first identifies all discourse and non-discourse relations, locates and labels their arguments, and then classifies their relation types. They reported overall system F-scores for partial matching of 46.80% with gold standard parser and 38.18% with full automation. Ghosh et al., (2011) in their work has taken a data driven approach to identify arguments of explicit discourse connectives using Conditional Random Fields (CRFs) technique and obtained an F-score of 57% for arg1 and 79% for arg2. Wellner et al., (2007) has worked on discovering the arguments of discourse connectives in the PDTB. Rather than identifying the full-extent of the arguments as annotated in the PDTB, they identified the argument heads. Using log-linear re-ranking model, they identified both the arguments correctly for over 74% of the connectives on held-out test data using gold standard parsers.

Stepanov and Riccardi (2014) in their paper has presented cross-domain evaluation of PDTB trained discourse relation parser and evaluated feature-level domain adaptation proficiencies on the argument span extraction sub task. They summed up that the corpora differences with respect to discourse connective usage affect the cross domain generalization of connective detection tasks negatively. Hence, it is necessary to develop a domain specific system for identification of discourse connectives. Ramesh et al., (2012) has developed a system for identification of discourse connectives in bio-medical domain and has obtained an F-score of 69%. But, they did not focus on identification of arguments. The main goal our work is to develop a system for automatic identification of explicit connectives, its sense and arguments using machine learning approach. The experiment outcomes manifested the efficiency of our system in discourse relation identification task.

3 Annotation Task

Our annotation task includes the annotation of discourse connectives, its sense and their arguments by following the guidelines of PDTB, a large-scale resource of annotated discourse relations and their arguments (Prasad et al., 2008a). We formulated our corpus by collecting abstracts from PubMed Central (PMC). PMC is a free full-text archive of bio-medical and life sciences journal literature developed by the U.S. National Institutes of Health’s National Library of Medicine (NIH/NLM). To analyze the distribution of discourse connectives and its arguments, a corpus containing 7670 sentences is used. The bio-medical corpus is annotated with Explicit, Implicit, Altlex, NoRel and Entrel relations. The syntactic classification of connectives includes Subordinators, Coordinators, Conjunct adverbs, and Correlative conjunction.

Example [3]

*The practical application of ESCs is throttled_{arg1} **because** it is unmanageable to derive and culture ESCs_{arg2}.*

In Example 3, the subordinator “because” occurs in the middle of the sentence connecting main clause with the subordinate clause. The sense of the discourse connectives provide a semantic description of the relation between the arguments of connectives. Based on the sense, these connectives are broadly classified into four top class levels viz Expansion, Temporal, Comparison and Contingency. These classes are further classified into various types and sub-types. For instance in Example 3, the connective “because” comes under sense class “Contingency” and belongs to the type “Cause” and sub-type “reason”.

The arguments are labeled as arg1 and arg2. The arguments include single clauses or multiple clauses and in some cases it may include whole sentences and even multiple sentences. The clause or sentence that is syntactically attached to connective is labeled as arg2 and the other clause or sentence is marked as arg1. Furthermore the arguments may be adjacent or non adjacent to the connectives. According to the concept of minimality, the minimum required argument for a relation is annotated. There are

no constraints for the linearity of the connectives and arguments. The linear order of the connectives and arguments need not be of the basic order arg1-con-arg2. It varies depending on the location of the connective. The connective can occur at sentence initial, medial or final position. In cases, where connective occurs in the initial position, the linear order is of the form con-arg2-arg1.

In case of implicit connectives, we have tagged the relation with a label “Implicit”. Implicit relation can be inferred, where a relation exists between two discourse units but not explicitly marked.

Example [4]

*In adipogenic cultures, both cell populations showed positive Oil Red O staining by day 21_{arg1}. **Implicit (Likewise, Similarly)** in chondrogenic cultures, both stem cells expressed the formation of proteoglycan_{arg2}.*

The above Example 4 shows that there exists an implicit relation between two discourse units. The relation can be established by connectives like “likewise”, “Similarly” etc.

AltLex relation is realized between adjacent sentences, where inserting an implicit connective may lead to redundancy in the expression of the relation.

