Identification and Disambiguation of Lexical Cues of Rhetorical Relations across Different Text Genres

Taraneh Khazaei
Department of Computer Science
University of Western Ontario
London, ON, Canada
tkhazae@uwo.ca

Lu Xiao
Faculty of Information & Media Studies
University of Western Ontario
London, ON, Canada
lxiao24@uwo.ca

Robert E. Mercer
Department of Computer Science
University of Western Ontario
London, ON, Canada
mercer@uwo.ca

Abstract

Lexical cues are linguistic expressions that can signal the presence of a rhetorical relation. However, such cues can be ambiguous as they may signal more than one relation or may not always function as a relation indicator. In this study, we first conduct a corpus-based analysis to derive a set of n-grams as potential lexical cues. These cues are then utilized in graph-based probabilistic models to determine the syntactic context in which the cue is signaling the presence of a particular relation. Evaluation results are reported for various cues of the CIRCUMSTANCE relation, confirming the value of syntactic features for the task of cue disambiguation in the context of Rhetorical Structure Theory. Moreover, using a graph to encode syntactic information is shown to be a more generalizable and effective approach compared to the direct usage of syntactic features.

1 Introduction

A semantically sound text consists of discourse units that are connected through discourse relations, which are also referred to as rhetorical relations. Despite the efforts to build robust theoretical foundations and taxonomies for such relations (Hobbs, 1990; Knott and Sanders, 1998; Lascarides and Asher, 1993; Mann and Thompson, 1988), current methods for their automatic analysis and discovery in written discourse have yet to improve. However, providing robust models to analyze and identify rhetorical relations can benefit various research directions in computational linguistics such as text generation (Hovy, 1993) and summarization (Marcu, 2000), and machine translation (Meyer et al., 2011).

One of the widely accepted frameworks for discourse analysis and understanding is Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In RST, discourse structure has a form of a tree, where the leaves correspond to elementary discourse units, and the internal nodes correspond to contiguous text spans. Each internal node is marked with a rhetorical relation that holds between its child nodes. Figure 1 provides an example of an RST tree taken from the RST corpus (Carlson et al., 2001). One of the notable differences of RST with other similar theories is that it is structured on the intentions of the writers to use those relations (Taboada, 2006). This distinctive feature can make it even more difficult to build models for automatic identification of rhetorical relations in the context of RST.

Rhetorical relations can be either explicit or implicit. Explicit relations are the ones that are signaled by cues, such as lexical cues, mood, modality, and intonation (Taboada, 2006), while no cue is present in implicit relations. In this study, we are focused on explicit relations in written text that are signaled by the presence of lexical cues. Lexical cues are defined as linguistic expressions that function as explicit indicators of a discourse relation (Hirschberg and Litman, 1993). For example, in the sentence provided in Figure 1, *but* and *because* can be considered lexical cues signaling the existence of the CONCESSION relation and the EXPLANATION-ARGUMENTATIVE relation, respectively.

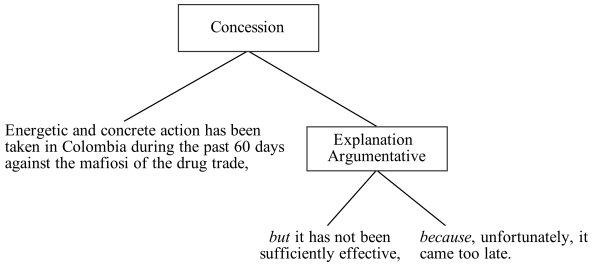

Figure 1: An example sentence parsed in the form of RST

Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 54–63,
Lisboa, Portugal, 18 September 2015.

Since this study is part of a larger project to identify rationales in written discourse, we focus on the three relations of CIRCUMSTANCE, EVALUATION, and ELABORATION that are commonly present in rationales (Xiao, 2013a). With the aim of proposing a cue-based approach to extract rhetorical relations, we have carried out some corpus-based experiments on RST annotated corpora. As a result of these experiments, we have generated a list of key n-grams as potential lexical cues for each relation. Such a corpus-based method may result in the discovery of underexplored lexical cues.

