Experiments on Sense Annotations and Sense Disambiguation of Discourse Connectives

Eleni Miltsakaki*, Nikhil Dinesh*, Rashmi Prasad*, Aravind Joshi*, and Bonnie Webber†
*University of Pennsylvania
Philadelphia, PA 19104 USA
{elenimi,nikhild,rjprasad,joshi}@linc.cis.upenn.edu
†University of Edinburgh
Edinburgh, EH8 9LW Scotland
bonnie@inf.ed.ac.uk

1 Introduction

Discourse connectives can be analyzed as discourse level predicates which project predicate-argument structure on a par with verbs at the sentence level. The Penn Discourse Treebank (PDTB) reflects this view in its design providing annotation of the discourse connectives and their arguments. Like verbs, discourse connectives have multiple senses. We present a set of manual sense annotation studies for three connectives whose arguments have been annotated in the PDTB. Using syntactic features computed from the Penn Treebank and a simple MaxEnt model, we have achieved some success in automatically disambiguating among their senses.

2 Background

The Penn Discourse Treebank (PDTB) project [11] builds on basic ideas presented originally in Webber and Joshi 1998 [13] – that connectives are discourse-level predicates which project predicate-argument structure on a par with verbs at the sentence level. In this framework, connectives are grouped into natural classes depending on how they project predicate-argument structure at the discourse level.

The PDTB corpus includes annotations of four types of connectives: subordinating conjunctions, coordinating conjunctions, adverbial connectives and implicit connectives.[1]

[1]Official release of the annotated corpus is expected by November 2005. The final number of annotations in the corpus will amount to approximately 25K: 15K annotations covering 96 explicit

Because discourse connectives (like verbs) can be polysemous, the final version of the corpus will also have annotated the semantic role of each argument of each type of connective. This paper presents our work to date on manual and automated sense annotation of discourse connectives as predicates.

3 Sense annotations of connectives

Senses can be distinguished or aggregated to a greater or lesser extent, depending on the needs of the application and the ability of annotators to distinguish them reliably. As a result of initial annotation experiments, we have grouped senses of the connectives *since*, *while* and *when* into the following classes (1) temporal senses that are not causally (contingently) related, (2) contrastive senses, (3) contingent senses, and (4) senses that are simultaneously temporal and causal.

Regarding *temporal* senses, we have not yet made finer distinctions [1]. The *contrastive* senses comprise *comparative*, *oppositive* and *concessive* senses, while the *contingent* senses comprise *causal* and *conditional* senses.

As one would expect, a **temporal** sense is identified when the events or situations expressed in the arguments of the connective are related temporally. All three connectives (*since*, *while* and *when*) have a temporal sense, as in the examples below. (In all examples, the first argument, Arg1, is shown in italics and the second argument, Arg2, in boldface. Arg2 is the argument which contains the clause that hosts the connective.)

(1) *there have been more than 100 mergers and acquisitions within the European paper industry* since **the most-recent wave of friendly takeovers was completed in the U.S. in 1986**.

(2) *The paper's local administrator, Maria Luz Lopez, was shot dead, and her mother wounded* while **her car was stopped for a red light**.

(3) *... the San Francisco earthquake hit* when **resources in the field already were stretched**.

Within the set of *contrastive* senses, a **comparative** sense is identified when two (or more) terms of the arguments are compared. *While* has such a comparative sense, as in (4) below.

(4) *The benchmark 11 3/4% Treasury bond due 2003/2007 rose 1/8 to 111 21/32 to yield 10.11%* while **the 12% issue of 1995 rose 3/32 to 103 23/32 to yield 11.01%**

An **oppositive** sense is identified when antithetical values are assigned to the terms of the arguments that are compared. A sense of *opposition* is identified for *while* and demonstrated in (5).

connectives identified in the corpus and 10K annotations of implicit connectives.

(5) *one ex-player claims he received $4,000 to $5,000 for his season football tickets* while **others said theirs brought only a few hundred dollars**

A **concessive** sense is identified when Arg1 violates an expectation raised in Arg2. Both *while* and *when* have a concessive sense, as shown in (6) and (7) respectively.

(6) While **the practice was discouraged in the past**, *the conference agreement is laced with veterans' hospitals, environmental projects and urban grants designated for specific communities.*

(7) *First Meridian's president, Roger V. Sala, portrayed himself as a "financial expert"* when **his qualifications largely consisted of a high-school diploma, work as a real-estate and insurance salesman, and a stint as supervisor at a highway toll booth**

Within the set of *contingent* senses, a **causal** sense is identified when the events or situations expressed in the arguments of the connective are causally related. As a diagnostic for this sense, we stipulated substitutability of the connective *because*. The connectives *since* and *when* both have causal senses, as in the examples below. In (9) *when* has a simultaneously *temporal* and *causal* sense (as was found to be the case for all causal interpretations of *when*).

