Classification of Discourse Coherence Relations: An Exploratory Study using Multiple Knowledge Sources

Ben Wellner[†*], **James Pustejovsky**[†], **Catherine Havasi**[†], **Anna Rumshisky**[†] and **Roser Saurí**[†]
[†]Department of Computer Science
Brandeis University
Waltham, MA USA
[*]The MITRE Corporation
202 Burlington Road
Bedford, MA USA
{wellner,jamesp,havasi,arum,roser}@cs.brandeis.edu

Abstract

In this paper we consider the problem of identifying and classifying discourse coherence relations. We report initial results over the recently released Discourse GraphBank (Wolf and Gibson, 2005). Our approach considers, and determines the contributions of, a variety of syntactic and lexico-semantic features. We achieve 81% accuracy on the task of discourse relation type classification and 70% accuracy on relation identification.

1 Introduction

The area of modeling discourse has arguably seen less success than other areas in NLP. Contributing to this is the fact that no consensus has been reached on the inventory of discourse relations nor on the types of formal restrictions placed on discourse structure. Furthermore, modeling discourse structure requires access to considerable prior linguistic analysis including syntax, lexical and compositional semantics, as well as the resolution of entity and event-level anaphora, all of which are non-trivial problems themselves.

Discourse processing has been used in many text processing applications, most notably text summarization and compression, text generation, and dialogue understanding. However, it is also important for general text understanding, including applications such as information extraction and question answering.

Recently, Wolf and Gibson (2005) have proposed a graph-based approach to representing informational discourse relations.[1] They demonstrate that tree representations are inadequate for modeling coherence relations, and show that many discourse segments have multiple parents (incoming directed relations) and many of the relations introduce crossing dependencies – both of which preclude tree representations. Their annotation of 135 articles has been released as the GraphBank corpus.

In this paper, we provide initial results for the following tasks: (1) automatically classifying the *type* of discourse coherence relation; and (2) identifying whether any discourse relation *exists* on two text segments. The experiments we report are based on the annotated data in the Discourse GraphBank, where we assume that the discourse units have already been identified.

In contrast to a highly structured, compositional approach to discourse parsing, we explore a simple, flat, feature-based methodology. Such an approach has the advantage of easily accommodating many knowledge sources. This type of detailed feature analysis can serve to inform or augment more structured, compositional approaches to discourse such as those based on Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) or the approach taken with the D-LTAG system (Forbes et al., 2001).

Using a comprehensive set of linguistic features as input to a Maximum Entropy classifier, we achieve 81% accuracy on classifying the correct type of discourse coherence relation between two segments.

2 Previous Work

In the past few years, the tasks of discourse segmentation and parsing have been tackled from different perspectives and within different frameworks. Within Rhetorical Structure Theory (RST), Soricut and Marcu (2003) have developed two

[1] The relations they define roughly follow Hobbs (1985).

Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pages 117–125,
Sydney, July 2006.

probabilistic models for identifying clausal elementary discourse units and generating discourse trees at the sentence level. These are built using lexical and syntactic information obtained from mapping the discourse-annotated sentences in the RST Corpus (Carlson et al., 2003) to their corresponding syntactic trees in the Penn Treebank.

Within SDRT, Baldridge and Lascarides (2005b) also take a data-driven approach to the tasks of segmentation and identification of discourse relations. They create a probabilistic discourse parser based on dialogues from the Redwoods Treebank, annotated with SDRT rhetorical relations (Baldridge and Lascarides, 2005a). The parser is grounded on headed tree representations and dialogue-based features, such as turn-taking and domain specific goals.

In the Penn Discourse TreeBank (PDTB) (Webber et al., 2005), the identification of discourse structure is approached independently of any linguistic theory by using discourse connectives rather than abstract rhetorical relations. PDTB assumes that connectives are binary discourse-level predicates conveying a semantic relationship between two abstract object-denoting arguments. The set of semantic relationships can be established at different levels of granularity, depending on the application. Miltsakaki, et al. (2005) propose a first step at disambiguating the sense of a small subset of connectives (*since, while*, and *when*) at the paragraph level. They aim at distinguishing between the temporal, causal, and contrastive use of the connective, by means of syntactic features derived from the Penn Treebank and a MaxEnt model.