Even though lexical cues can be exploited to label rhetorical relations, they are not always unambiguous (Pitler and Nenkova, 2009). Some linguistic expressions may or may not function as a lexical cue, or they may signal different types of relations in different sentences. Hence, here, we propose a graph-based probabilistic model that takes into account the syntactic features of sentences. These models are intended to determine in what syntactic context a lexical cue is indeed signaling the presence of a particular relation. The models are then applied and tested on two corpora that belong to different text genres: news articles and online reviews.

The evaluation results of the approach are presented and discussed for the CIRCUMSTANCE relation. The CIRCUMSTANCE relation exists when a context of time or situation is presented, wherein the main events and ideas provided in the sentence can be interpreted in. CIRCUMSTANCE is chosen as the relation of focus since (Khazaei and Xiao, 2015) revealed that the cue-based approaches can be well-suited for the detection of CIRCUMSTANCE across different genres, while the ELABORATION relation is not normally signaled. In addition, the features of the underlying text genre can significantly influence how EVALUATION is signaled (Khazaei and Xiao, 2015).

The remainder of this paper is organized as follows: An overview of the previous research on lexical cue disambiguation is provided in Section 2. In Section 3, an explanation of the underlying corpora and the methods used to extract and disambiguate the cues is provided. The evaluation results are presented in Section 4. A discussion of the findings is given in Section 5, followed by a conclusion of the study in Section 6.

2 Related Work

The majority of studies focusing on discourse parsing and discourse relation classification report results achieved from both explicit and implicit relations (Soricut and Marcu, 2003; Wellner et al., 2006; Versley, 2013). Among the works that are particularly focused on lexical cue disambiguation, a large proportion are conducted on the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), while fewer studies have been conducted to study other discourse theories and frameworks.

PDTB annotation is lexically-grounded, and it is theory-neutral with respect to higher-level discourse structure (Rashmi et al., 2014). In the course of the annotation, the annotators were asked to seek lexical items that can signal relations and then annotate their corresponding arguments and relations (Rashmi et al., 2014). Even for implicit relations, annotators were asked to look for adjacent sentences that lacked one of these signals. When a relation could be inferred, they were asked to first label the relation with a lexical item that could serve as a signal and then annotate the relation sense. Such a lexically oriented approach to annotate relations has motivated a lot of work on disambiguation of lexical cues in PDTB.

For example, Miltsakaki et al. (Miltsakaki et al., 2005) have utilized a set of syntactic features along with a supervised model to disambiguate three discourse cues of *while*, *since*, and *when*. Their feature set includes form of the auxiliary *have*, form of the auxiliary *be*, form of the head, and presence of a modal. They obtained an accuracy of 75.5% to classify *since*, 71.8% for *while*, and 61.6% for *when*.

Pitler and Nenkova (Pitler and Nenkova, 2009) used a set of syntactic features to disambiguate cues regarding their discourse and non-discourse usage and sense disambiguation. Their features consist of the syntactic category of the marker, its parent, and its siblings. Two binary features are also taken into account to indicate whether the right sibling contains a VP and/or a trace. Their best feature set also included pairwise interaction features between the cues and syntactic features, and between the syntactic features themselves. Their learning algorithm resulted in an F-score of 92.28% for discourse versus non-discourse usage and an accuracy of 94.15% for sense classification. These results were later improved in (Ibn Faiz and Mercer, 2013), where a

set of surface-level and syntactic features are introduced and are combined with the feature set presented in (Pitler and Nenkova, 2009). The results of a classifier trained on this feature set resulted in an F-score of 96.22%.

Within a broader context of building an end-to-end discourse parser for PDTB, Lin et al. (Lin et al., 2014) built a cue classifier to identify whether a lexical item functions as a discourse cue or not. In addition to the features used in (Pitler and Nenkova, 2009), they also included part-of-speech features as well as features related to the syntactic parse path from the cue to the root of the tree. Using their set of lexico-syntactic and path features resulted in an F-score of 95.36%.