(8) *It was a far safer deal for lenders* since **NWA had a healthier cash flow and more collateral on hand**

(9) When **the Trinity Repertory Theater named Anne Bogart its artistic director last spring**, *the nation's theatrical cognoscenti arched a collective eyebrow*

A **conditional** sense is identified when Arg2 sets up a truth condition for Arg1. In many cases conditional and causal interpretations were hard to distinguish. As a diagnostic of a conditional sense, we stipulated substitutability of the connective *if* but not *because*. Of the three connectives, only *when* has a conditional sense.

(10) However, when **powerful forces start shaking the market's structure**, *the more "earthquake-resistant" it is, the better its chance for survival.*

3.1 *Since*

For the subordinate conjunction *since* we identified the following three senses described above: a purely *temporal* sense, a purely *causal* sense, and the simultaneously *temporal and causal* sense.

An example of *since* expressing a temporal relation is shown earlier in (1). Example (8) demonstrates the causal sense of *since*. *Since* was only annotated as having a *causal* sense when only that interpretation was entertained. When both a temporal and a causal interpretation were possible the annotators were instructed to use the tag *temporal/causal* – (e.g., (11)). The annotators were instructed to use the tag *uncertain* when none of the given sense tags seemed appropriate.

(11) *and domestic car sales have plunged 19%* since **the Big Three ended many of their programs Sept. 30**

Two annotators independently carried out sense annotation of the 186 tokens of the connective *since* in the PDTB on which there was syntactic agreement about its arguments. Table (1) shows the distribution of *since*-senses per annotator. From the low number of *uncertain* labels, we take the three significant sense options as being sufficient to cover the range of interpretations of *since* in the PDTB corpus.

	Annot. 1	Annot.2
Temporal	74 (39.8%)	76 (40.9%)
Causal	90 (48.4%)	93 (50%)
T/C	21 (11.3%)	16 (8.6%)
Uncertain	1 (0.5%)	1 (0.5%)
Total	186	186

Table 1: Distribution of *since*-senses

To check reliability, we computed inter-annotator agreement between the two annotators, excluding the cases for which the annotators were not certain. Table (2) shows the inter-annotator agreement achieved between the two annotators. For 91.3% of the tokens the two annotators picked the same sense tag. Another 7.5% of the tokens had partial agreement, with one annotator assigning the combined T/C tag and the other annotator assigning either T or C. Disagreement was very low (1.1%).

Exact agree	169 (91.3%)
Partial agree	14 (7.5%)
Total agree	183 (98.9%)
Disagree	2 (1.1%)
Total	185

Table 2: Inter-annotator agreement for *since*- senses

3.2 *While*

For the connective *while*, we identified a *temporal*, as well as all three *contrastive* senses – *comparison*, *opposition* and *concession*. In the comparative sense of *while* two or more terms were compared. An example was shown in (4) earlier. Example (5) earlier illustrated the contrastive sense *opposition*. *Opposition* does not trigger

the inference that given Arg1, Arg2 is unexpected or contradictory. In this, it differs from the *concessive* sense of *while*, which expresses violation of expectation. In (12), for example, Arg2 creates the expectation that any collaboration between *Delmed* and *National Medical Care* will be discontinued, which is then challenged in Arg1.

(12) While **the discussions between Delmed and National Medical Care have been discontinued**, *Delmed will continue to supply dialysis products through National Medical after their exclusive agreement ends in March 1990*, Delmed said.

As before, sense annotators were instructed to use the tag *uncertain* if none of the available senses of *while* seemed appropriate. Two annotators annotated the senses of the first 100 tokens of *while* in the PDTB for which there was complete agreement on its arguments. Table (3) shows the distribution of *while*-senses per annotator.

	Annot. 1	Annot. 2
Temporal	22	19
Comparison	16	11
Opposition	43	30
Concession	8	31
Uncertain	11	9
Total	100	100

Table 3: Distribution of *while* senses

Table (4) shows the inter-annotator agreement for the annotation of *while*-senses, excluding cases for which the annotators were not certain (a total of 20). Note that agreement for the tokens that we annotated with a sense tag is reasonably high but the number of tokens marked as *uncertain* is also high, indicating that in several cases the proposed sense distinctions were hard to make. Specifically, 11 out of the 13 cases of disagreement were tagged as *concession* by Annot. 1 and as *opposition* by Annot. 2. The remaining two cases involved disagreement between the *opposition* and *temporal* tags.

Agree	67 (84%)
Disagree	13 (16%)
Total	80

Table 4: Inter-annotator agreement for *while*-senses

Earlier, we identified three senses of *while* - *concession*, *comparative* and *opposition* under the umbrella of **contrast**. With respect to *opposition*, Lakoff [9] defines **semantic opposition** as a form of contrast in which symmetric predicates (tall vs short) are predicated of distinct but comparable entities (Peter vs Bill), as in example (13)

(13) While Peter is tall, Bill is short.

(14) While Peter is intelligent, he is not a genius.