Example [5]

*Dichaete potentially regulates many more genes in the Drosophila genome and was found to be associated with over 2000 mapped regulatory elements. **AltLex [Our analysis suggests]** that Dichaete acts as a transcriptional hub, controlling multiple regulatory pathways during CNS development.*

In the above Example 5 the AltLex relation is shown. Here, inserting a connective may cause redundancy. Hence AltLex relation is marked in such cases.

EntRel relation exists where the implicit relation between adjacent sentences is not between their abstract object interpretations, but form an entity based coherence. The entity is realized in both the sentences directly or indirectly.

Example [6]

*Brg1 is required for stem cell maintenance in the murine intestinal epithelium in a tissue-specific manner. **EntRel** Brg1 is a chromatin remodeling factor involved in mediation of a plethora of signaling pathways leading to its participation in various physiological processes both during development and in adult tissues.*

In Example 6, the entity “Brg1” in first discourse is directly realized in the second unit.

Norel relation exists where there is no discourse relation or entity based coherence relation between adjacent sentences. Since our corpus contains abstracts, all sentences within abstracts are related explicitly or implicitly or entity based coherence are found. However, in our corpus NOREL relation can be tagged between abstracts, as shown in Example 7.

Example [7]

*TGF-1 alone and in combination with PDGF also amplified surface integrin expression and adhesivity of MSCs with extracellular matrix proteins. These findings will provide a more mechanistic insight for modeling tissue-level rigidity in fibrotic tissues and tumors. **NoRel** Lung cancer tumorigenicity and drug resistance are maintained through ALDH(hi)CD44(hi) tumor initiating cells.*

The data statistics are shown in the Table 1.

S.No	Connective Types	No of Connectives
1	Explicit	2957
1.a	Intra-sentential	1742
1.b	Inter-sentential	1215
2	Implicit	1610
3	Altlex	616
4	EntRel	802
5	NoRel	585

Table 1: Data statistics.

We annotated our corpus with above relations and cross checked the annotation quality by tagging

3000 sentences from the data using second annotator. We calculated the inter-annotator agreement using Cohens Kappa coefficient (Viera and Garrett, 2005), which is a statistical measure.

$$K = (p_o - p_c) / (1 - p_c)$$

where p_o is the agreement rate between two human annotators and p_c is chance agreement between two annotators. The agreement between the annotators is almost perfect for connectives. We obtained Cohens kappa score of .94 for explicit connectives. In the case of annotation of arguments there is a substantial agreement between the annotators for all the argument boundaries. The overall agreement in identifying both the arguments of explicit connectives is 0.86. The variation in agreement rate is due to various structural inter-dependencies that occur between discourse relations.

4 Method

This section describes the experiments performed for extraction of discourse relations using machine learning (ML) approach. In this work we have concentrated on the automatic identification of explicit relations. We performed two sets of tasks. In the first task the connectives and its sense were identified and in the second task the argument boundaries were identified. In further sections experiments are explained in detail.

4.1 Features Used

For our work we have used simple and minimal number of features given in Table 2 and Table 3. The connectives are mostly conjunction and hence the PoS features contribute most to the identification of connectives. Chunk features help to identify the boundary of the connectives and arguments. Since a discourse connective connects two clauses, clause start and end can be used as feature for connective identification. For sense identification, we have used syntactic features. In addition to features used for connective identification, connective itself is used as a feature for identifying the sense of the connective.

S.No	Features for Connectives	Examples
1	Word	previous word, current word, next word
2	PoS (P)	PoS of previous word, PoS of current word, PoS of next word
3	Combination of PoS and Chunk	PoS and Chunk of current word
4	Combination of 1, 2, 3	Current word, PoS of current word, Chunk of current word
5	Connective	Connective is an exceptional feature for sense identification

Table 2: Features used for connective identification (syntactic and sense).

S.No	Features for Arguments	Examples
1	Sentence position with respect to connective	arg1 end mostly will be before connective and arg2 start after connective
2	Sentence boundary	arg1 start mostly start of the sentence, arg2 end- mostly end of the sentence
3	Clause	Arguments may be clause and hence clause boundary is used as feature
4	Previously identified Argument boundaries	Previously identified arguments arg1 end, arg2 start, arg 1 start

Table 3: Features used for argument identification.