Meyer et al. (Meyer et al., 2011) used their own annotation schema developed based on parallel corpora and translation spotting to annotate three cues, two in English and one in French. Their annotation roughly follows a PDTB-like annotation and is likewise lexically-grounded. The annotated corpora was then used to train a learning model based on a set of features deemed valuable, including POS-tagged and parsed sentences, to disambiguate the lexical cues. As their best result, they achieved an accuracy of 85.7%.

Even though RST is one of the most widely accepted frameworks for discourse analysis, relatively little attention has been paid to RST annotated corpora in regards to lexical cue analysis and disambiguation. Unlike PDTB, annotations following RST are not lexically-grounded, and every relation is defined in terms of intentions that lead authors to use those specific relations (Taboada, 2006). Therefore, an RST diagram represents some of the authors' purposes or intentions for including each part of the text (Taboada, 2006). Such attributes of RST annotations make it a challenging task to study the role of lexical items in relation classification and to disambiguate them.

Marcu (Marcu, 2000) attempted to create a rhetorical parsing algorithm. A corpus study was conducted to understand how cues can be used to identify elementary discourse units and hypothesize their corresponding relation. By utilizing prior studies on discourse analysis, he created a list of 450 discourse cues to start with. An average of 17 text spans associated with each cue was then collected from the Brown corpus. All of the sentences were then annotated with two sets of metadata: discourse-related information and algorithmic features. Using these annotations, which mostly capture the orthographic environments of the cues, a set of regular expressions was created manually to recognize potential cues. If a cue had different discourse functions in different orthographic environments, a separate regular expression was made for each case. The algorithm resulted in an 84.9% F-score for the sub-task of cue identification. For the sub-task of relation classification, they achieved an F-score of 58.76%.

HILDA (Hernault et al., 2010) is a discourse parser developed to automatically construct the RST tree by performing the two core tasks of text segmentation and relation labeling. The relation labeling model takes into account the textual organization, structural organization, and lexical information of text as the underlying feature set. The performance results of a supervised model built on this feature set varies widely across different relations, ranging from 95% to 3.9% in F-score. The results are only reported for a subset of relations, within which the CIRCUMSTANCE relation is not present. On average, they achieved an F-score of 47.7% for labeling rhetorical relations.

In (da Cunha, 2013), a set of cues is first extracted from the database of Spanish discourse cues. The context of each cue is then extracted from the RST Spanish Treebank and is given to a syntactic parser. The syntactic features of the context of each cue are then manually analyzed to identify potential linguistic regularities and patterns. By using the results of the analysis, linguistic rules are developed to disambiguate the lexical cues. Their rules achieved an accuracy of 60.65%.

More recent studies on relation labeling in the context of RST have improved these results. For example, (Joty et al., 2013) has obtained an F-score of roughly 55% for their relation detection task. They made use of various organization features, textual features, lexio-syntactic features, lexical chains, as well as a lexical n-gram dictionary. These results are slightly improved by Ji and Eisenstein (Ji and Eisenstein, 2014), as they achieved an F-score of roughly 61% to detect rhetorical relations. They proposed a feature representation learning method in a shift-reduce discourse parser.

Many of the prior works on RST relation annotation are semi-automated and include manual steps. The few approaches that provide fully-automated cue-based techniques (Ji and Eisen-

stein, 2014; Joty et al., 2013; Hernault et al., 2010) have focused both their training and test process on a similar text genre. Even when focused on a single genre, to the best of our knowledge, the previous state-of-the art in relation labeling have resulted in an F-score of 61.75% (Ji and Eisenstein, 2014). Our work is intended to provide an automated approach to detect potential lexical cues that can indicate rhetorical relations, and to analyze whether their syntactic context can be of value for cue disambiguation across two different text genres.

3 Approach

In this section, we first describe the two RST annotated corpora that are used in the present work: RST corpus (Carlson et al., 2001) and Simon Fraser University (SFU) review dataset (Taboada et al., 2006). Then, an explanation of the approach used to extract a set of key n-grams as potential lexical cues is presented, which is followed by a description of our graph-based approach to disambiguate lexical cues.