3 GraphBank

3.1 Coherence Relations

For annotating the discourse relations in text, Wolf and Gibson (2005) assume a clause-unit-based definition of a discourse segment. They define four broad classes of coherence relations:

(1) 1. Resemblance: similarity (par), contrast (contr), example (examp), generalization (gen), elaboration (elab);
2. Cause-effect: explanation (ce), violated expectation (expv), condition (cond);
3. Temporal (temp): essentially narration;
4. Attribution (attr): reporting and evidential contexts.

The textual evidence contributing to identifying the various resemblance relations is heterogeneous at best, where, for example, *similarity* and *contrast* are associated with specific syntactic constructions and devices. For each relation type, there are well-known lexical and phrasal cues:

(2) a. *similarity*: and;
b. *contrast*: by contrast, but;
c. *example*: for example;
d. *elaboration*: also, furthermore, in addition, note that;
e. *generalization*: in general.

However, just as often, the relation is encoded through lexical coherence, via semantic association, sub/supertyping, and accommodation strategies (Asher and Lascarides, 2003).

The cause-effect relations include conventional *causation* and *explanation* relations (captured as the label *ce*), such as (3) below:

(3) **cause:** SEG1: crash-landed in New Hope, Ga.,
effect: SEG2: and injuring 23 others.

It also includes *conditionals* and *violated expectations*, such as (4).

(4) **cause:** SEG1: an Eastern Airlines Lockheed L-1011 en route from Miami to the Bahamas lost all three of its engines,
effect: SEG2: and land safely back in Miami.

The two last coherence relations annotated in GraphBank are *temporal (temp)* and *attribution (attr)* relations. The first corresponds generally to the *occasion* (Hobbs, 1985) or *narration* (Asher and Lascarides, 2003) relation, while the latter is a general annotation over attribution of source.[2]

3.2 Discussion

The difficulty of annotating coherence relations consistently has been previously discussed in the literature. In GraphBank, as in any corpus, there are inconsistencies that must be accommodated for learning purposes. As perhaps expected, annotation of attribution and temporal sequence relations was consistent if not entirely complete. The most serious concern we had from working with

[2]There is one non-rhetorical relation, *same*, which identifies discontiguous segments.

the corpus derives from the conflation of diverse and semantically contradictory relations among the *cause-effect* annotations. For canonical causation pairs (and their violations) such as those above, (3) and (4), the annotation was expectedly consistent and semantically appropriate. Problems arise, however when examining the treatment of purpose clauses and rationale clauses. These are annotated, according to the guidelines, as cause-effect pairings. Consider (5) below.

(5) **cause:** SEG1: to upgrade lab equipment in 1987.
effect: SEG2: The university spent $ 30,000

This is both counter-intuitive and temporally false. The rationale clause is annotated as the cause, and the matrix sentence as the effect. Things are even worse with purpose clause annotation. Consider the following example discourse:[3]

(6) John pushed the door to open it, but it was locked.

This would have the following annotation in GraphBank:

(7) **cause:** to open it
effect: John pushed the door.

The guideline reflects the appropriate intuition that the intention expressed in the purpose or rationale clause must precede the implementation of the action carried out in the matrix sentence. In effect, this would be something like

(8) [INTENTION TO SEG1] CAUSES SEG2

The problem here is that the cause-effect relation conflates real event-causation with *telos*-directed explanations, that is, action directed towards a goal by virtue of an intention. Given that these are semantically disjoint relations, which are furthermore triggered by distinct grammatical constructions, we believe this conflation should be undone and characterized as two separate coherence relations. If the relations just discussed were annotated as *telic*-causation, the features encoded for subsequent training of a machine learning algorithm could benefit from distinct syntactic environments. We would like to automatically generate temporal orderings from cause-effect relations from the events directly annotated in the text. Splitting these classes would preserve the soundness of such a procedure, while keeping them lumped generates inconsistencies.

[3]This specific example was brought to our attention by Alex Lascarides (p.c).

4 Data Preparation and Knowledge Sources

In this section we describe the various linguistic processing components used for classification and identification of GraphBank discourse relations.