4.2 Experiments Performed

The annotated corpus was preprocessed, before training it using ML algorithms. The sentences were tokenized and PoS tags and chunks were added using the GENIA tagger (Tsuruoka et al., 2005). After analyzing the corpus, features were extracted. Based on the extracted features, language models were built using CRFs algorithm. We used CRF++ tool (Kudo, 2005), an open source implementation of CRFs algorithm. Using the language model the explicit connectives, its sense and argument boundaries were automatically identified from the test set. The experiments were performed as two tasks.

Connective classification and sense identification: In the first task, the system was trained for classification of connectives. Using the features described in section 4.1, the model was built for connective identification. The built model was used for identifying the connectives from the test data. Further, post-processing rules were applied to improve the system’s performance. After classifying the tokens as connectives and non-connectives, sense of the connectives were identified. In our work we have also identified the type and sub-type of the connective sense. We developed one-stage model and multi-stage model to identify the sense. For one stage model, a single model is developed for all types of sense. While for multi-stage model, four separate models were developed for each type of sense based on its upper level class. After identifying the senses separately, the output was combined from each model. The connectives having overlapping senses were merged based on confidence scores (i.e. sense having large probability).

Argument identification: After identifying the connectives, the second task was performed, where the arguments were identified. To overcome the problem of overlapping sequence, we processed each connective separately for argument identification. We developed two types of models, first by partitioning these sentences into inter and intra-sentential relations and second as a one-stage model without partitioning the sentences. Intra-sentential connectives are those that occur within a sentence, while inter-sentential connectives are those that occur outside the sentence. For inter-sentential connectives mostly previous sentence acts as argument. In few cases the arguments span across sentences. For the one-stage model the output from connective identification was given as such after extracting the features without dividing them as inter-sentential or intra-sentential connectives. We developed gold standard parser and automatic parser for argument identification using features mentioned in Table 3 .

For gold standard parser the gold standard connectives were used to train the system. For automatic parser the output from the connective identification task was fed as input to the argument identification task. For identifying the arguments we followed the method used by Menaka et al., (2011). They presented their work on automatic identification of the cause-effect relation from Tamil text. In their work they developed separate models for each boundary. Similarly, we built 4 models for each boundary of the arguments, i.e. identification of arg1 start and end and arg2 start and end. The argument boundaries were identified in the following series, arg2 start, arg1 end, arg1 start and arg2 end. The output from one model is fed as input to the next model. The choice of order of identification of bounds was made with the idea that it is easier to identify the boundaries that are close to the connective. After identifying the boundaries, the outputs were merged. Thus connective, sense and arguments were identified. The results are detailed in the next section.

5 Results and Discussion

We evaluated the performance of our system using precision, recall and F-score measure. Precision is the number of labels correctly perceived by the system from the total number of labels identified, Recall is the number of labels correctly detected by the system by the total number of labels contained in the stimulus text and F-score is merely the mean of precision and recall. We performed 10 fold cross-validations for connective identification using CRFs. The partition of the corpus was done randomly to test all the relations at least once. The average 10 fold cross-validation F-score obtained for CRFs is 86.42%. The result for connective classification obtained from best model is presented in Table 4. For sense identification the results for one-stage model and multi-stage model are presented in Table 5. For sense identification one-stage model gives better results than multi-stage model for both gold and automatic parser.

Method	Precision (%)	Recall (%)	F-score (%)
CRFs	92.49	83.83	88.16
CRFs + Post-processing	93.01	88.05	90.53

Table 4: Results for connective classification.

Model	Gold Parser			Automatic Parser		
	Precision	Recall	F-score	Precision	Recall	F-score
One-stage model	98.03	93.26	95.65	90.79	79.25	85.02
Multi-stage model	97.79	91.22	94.51	90.87	78.13	84.5

Table 5: Results for sense identification in %.