3.1 Corpora

We used two human-annotated corpora as our underlying datasets for the experiments: the RST corpus (Carlson et al., 2001) and the SFU review dataset (Taboada et al., 2006). Both corpora are annotated in the RST framework and are constructed using the RSTTool[1].

The RST corpus, which has been made available by the Linguistic Data Consortium over the years, includes 385 *Wall Street Journal* articles and covers more than 178,000 words. Among the relation instances in the RST corpus, there exist around 700 instances of CIRCUMSTANCE, which constitutes almost 3% of the total number of relation instances.

The SFU review corpus is a collection of 400 review documents from movie, book, and consumer products. This dataset contains over 303,000 words and was collected in 2004 from the Epinions Web site[2]. There exist around 1300 CIRCUMSTANCE instances, constituting almost 7% of the annotated instances in the corpus.

[1] http://www.wagsoft.com/RSTTool

[2] http://www.epinions.com/

3.2 Lexical Cue Selection

The news text has a well-structured formal writing style, whereas the online reviews are relatively less structured and informal, written by users with a wide range of writing abilities. Therefore, to extract lexical cues associated with a given relation, we used the RST corpus.

First, all the relation instances are extracted from the RST corpus and are collected in a relation document named after the corresponding relation. Then, following the approach proposed in (Biran and Rambow, 2011), all the n-grams (up to trigrams) are extracted from the composed relation document. For each n-gram, an altered version of TF-IDF metric is then calculated. The IDF measure is still calculated based on the number of documents that contain the n-gram and the total number of documents in the corpus. However, since each line corresponds to one instance of the relation, the TF metric is calculated based on the number of lines that contain at least one instance of the n-gram. This altered metric allows us to offset the potential bias that may be caused by the TF metric for the words appearing more than once in a relation instance.

The list of the extracted n-grams (i.e., lexical cues) is then filtered to only include the n-grams with their TF-IDF above 0.5. To filter any corpus-specific n-grams that may appear in the list, the n-grams extracted from the RST corpus are applied to the SFU review dataset to identify the corresponding relation. The F-score of each n-gram is then calculated independently. Finally, the n-grams with an F-score of above 0.1 are selected as potential lexical cues. The aforementioned procedure resulted in the selection of seven lexical cues for the CIRCUMSTANCE relation: *When*, *after*, *on*, *before*, *with*, *out*, *as*.

3.3 Lexical Cue Disambiguation

Our cue disambiguation approach is mainly inspired by the work of (Hassan et al., 2010) on the detection of sentences with attitudes. In their study, the text fragment that includes a second pronoun is first extracted as the most relevant part of a sentence. These fragments are then represented using different patterns, capturing their syntactic features and semantic orientation. For every kind of pattern, graph models are built based on sentences with and without attitude. Finally, the likelihood of a new sentence being generated from

these models is used to predict the existence of an attitude. We adopted their approach for lexical cue disambiguation. Our graph models are built on the RST corpus and evaluated on the SFU review corpus and vice versa. Therefore, the graph building procedure explained below is conducted on both underlying corpora.

3.3.1 Data Collection

For every extracted cue, we first create two corresponding documents from the annotated corpora. One document consists of all of the relation instances that contain the cue and are annotated with the relation of focus (e.g., all of the CIRCUMSTANCE instances that are signaled by *when*). From now on, such instances will be referred to as positive instances. The other document consists of all the relation instances that contain the cue and are annotated as any relation except for the relation of focus (e.g., all of the non-circumstance instances that contain *when*). In the rest of this manuscript, we will refer to these instances as negative instances.

RST postulates a hierarchal structure on text, where a relation instance can be embedded in other instances. Therefore, during the extraction of the instances, we ensured not to collect negative instances that include any positive or negative sub-instance. We also ensured not to collect any positive instances that include negative sub-instances. The inclusion of such embedded instances would have resulted in redundant and incorrect data points. For example, consider the following positive instance from the RST corpus:

[*When* Mr. Gandhi came to power,]

[he ushered in new rules for business]$_{\text{circumstance}}$

When collecting negative instances, it was revealed that this instance was embedded in ten negative instances. However, since *when* is in fact functioning as a circumstance cue in all of them, those ten instances could not qualify as negative instances and so were excluded.