4.1 Pre-Processing

We performed tokenization, sentence tagging, part-of-speech tagging, and shallow syntactic parsing (chunking) over the 135 GraphBank documents. Part-of-speech tagging and shallow parsing were carried out using the Carafe implementation of Conditional Random Fields for NLP (Wellner and Vilain, 2006) trained on various standard corpora. In addition, full sentence parses were obtained using the RASP parser (Briscoe and Carroll, 2002). Grammatical relations derived from a single top-ranked tree for each sentence (headword, modifier, and relation type) were used for feature construction.

4.2 Modal Parsing and Temporal Ordering of Events

We performed both modal parsing and temporal parsing over *events.* Identification of events was performed using EvITA (Saurí et al., 2006), an open-domain event tagger developed under the TARSQI research framework (Verhagen et al., 2005). EvITA locates and tags all event-referring expressions in the input text that can be temporally ordered. In addition, it identifies those grammatical features implicated in temporal and modal information of events; namely, tense, aspect, polarity, modality, as well as the event class. Event annotation follows version 1.2.1 of the TimeML specifications.[4]

Modal parsing in the form of identifying subordinating verb relations and their type was performed using SlinkET (Saurí et al., 2006), another component of the TARSQI framework. SlinkET identifies subordination constructions introducing modality information in text; essentially, infinitival and *that*-clauses embedded by factive predicates (*regret*), reporting predicates (*say*), and predicates referring to events of attempting (*try*), volition (*want*), command (*order*), among others.

[4]See http://www.timeml.org.

SlinkET annotates these subordination contexts and classifies them according to the modality information introduced by the relation between the embedding and embedded predicates, which can be of any of the following types:

- **factive:** The embedded event is presupposed or entailed as true (e.g., *John managed to leave the party*).
- **counter-factive:** The embedded event is presupposed as entailed as false (e.g., *John was unable to leave the party*).
- **evidential:** The subordination is introduced by a reporting or perception event (e.g., *Mary saw/told that John left the party*).
- **negative evidential:** The subordination is a reporting event conveying negative polarity (e.g., *Mary denied that John left the party*).
- **modal:** The subordination creates an intensional context (e.g., *John wanted to leave the party*).

Temporal orderings between events were identified using a Maximum Entropy classifier trained on the TimeBank 1.2 and Opinion 1.0a corpora. These corpora provide annotated events along with temporal links between events. The link types included: *before* (e_1 occurs before e_2) , *includes* (e_2 occurs sometime during e_1), *simultaneous* (e_1 occurs over the same interval as e_2), *begins* (e_1 begins at the same time as e_2), *ends* (e_1 ends at the same time as e_2).

4.3 Lexical Semantic Typing and Coherence

Lexical semantic types as well as a measure of lexical similarity or coherence between words in two discourse segments would appear to be useful for assigning an appropriate discourse relationship. *Resemblance* relations, in particular, require similar entities to be involved and lexical similarity here serves as an approximation to definite nominal coreference. Identification of lexical relationships between words across segments appears especially useful for *cause-effect* relations. In example (3) above, determining a (potential) cause-effect relationship between *crash* and *injury* is necessary to identify the discourse relation.

4.3.1 Corpus-based Lexical Similarity

Lexical similarity was computed using the Word Sketch Engine (WSE) (Killgarrif et al., 2004) similarity metric applied over British National Corpus. The WSE similarity metric implements the word similarity measure based on grammatical relations as defined in (Lin, 1998) with minor modifications.

4.3.2 The Brandeis Semantic Ontology

As a second source of lexical coherence, we used the Brandeis Semantic Ontology or BSO (Pustejovsky et al., 2006). The BSO is a lexically-based ontology in the Generative Lexicon tradition (Pustejovsky, 2001; Pustejovsky, 1995). It focuses on contextualizing the meanings of words and does this by a rich system of types and qualia structures. For example, if one were to look up the phrase RED WINE in the BSO, one would find its type is WINE and its type's type is ALCOHOLIC BEVERAGE. The BSO contains ontological qualia information (shown below). Using the BSO, one

wine
CONSTITUTIVE = **Alcohol**
HAS ELEMENT = **Alcohol**
MADE OF = **Grapes**
INDIRECT TELIC = **drink activity**
INDIRECT AGENTIVE = **make alcoholic beverage**

is able to find out where in the ontological type system WINE is located, what RED WINE's lexical neighbors are, and its full set of part of speech and grammatical attributes. Other words have a different configuration of annotated attributes depending on the type of the word.