Then the arguments of the connectives were identified using intra-sentential and inter-sentential model and also using one-stage model. It is observed that the F-score for identification of intra-sentential and inter-sentential argument boundaries arg1 end and arg2 start is better than arg1 start and arg2 end. This is because the argument boundaries arg1 end and arg2 start are nearer to discourse connective in intra-sentential model. While, in inter-sentential model the argument boundary arg1 end will be mostly the sentence or clause end and arg2 start will succeed the connective. The argument identification results for intra-sentential argument and inter-sentential arguments are shown in Table 6 and 7 respectively. The results for one-stage model is given in Table 8.

Arguments	Gold Parser			Automatic Parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 start	80.95	81.20	81.08	81.27	76.39	78.83
Arg1 end	94.11	89.24	91.68	84.31	81.2	82.76
Arg2 start	94.20	93.24	93.72	93.1	90.89	91.99
Arg2 end	85.31	80.54	82.93	83.3	81.56	81.93

Table 6: Results for intra-sentential argument identification in %.

The system achieved significant performance even with minimal features. The performance of existing systems described in Section 2.2, shows that the performance of our system is comparable to state-of-art systems. The errors in the identification of connectives and arguments were analysed. After identifying the errors we developed post-processing rules to improve the results. We analyzed the output obtained from the system and observed that the decrease in measures in automatic identification of connectives is due to data sparsity. The connective patterns that exist in test data may not exist in the training data. Also, the difficulty in identification of connectives arises due to propagation of errors from preprocessing modules. As we use PoS as a feature for connective identification the error introduced in PoS module decreases the measure. Importantly, conjunctions are not connectives.

Example [8]

BMECs are important components of the hematopoietic microenvironment in the bone marrow_{arg1} and they can secrete several types of cytokines_{arg2}.

In Example 1 (b) the conjunction “and” connect two noun “ALDH” and “CD44 co-expressing cells”. Hence “and” does not act as a discourse connective. Whereas, in Example 8 “and” connects two discourse units as a whole and hence in this case it acts as a discourse connective. This ambiguity in the identification of connectives creates false positive results. To overcome this problem we have formulated some linguistic rules and have applied to the CRFs output. We explain an example for linguistic rule used in our work.

Rule1:

If the current token is “and”, previous token is “;”, PoS of previous to previous token is “VBN” and PoS of the next token is a “noun”, then “and” is a discourse connective.

Arguments	Gold Parser			Automatic Parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 start	86	78	82	84	76.6	80.3
Arg1 end	86.5	83.4	84.9	85.6	82.4	84
Arg2 start	90.5	90.5	90.5	88	86.8	87.5
Arg2 end	82.1	81.5	81.8	80.5	76.3	78.3

Table 7: Results for inter-sentential argument identification in %.

Arguments	Gold Parser			Automatic Parser		
	Precision	Recall	F-score	Precision	Recall	F-score
Arg1 start	82.9	80.1	81.5	83.4	76.8	80.1
Arg1 end	96.3	95.5	95.9	86.6	86.8	86.7
Arg2 start	99.4	98.2	98.8	92.1	85.4	88.8
Arg2 end	87.3	87.5	87.4	87.4	82.1	84.7

Table 8: Results for one-stage model argument identification in%.

Rule2:

If the current token is “also”, PoS of previous token is “JJ” and PoS of the next token is “RB”, then “also” is a discourse connective.

In the identification of arguments, the paired connectives generate errors. The paired connectives or co-relative conjunction includes two connectives that share same arguments. The error occurs in the argument boundary identification, when the argument includes multiple sentences. As we consider only the sentence with intra-sentential connective as input to intra-sentential model for argument identification the difficulty occurs in identification of argument with multiple sentences. Connectives can occur at the beginning or in the middle of the sentences or sometimes at the end. Consider below Example 9, where connective “also” occurs in the middle of the argument. This creates an error in identification of arg2 start.

Example [9]

*IL-17 activates several downstream signaling pathways including NF-B, MAPKs and C/EBPs to induce gene expression of antibacterial peptides, proinflammatory chemokines and cytokines and matrix metalloproteinases (MMPs)_{arg1}. IL-17 can **also** stabilize mRNAs of genes induced by TNF_{arg2}.*

The error analysis indicates the need for more sophisticated features to further improve the precision of the system.