3.3.2 Syntactic Representations

After creation of the documents, each instance is processed and transformed into two different representations, capturing the syntactic features of the instance. To create the first syntactic representation, words in instances are replaced with their corresponding Part-Of-Speech (POS) tags, while the cue itself is kept as is. The second representation includes the shortest path from the root element to the cue in the dependency parse tree. The following is an example of the CIRCUMSTANCE relation, along with its two corresponding syntactic representations:

- Positive instance with *when* as the cue:
  When Mr. Gandhi came to power, he ushered in new rules for business
- POS-based representation:
  When NNP NNP VBD TO NN PRP VBD IN JJ NNS IN NN
- Shortest path representation:
  root advmod

We used the OpenNLP[3] toolkit to tokenize and POS tag the instances and the Stanford dependency parser to generate the parse trees (Klein and Manning, 2003).

3.3.3 Graph Modeling

We encoded the syntactic information of the instances in graph models. We build the directed weighted graph $G = (V, E), w$, where:

- V is the set of all possible tokens that may appear in the representations. For example, for the POS representations, V is the union of the set of all POS tags and the cue set.
- $E = V \times V$ is the set of all possible ordered transitions between any two tokens.
- $w \rightarrow [0 - 1]$ is a weighting function that assigns a probability value to an edge (i, j), which represents the probability of a transition from token i to token j.

Given a set of syntactic representations, the probability of a transition from token i to token j is calculated following a maximum likelihood estimation. Thus, the probability is calculated by dividing the number of times that token i is immediately followed by token j by the number of times that token i itself appears in the set.

This method of building the graphs is similar to language modeling but is conducted over a set of syntactic representations (Hassan et al., 2010). For every kind of representation, we build one graph based on the set of positive instances, and one based on the set of negative instances. As a result, given a cue (e.g., *when*) and its corresponding relation (e.g., CIRCUMSTANCE), we build four graph

[3] https://opennlp.apache.org/

models based on the following sets: POS representations of positive instances, POS representations of negative instances, dependency parsed representations of positive instances, and dependency parsed representations of negative sentences.

3.3.4 Cue Disambiguation Model

Finally, for our final cue disambiguation model, we utilize the probability values obtained from our graph models as the feature set for a standard machine-learning model. Given an instance and a graph, we calculate the likelihood of its syntactic representations to be generated from the corresponding syntactic graphs. The probability of a syntactic representation R that consists of a sequence of tokens $T_1, T_2, ..., T_n$ being generated from graph G is estimated using the following formula. Note that W is the weighting or probability transition function.

$$P_G(R) = \prod_{i=2}^{n} P(T_i|T_1, ..., T_{i-1}) = \prod_{i=2}^{n} W(T_{i-1}, T_i)$$

Given that we have four graph models, we can generate four probability values as our feature set. These features are further used in a standard supervised learning algorithm to disambiguate the cue and to classify the relation of a given instance. Figure 2 provides a high-level description of the entire process of cue extraction and disambiguation.

4 Evaluation

Given that our ultimate goal is to detect rationales from written discourse, our approach is evaluated for the CIRCUMSTANCE relation as it is the only cue-based relation that is known to be frequently present in rationales (Khazaei and Xiao, 2015; Xiao, 2013a). We carried out experiments using different forms of POS representations based on the number of POS tags surrounding the cue and the granularity of the tags. We conducted experiments using the entire POS tagged instance, using two POS tags before and two tags after the cue, and using one before and one after the cue. We also used three levels of POS tag granularity, including the finest, that is, the Penn English Treebank[4] tagset used by OpenNLP. We also used a medium and a coarse granularity that are created by gathering together similar tags into one high-level tag. Table 1 shows the tags that are grouped in each of these two granularity levels. Note that the tags not mentioned in the granularity levels are used as is.