We used the BSO typing information to semantically tag individual words in order to compute lexical paths between word pairs. Such lexical associations are invoked when constructing cause-effect relations and other implicatures (e.g. between *crash* and *injure* in Example 3).

The type system paths provide a measure of the connectedness between words. For every pair of head words in a GraphBank document, the shortest path between the two words within the BSO is computed. Currently, this metric only uses the type system relations (i.e., inheritance) but preliminary tests show that including qualia relations as connections is promising. We also computed the earliest common ancestor of the two words. These metrics are calculated for every possible sense of the word within the BSO.

The use of the BSO is advantageous compared to other frameworks such as Wordnet because it focuses on the connection between words and their semantic relationship to other items. These connections are captured in the qualia information and the type system. In Wordnet, qualia-like information is only present in the glosses, and they do not provide a definite semantic path between any two lexical items. Although synonymous in some ways, synset members often behave differently in many situations, grammatical or otherwise.

5 Classification Methodology

This section describes in detail how we constructed features from the various knowledge sources described above and how they were encoded in a Maximum Entropy model.

5.1 Maximum Entropy Classification

For our experiments of classifying relation types, we used a Maximum Entropy classifier[5] in order to assign labels to each pair of discourse segments connected by some relation. For each instance (i.e. pair of segments) the classifier makes its decision based on a set of *features*. Each feature can query some arbitrary property of the two segments, possibly taking into account external information or knowledge sources. For example, a feature could query whether the two segments are adjacent to each other, whether one segment contains a discourse connective, whether they both share a particular word, whether a particular syntactic construction or lexical association is present, etc. We make strong use of this ability to include very many, highly interdependent features[6] in our experiments. Besides binary-valued features, feature values can be real-valued and thus capture frequencies, similarity values, or other scalar quantities.

5.2 Feature Classes

We grouped the features together into various *feature classes* based roughly on the knowledge source from which they were derived. Table 1 describes the various feature classes in detail and provides some actual example features from each class for the segment pair described in Example 5 in Section 3.2.

[5]We use the Maximum Entropy classifier included with Carafe available at http://sourceforge.net/projects/carafe

[6]The total maximum number of features occurring in our experiments is roughly 120,000.

6 Experiments and Results

In this section we provide the results of a set of experiments focused on the task of discourse relation classification. We also report initial results on relation *identification* with the same set of features as used for classification.

6.1 Discourse Relation Classification

The task of discourse relation classification involves assigning the correct label to a pair of discourse segments.[7] The pair of segments to assign a relation to is provided (from the annotated data). In addition, we assume, for asymmetric links, that the nucleus and satellite are provided (i.e., the *direction* of the relation). For the *elaboration* relations, we ignored the annotated subtypes (person, time, location, etc.). Experiments were carried out on the full set of relation types as well as the simpler set of coarse-grained relation categories described in Section 3.1.

The GraphBank contains a total of 8755 annotated coherence relations. [8] For all the experiments in this paper, we used 8-fold cross-validation with 12.5% of the data used for testing and the remainder used for training for each fold. Accuracy numbers reported are the average accuracies over the 8 folds. Variance was generally low with a standard deviation typically in the range of 1.5 to 2.0. We note here also that the inter-annotator agreement between the two GraphBank annotators was 94.6% for relations *when they agreed on the presence of a relation.* The majority class baseline (i.e., the accuracy achieved by calling all relations *elaboration*) is 45.7% (and 66.57% with the collapsed categories). These are the upper and lower bounds against which these results should be based.

To ascertain the utility of each of the various feature classes, we considered each feature class independently by using only features from a single class in addition to the Proximity feature class which serve as a baseline. Table 2 illustrates the result of this experiment.

We performed a second set of experiments shown in Table 3 that is essentially the converse of the previous batch. We take the union of all the

[7]Each segment may in fact consist of a sequence of segments. We will, however, use the term *segment* loosely to refer to segments or segment sequences.

[8]All documents are doubly annotated; we used the *annotator1* annotations.