6 Conclusion

We have developed a discourse relation bio-medical corpus, annotated with discourse connectives and its arguments. In this paper, we have explained the guidelines used for annotating our corpus. The inter-annotator agreement is calculated to check the annotation quality. We obtained almost perfect agreement between annotators for connectives and substantial agreement for argument boundaries. Also, a method for identification of discourse connectives and arguments from the annotated corpus using the ML approach is presented. We obtained encouraging results even with minimal features. In our future work, we will try to resolve the errors identified, thereby improving the overall results of the parser. Also, we will evaluate our system using BioDRB corpus, so that performance of our system on full articles can be verified. We will also extend our work in developing a discourse parser for identifying the implicit, EntRel and Altlex relations.

References

- Al-Saif, Amal, and Katja Markert. 2010. The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. *Proceedings of the 6th International Conference on Language Resources and Evaluation*.
- K Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder and Lawrence .E Hunter 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11:492.
- Anthony J. Viera and Joanne M. Garrett. 2005. Understanding Interobserver Agreement: The Kappa Statistic. *Research Series*, 37(5):360–363.
- Balaji P. Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *J Am Med Inform Assoc*, 19(5):800–808.
- Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 92–101.
- Charlotte Roze, Laurence Danlos and Philippe Muller. 2010. LEXCONN: a French lexicon of discourse connectives. *Proceedings of Multidisciplinary Approaches to Discourse*, Moissac, France, 114–125.
- Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2.
- Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. *International Joint Conference on Natural Language Processing*, Hyderabad, India.
- Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi and Bonnie Webber. 2005. Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives. *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, Barcelona.
- Evgeny A. Stepanov and Giuseppe Riccardi. 2014. Towards Cross-Domain PDTB-Style Discourse Parsing. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis*, Gothenburg, Sweden, 30–37.
- Marketa Lopatkova, Natalia Klyueva and Petr Homola. 2009. Capturing the Relationship among Clauses in Czech Sentences. *Proceedings of the Third Linguistic Annotation Workshop*, Suntec, Singapore, 74–81.
- S Menaka, Patabhi RK. Rao and Sobha L. Devi. 2011. Automatic identification of cause-effect relations in Tamil using CRFs. *Proceedings of Computational Linguistics and Intelligent Text Processing*, Springer, Berlin, Heidelberg, 316–327.
- Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi and Bonnie Webber. 2005. Attribution and the (Non-)Alignment of Syntactic and Discourse Arguments of Connectives. *Proceeding Corpus Anno '05 Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, Stroudsburg, PA, USA, 29–36.
- Ravi T. Rachakonda, and Dipti M. Sharma. 2011. Creating an Annotated Tamil Corpus as a Discourse Resource. *Proceedings of the Fifth Law Workshop (LAW V)*, Portland, Oregon, 119-123.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008a. The Penn discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation*.
- Rashmi Prasad, Husain S, Dipti M. Sharma and Aravind Joshi. 2008b. Towards an Annotated Corpus of Discourse Relations in Hindi. *International Joint Conference on Natural Language Processing*. Hyderabad, India.
- Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi and Hong Yu. 2011. The biomedical discourse relation bank. *BMC Bioinformatics*, 12:188.
- Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. *Proceedings of the IEEE International Conference on Semantic Computing*, Washington, DC, USA, 198–205.
- Sobha L. Devi, Lakshmi Sreedhar and Sindhuja Gopalan. 2014. Discourse Tagging for Indian Languages. *Lecture Notes in Computer Science*, 8404:470–481.

- Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi, and Sara Tonelli. 2011. Shallow Discourse Parsing with Conditional Random Fields. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 1071–1079.
- Taku Kudo. 2005. CRF++, an open source toolkit for CRF. <http://crfpp.sourceforge.net>.
- Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. *Lecture Notes in Computer Science*, 3746:382–392.
- Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata and Jun-ichi Tsujii. 2000. Building an Annotated Corpus in the Molecular-Biology Domain. *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, 28–36.
- Ziheng Lin, Hwee Tou Ng and Min-Yen Kan. 2012. A PDTB-Styled End-to-End Discourse Parser. *Natural Language Engineering*, 1(1):1–35.