Label	Medium Granularity	Coarse Granularity
JJ	JJ, JJR, JJS	JJ, JJR, JJS, DT, WDT
NN	NN, NNS, NNP, NNPS	NN, NNS, NNP, NNPS
PRP	PRP, PRP$	PRP, PRP$, WP, WP$
RB	RB, RBR, RBS	RB, RBR, RBS, WRB
WP	WP, WP$	-
VB	VB, VBD, VBN, VBP, VBZ	VB, VBD, VBN, VBP, VBZ, MD, VBG

Table 1: In addition to the POS tags in the Penn English Treebank tag set, experiments are conducted using tags grouped according to different levels of granularity.

Using these three variations of the two POS tag attributes resulted in nine different experimental settings. We achieved our best results on both corpora using one tag before and one tag after the cue and the medium granularity level. In this section, we report results for experiments using this particular POS setting.

The final algorithm built on probability values is evaluated using the Weka workbench[5]. It classifies instances via regression[6], and a stratified ten-fold cross validation is followed to evaluate the model. To gain insight into the effectiveness of the model in the disambiguation of different cues, results are reported for each of the seven cues independently. The SMOTE filter was used when significant class imbalance was encountered.

Table 2 demonstrates the results when the RST corpus was used to build the graphs, and the SFU corpus was used to build and test the final model. Table 3 shows the results of the evaluation, where graphs are built on the SFU corpus and used on the RST dataset. As can be seen, the measures of precision, recall, and F-score are reported, along with their average value. The weighted average of F-score is also provided, taking into account the distribution of relation instances that contain the cues in the test set. This metric is provided while bearing in mind that the test set may not be an accurate representative of the general distribution of relations. According to the results, on average, we were able to classify CIRCUMSTANCE with an F-

[4]http://www.cis.upenn.edu/ treebank/

[5]http://www.cs.waikato.ac.nz/ml/weka/

[6]ClassificationViaRegression algorithm is used with default parameters

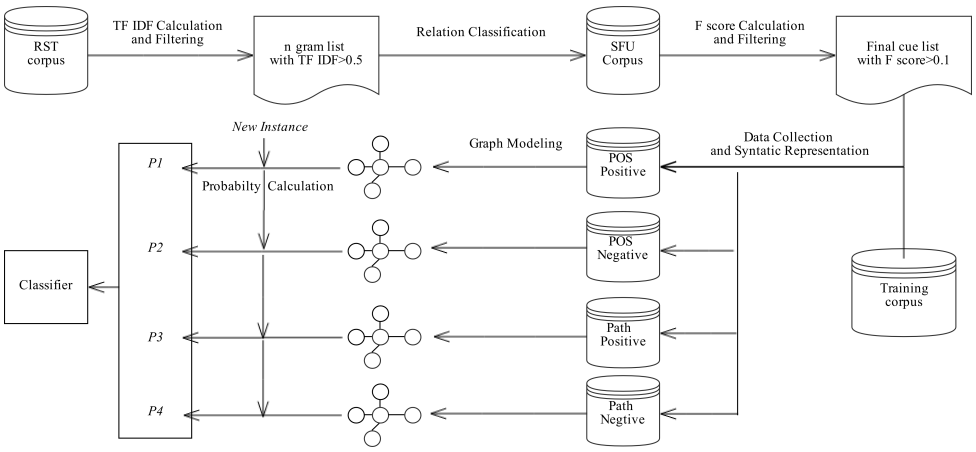

Figure 2: A high-level overview of the cue extraction and disambiguation approach

Cue	Precision	Recall	F-score
When	0.55	0.79	0.65
After	0.46	0.56	0.51
On	0.73	0.75	0.74
Before	0.62	0.60	0.61
With	0.76	0.80	0.78
Out	0.60	0.54	0.57
As	0.71	0.82	0.76
Average	0.63	0.69	0.66
Weighted Average			0.71

Table 2: Classification results of a ten-fold cross validation on the SFU corpus. Probability values used as the underlying feature set are inferred from the graph models built on the RST corpus.