Feature Class	Description	Example
C	Words appearing at beginning and end of the two discourse segments - these are often important discourse cue words.	first1-is-*to*; first2-is-*The*
P	Proximity and direction between the two segments (in terms of segments) - binary features such as *distance less than 3*, *distance greater than 10* were used in addition to the distance value itself; the distance from beginning of the document using a similar binning approach	adjacent; dist-less-than-3; dist-less-than-5; direction-reverse; samesentence
BSO	Paths in the BSO up to length 10 between non-function words in the two segments.	ResearchLab $\rightarrow$ EducationalActivity $\rightarrow$ University
WSE	WSE word-pair similarities between words in the two segments were binned as (> 0.05, > 0.1, > 0.2). We also computed sentence similarity as the sum of the word similarities divided by the sum of their sentence lengths.	WSE-greater-than-0.05; WSE-sentence-sim = 0.005417
E	Event head words and event head word pairs between segments as identified by EvITA.	event1-is-*upgrade*; event2-is-*spent*; event-pair-*upgrade-spent*
SlinkET	Event attributes, subordinating links and their types between event pairs in the two segments	seg1-class-is-*occurrence*; seg2-class-is-*occurrence*; seg1-tense-is-*infinitive*; seg2-tense-is-*past*; seg2-*modal*-seg1
C-E	Cuewords of one segment paired with events in the other.	first1-is-*to*-event2-is-*spent*; first2-is-*The*-event1-is-*upgrade*
Syntax	Grammatical dependency relations between two segments as identified by the RASP parser. We also conjoined the relation with one or both of the headwords associated with the grammatical relation.	gr-*ncmod*; gr-*ncmod*-head1-*equipment*; gr-*ncmod*-head-2-*spent*; etc.
Tlink	Temporal links between events in the two segments. We included both the link types and the number of occurrences of those types between the segments	seg2-*before*-seg1

Table 1: Feature classes, their descriptions and example feature instances for Example 5 in Section 3.2.

Feature Class	Accuracy	Coarse-grained Acc.
Proximity	60.08%	69.43%
P+C	76.77%	83.50%
P+BSO	62.92%	74.40%
P+WSE	62.20%	70.10%
P+E	63.84%	78.16%
P+SlinkET	69.00%	75.91%
P+CE	67.18%	78.63%
P+Syntax	70.30%	80.84%
P+Tlink	64.19%	72.30%

Table 2: Classification accuracy over standard and coarse-grained relation types with each feature class added to Proximity feature class.

feature classes and perform ablation experiments by removing one feature class at a time.

Feature Class	Accuracy	Coarse-grain Acc.
All Features	81.06%	87.51%
All-P	71.52%	84.88%
All-C	75.71%	84.69%
All-BSO	80.65%	87.04%
All-WSE	80.26%	87.14%
All-E	80.90%	86.92%
All-SlinkET	79.68%	86.89%
All-CE	80.41%	87.14%
All-Syntax	80.20%	86.89%
All-Tlink	80.30%	87.36%

Table 3: Classification accuracy with each feature class removed from the union of all feature classes.

6.2 Analysis

From the ablation results, it is clear that overall performance is most impacted by the *cue-word* features (C) and *proximity* (P). Syntax and SlinkET also have high impact improving accuracy by roughly 10 and 9 percent respectively as shown in Table 2. From the ablation results in Table 3, it is clear that the utility of most of the individual features classes is lessened when all the other feature classes are taken into account. This indicates that multiple feature classes are responsible for providing evidence any given discourse relations. Removing a single feature class degrades performance, but only slightly, as the others can compensate.

Overall precision, recall and F-measure results for each of the different link types using the set of all feature classes are shown in Table 4 with the corresponding confusion matrix in Table A.1. Performance correlates roughly with the frequency of the various relation types. We might therefore expect some improvement in performance with more annotated data for those relations with low frequency in the GraphBank.

Relation	Precision	Recall	F-measure	Count
elab	88.72	95.31	91.90	512
attr	91.14	95.10	93.09	184
par	71.89	83.33	77.19	132
same	87.09	75.00	80.60	72
ce	78.78	41.26	54.16	63
contr	65.51	66.67	66.08	57
examp	78.94	48.39	60.00	31
temp	50.00	20.83	29.41	24
expv	33.33	16.67	22.22	12
cond	45.45	62.50	52.63	8
gen	0.0	0.0	0.0	0

Table 4: Precision, Recall and F-measure results.