Cue	Precision	Recall	F-score
When	0.62	0.69	0.65
After	0.61	0.57	0.59
On	0.77	0.81	0.79
Before	0.70	0.40	0.51
With	0.77	0.81	0.79
Out	0.75	0.70	0.73
As	0.73	0.67	0.70
Average	0.71	0.67	0.68
Weighted Average			0.69

Table 3: Classification results of a ten-fold cross validation on the RST dataset. Probability values used as the underlying feature set are inferred from the graph models built on the SFU corpus.

score of 0.66% in the SFU review dataset, while the weighted average of F-score is 0.71%. In addition, an average F-score of 0.68% and a weighted average of 0.69% are achieved for the RST corpus.

5 Discussion

The use of syntactic context to disambiguate lexical cues has been shown to be useful to disambiguate cues in lexically oriented relations (e.g., PDTB relations). In this study, we have focused our efforts on RST annotated corpora and explored the potential of syntactic context for cue disambiguation in the RST framework. We have demonstrated that syntactic features can be of great value in the classification of explicit rhetorical relations. In addition, unlike the majority of prior studies on cue disambiguation, we encoded the syntactic context of cues in the form of graphs. This graph-based approach was expected to provide a more generalizable and effective approach.

Earlier studies on the detection of relations in the context of RST are focused on a single text genre for their training phase as well as their test process; hence, their results are not directly comparable with our approach. In addition, they have not reported results for the CIRCUMSTANCE relation separately. Even though our average results for the detection of CIRCUMSTANCE (66% for the SFU corpus and 68% for the RST corpus) are higher than the state-of-the-art (61.75% (Ji and Eisenstein, 2014)), further experiments are required with other RST relations and different genres to make a sound comparison with such earlier works. To highlight the contribution of our work, we conducted experiments to compare the graph-based model with the direct usage of syntactic features. A logistic model was first trained on the RST corpus and tested on the SFU dataset (see Table 4), and then trained on the SFU corpus and tested on the RST corpus (see Table 5). The results are consistently lower for all of the three measures, confirming the superiority of our approach.

Based on the results of our proposed approach, it can be seen that the three lexical cues of *when*, *after*, and *before* have the lowest performance in the RST corpus (see Table 3). They are also

Cue	Precision	Recall	F-score
When	0.50	0.41	0.45
After	0.39	0.15	0.22
On	0.65	0.28	0.39
Before	0.52	0.49	0.51
With	0.53	0.34	0.41
Out	0.51	0.19	0.28
As	0.49	0.24	0.32
Average	0.51	0.30	0.37

Table 4: Classification results on the SFU corpus when the syntactic features are used directly to train a model on the RST corpus

Cue	Precision	Recall	F-score
When	0.53	0.65	0.58
After	0.31	0.14	0.20
On	0.36	0.17	0.23
Before	0.33	0.22	0.26
With	0.62	0.20	0.30
Out	0.57	0.53	0.55
As	0.52	0.25	0.34
Average	0.52	0.25	0.35

Table 5: Classification results on the RST corpus when the syntactic features are used directly to train a model on the SFU corpus

among the four cues with lowest F-score in the SFU dataset (see Table 2). This finding could be attributed to the fact that, for these three cues, the corresponding datasets were among the smallest cue sets. Possibly more importantly, these three cues can function as temporal indicators, which may make it particularly difficult to disambiguate them. For example, consider the following instances extracted from the SFU corpus:

- Positive instance:
  When I have time to kill between flights, I like to wander through and browse

- Negative instance:
  I was surprised *when* he told me that all the equipment was standard even on the base model

The first sentence is an instance of the CIRCUMSTANCE relation signaled by *when*, while in the second one, *when* implies the temporal aspect of the sentence and is not signaling CIRCUMSTANCE. We expect that certain linguistic and contextual features associated with the text, such as verb tense, might be useful in the disambiguation of such lexical cues. Further studies are required to explore these features.

Since RST places an emphasis on the writer's intentions and the effect of the relation on the reader (Taboada, 2006), RST annotations are inherently subjective and are based on the readers' understanding of the text (Taboada, 2006). Hence, there can be differences across the two corpora due to the different knowledge possessed by each set of annotators (Taboada, 2006). Despite the genre disparity and annotation issues, we obtained encouraging results using the proposed model. However, the results are expected to improve when the models are built on corpora from similar genres and are annotated using ground truth rules.

6 Conclusion

The study and analysis of rhetorical relations, as the building blocks of coherence in discourse, can contribute toward the development of sophisticated applications and algorithms. With the aim of facilitating automatic discovery of explicit rhetorical relations in text, we developed an algorithm to first detect potential lexical cues and to later disambiguate them by predicting the relation.

An altered version of TF-IDF was used to extract the cues, and a graph-based model built on syntactic features was used to address the cue disambiguation task. Overall, the evaluation results indicate the effectiveness of syntactic features in the disambiguation of cues and prediction of explicit rhetorical relations across different genres. Our experiments revealed the superiority of encoding such syntactic features in a probabilistic graph compared to their direct usage.

This study is our first attempt toward the identification of rationales in text. A rationale is an explanation of the reasons underlying decisions, conclusions, and interpretations. Prior studies on rationale articulation and sharing suggest that it contributes to quality control, knowledge management, and knowledge reuse (Xiao, 2014; Xiao, 2013b). However, there exists only a few automated methods to identify rationales from ill-structured text (Ghosh et al., 2014; Boltužić and Šnajder, 2014). Our future research efforts are focused on the development of algorithms to extract lightly-signaled and implicit relations and to further explore the potential and limitations of using rhetorical relations in the detection of rationales.

Acknowledgments

This project is funded through the Discovery program of Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the IEEE International Conference on Semantic Computing*, pages 162–168.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 1–10.

Iria da Cunha. 2013. A symbolic corpus-based approach to detect and solve the ambiguity of discourse markers. *Research in Computing Science*, 70:93–104.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.

Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. What's with the attitude? Identifying sentences with attitude in online discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1245–1255.

Hugo Hernault, Helmut Prendinger, David Duverle, Mitsuru Ishizuka, and Tim Paek. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 3(1):1–33.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

J.R. Hobbs. 1990. *Literature and Cognition*. Center for the Study of Language and Information - Lecture Notes. Cambridge University Press.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63(1-2):341–385.

Syeed Ibn Faiz and Robert Mercer. 2013. Identifying explicit discourse connectives in text. In *Advances in Artificial Intelligence*, pages 64–76.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13–24.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 486–496.

Taraneh Khazaei and Lu Xiao. 2015. Corpus-based analysis of rhetorical relations: A study of lexical cues. In *Proceedings of the IEEE Conference on Semantic Computing*, pages 417–423.

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

Alistair Knott and Ted Sanders. 1998. The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics*, 30(2):135 – 175.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *SIGdial Meeting on Discourse and Dialogue*, pages 194–203.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference (Short Papers)*, pages 13–16.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 2961–2968.

Prasad Rashmi, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 149–156.

Maite Taboada, Caroline Anthony, and Kimberly Voll. 2006. Methods for creating semantic orientation dictionaries. In *Proceedings of the Conference on Language Resources and Evaluation*, pages 427–432.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567 – 592.

Yannick Versley. 2013. Subgraph-based classification of explicit and implicit discourse relations. In *Proceedings of the International Conference on Computational Semantics*, pages 264–275.

Ben Wellner, James Pustejovsky, Catherine Havasi, Anna Rumshisky, and Roser Saurí. 2006. Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 117–125.

Lu Xiao. 2013a. Do members converge to similar reasoning styles in teamwork? A study of shared rationales in small team activities. In *Proceedings of the iConference*, pages 524–530.

Lu Xiao. 2013b. The effects of a shared free form rationale space in collaborative learning activities. *Journal of Systems and Software*, 86(7):1727 – 1737.

Lu Xiao. 2014. Effects of rationale awareness in online ideation crowdsourcing tasks. *Journal of the Association for Information Science and Technology*, 65(8):1707–1720.