6.3 Coherence Relation Identification

The task of identifying the presence of a relation is complicated by the fact that we must consider all $\binom{n}{2}$ potential relations where n is the number of segments. This presents a troublesome, highly-skewed binary classification problem with a high proportion of negative instances. Furthermore, some of the relations, particularly the *resemblance* relations, are transitive in nature (e.g. $parallel(s_i, s_j) \wedge parallel(s_j, s_k) \rightarrow parallel(s_i, s_k)$). However, these transitive links are not provided in the GraphBank annotation - such segment pairs will therefore be presented incorrectly as negative instances to the learner, making this approach infeasible. An initial experiment considering all segment pairs, in fact, resulted in performance only slightly above the majority class baseline.

Instead, we consider the task of identifying the presence of discourse relations between segments within the same sentence. Using the same set of all features used for relation classification, performance is at 70.04% accuracy. Simultaneous identification and classification resulted in an accuracy of 64.53%. For both tasks the baseline accuracy was 58%.

6.4 Modeling Inter-relation Dependencies

Casting the problem as a standard classification problem where each instance is classified independently, as we have done, is a potential drawback. In order to gain insight into how collective, dependent modeling might help, we introduced additional features that model such dependencies: For a pair of discourse segments, s_i and s_j, to classify the relation between, we included features based on the *other* relations involved with the two segments (from the gold standard annotations): $\{R(s_i, s_k)|k \neq j\}$ and $\{R(s_j, s_l)|l \neq i\}$. Adding these features improved classification accuracy to 82.3%. This improvement is fairly significant (a 6.3% reduction in error) given that this dependency information is only encoded weakly as features and not in the form of model constraints.

7 Discussion and Future Work

We view the accuracy of 81% on coherence relation classification as a positive result, though room for improvement clearly remains. An examination of the errors indicates that many of the remaining problems require making complex lexical associations, the establishment of entity and event anaphoric links and, in some cases, the exploitation of complex world-knowledge. While important lexical connections can be gleaned from the BSO, we hypothesize that the current lack of word sense disambiguation serves to lessen its utility since lexical paths between *all* word sense of two words are currently used. Additional feature engineering, particularly the crafting of more specific *conjunctions* of existing features is another avenue to explore further - as are automatic feature selection methods.

Different types of relations clearly benefit from different feature types. For example, resemblance relations require similar entities and/or events, indicating a need for robust anaphora resolution, while cause-effect class relations require richer lexical and world knowledge. One promising approach is a pipeline where an initial classifier assigns a coarse-grained category, followed by separately engineered classifiers designed to model the finer-grained distinctions.

An important area of future work involves incorporating additional *structure* in two places. First, as the experiment discussed in Section 6.4 shows, classifying discourse relations collectively shows potential for improved performance. Secondly, we believe that the tasks of: 1) identifying which segments are related and 2) identifying the discourse segments themselves are probably best approached by a parsing model of discourse. This view is broadly sympathetic with the approach in (Miltsakaki et al., 2005).

We furthermore believe an extension to the GraphBank annotation scheme, with some minor changes as we advocate in Section 3.2, layered on top of the PDTB would, in our view, serve as an interesting resource and model for informational

discourse.

Acknowledgments

This work was supported in part by ARDA/DTO under grant number NBCHC040027 and MITRE Sponsored Research. Catherine Havasi is supported by NSF Fellowship # 2003014891.

References

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, England.

J. Baldridge and A. Lascarides. 2005a. Annotating discourse structures for robust semantic interpretation. In *Proceedings of the Sixth International Workshop on Computational Semantics*, Tilburg, The Netherlands.

J. Baldridge and A. Lascarides. 2005b. Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, Ann Arbor, USA.

T. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002*, pages 1499–1504.

L. Carlson, D. Marcu, and M. E. Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Janvan Kuppelvelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*. Kluwer Academic Publishers.

K. Forbes, E. Miltsakaki, R. Prasad, A. Sakar, A. Joshi, and B. Webber. 2001. D-LTAG system: Discourse parsing with a lexicalized tree adjoining grammar. In *Proceedings of the ESSLLI 2001: Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

J. Hobbs. 1985. On the coherence and structure of discourse. In *CSLI Technical Report 85-37*, Stanford, CA, USA. Center for the Study of Language and Information.

A. Killgarrif, P. Rychly, P. Smrz, and D. Tugwell. 2004. The sketch engine. In *Proceedings of Euralex, Lorient, France*, pages 105–116.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, Montreal, Canada.

E. Miltsakaki, N. Dinesh, R. Prasad, A. Joshi, and B. Webber. 2005. Experiments on sense annotation and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Catalonia.

J. Pustejovsky, C. Havasi, R. Saurí, P. Hanks, and A. Rumshisky. 2006. Towards a Generative Lexical resource: The Brandeis Semantic Ontology. In *Language Resources and Evaluation Conference, LREC 2006*, Genoa, Italy.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

J. Pustejovsky. 2001. Type construction and the logic of concepts. In *The Language of Word Meaning*. Cambridge University Press.

R. Saurí, M. Verhagen, and J. Pustejovsky. 2006. Annotating and recognizing event modality in text. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the HLT/NAACL Conference*, Edmonton, Canada.

M. Verhagen, I. Mani, R. Saurí, R. Knippen, J. Littman, and J. Pustejovsky. 2005. Automating temporal annotation within TARSQI. In *Proceedings of the ACL 2005*.

B. Webber, A. Joshi, E. Miltsakaki, R. Prasad, N. Dinesh, A. Lee, and K. Forbes. 2005. A short introduction to the penn discourse TreeBank. In *Copenhagen Working Papers in Language and Speech Processing*.

B. Wellner and M. Vilain. 2006. Leveraging machine readable dictionaries in discriminative sequence models. In *Language Resources and Evaluation Conference, LREC 2006*, Genoa, Italy.

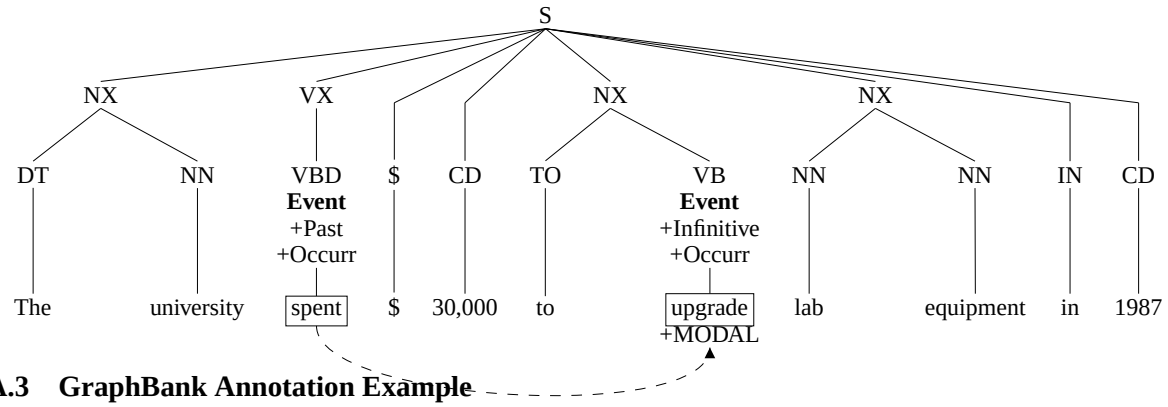
F. Wolf and E. Gibson. 2005. Representing discourse coherence: A corpus-based analysis. *Computational Linguistics*, 31(2):249–287.

A Appendix

A.1 Confusion Matrix

	elab	par	attr	ce	temp	contr	same	examp	expv	cond	gen
elab	488	3	7	3	1	0	2	4	0	3	1
par	6	110	2	2	0	8	2	0	0	2	0
attr	4	0	175	0	0	1	2	0	1	1	0
ce	18	9	3	26	3	2	2	0	0	0	0
temp	6	8	2	0	5	3	0	0	0	0	0
contr	4	12	0	0	0	38	0	0	3	0	0
same	3	9	2	2	0	2	54	0	0	0	0
examp	15	1	0	0	0	0	0	15	0	0	0
expv	3	1	1	0	1	4	0	0	2	0	0
cond	3	0	0	0	0	0	0	0	0	5	0
gen	0	0	0	0	0	0	0	0	0	0	0

A.2 SlinkET Example



A.3 GraphBank Annotation Example