Example (14) differs from (13) in two ways: it talks about a single entity, and its Arg2 raises an expectation that is denied in Arg1. As noted earlier, a *contrastive* sense that raises an expectation is considered *concessive*. Instances of a *contrastive* sense that lack the requirements of either *opposition* or *concession* are considered simply *comparative*.

Subsequent to Lakoff [9], there has been debate in the literature as to whether these three are indeed different categories. Blakemore [2] argues for the merging of these categories to provide a unified analysis of *but*, while Jayez and Rossari [6] argue for maintaining a distinction on the basis of considerations in French. Because we allowed annotators to give more than one label to a sense, we felt that we could only gain by retaining all three. However, as the high number of uncertain cases suggests, a more careful analysis of the differences or lack thereof between the two senses is necessary before a final decision can be taken on them.

3.3 *When*

For the connective *when*, we identified the following four senses described above: a purely *temporal* sense, a simultaneously *temporal and causal* sense, a *conditional* sense and a *concessive* sense. As with *since*, the causal sense of *when* is identified when the situations expressed in its arguments are causally related. Unlike *since*, however, the combined tag *temporal/causal* (henceforth *T/C*) was in fact used because there were no instances in our data of a causal-only interpretation of *when*. An example of this combination of senses is given in (15).

(15) *Use of dispersants was approved* when **a test on the third day showed some positive results**, officials said.

Despite the significant overlap of causal and conditional relations, we found it useful to identify a *conditional* sense of *when*. As mentioned earlier, the conditional tag was used only when a causal paraphrase was not possible, as in (16), where substituting *because* for *when* gives an odd interpretation.

(16) When **you reach a point where a policy-making body is trying to shape administrative decisions**, *then that's a no-no in my book,*

	Annot. 1	Annot. 2
Temporal	44	37
T/C	22	28
Conditional	29	31
Concessive	1	2
Uncertain	4	2
Total	100	100

Agree	75 (79%)
Disagree	20 (21%)
Total	95

Table 5: Distribution of 'when' senses (left) and inter-annotator agreement for *when*- senses (right).

The concessive sense of *when* is identified when Arg1 violates an expectation raised in Arg2, as in (7). For the annotation of *when*-senses, two annotators annotated the senses of the first 100 tokens of *when* in the PDTB for which there was complete agreement on argument selection. Table (5) below shows the distribution of sense tags per annotator.

Table (5, right) shows the inter-annotator agreement achieved for the annotation of *when*-senses, excluding tokens for which the annotators were uncertain. There were a total of 5 tokens for which one or both annotators were uncertain. Out of the 20 tokens of disagreement, 8 involved disagreement between the *T/C* and *conditional* tags, 7 between the *temporal* and *conditional* tags and 5 between the *temporal* and *T/C* tags.

4 Sense Disambiguation

In this section, we describe experiments that attempt to automatically predict the sense of a connective given its arguments. We use the following notation to describe experiments. Suppose a connective has sense labels x,y, and z, then we denote an experiment to do the 3-way classification by (x,y,z). There are two variations we explored:

- Sense groups - In this case a sense could be a member of atmost one group. If we decided to group x and z in an experiment, we denote it by ({x,z},y) and in such experiments, we relabel z as x or vice-versa while training and testing the classifier.

- Sense subsets - In these experiments, we eliminated one or more senses from the training and test data. For example, if we were interested in how the well the classifier was able to distinguish between x and y, then we would denote this experiment by (x,y).

All experiments were carried out using a Maximum Entropy classifier as implemented by Mallet [10]. The reported results for all experiments are average accuracy in 10-fold cross-validation. For all experiments, we use a simple baseline, namely predict the most frequent sense. The accuracy of the baseline is enclosed in parantheses adjacent to it.

4.1 Feature Selection

We used as a guide in the search of features and the interpretation of results, the literature on resolving the temporal relations that hold between clauses in a discourse, which we specify in the full paper [7], [12].

For each argument of a connective, we extract the following four-dimensional vector from the gold-standard annotations of the Penn Treebank:

1. Form of auxiliary *have* - *Has, Have, Had* or *Not Found*.
2. Form of auxiliary *be* - *Present*(am, is, are), *Past* (was, were), *Been,* or *Not Found*.
3. Form of the head - *Present* (part-of-speech VBP or VBZ), *Past* (VBD), *Past Participal* (VBN), *Present Participal* (VBG).
4. Presence of a modal - *Found* or *Not Found*. The number of instances with a modal tense were few, so distinguishing between the various kinds of modals did not aid in increasing accuracy.

A sentence like *He has been going to the mall* would thus be assigned the vector [Has, Been, HeadPresentParticipal, ModalNotFound], while the sentence *He had gone to the mall* would be assigned the vector [Had, BeNotFound, HeadPastParticipal, ModalNotFound].

This feature helped in the disambiguation of all the connectives in this study, in varying degrees. The other feature used in all our experiments, tracked the presence of explicit temporal markers in Arg2, as in (1). These are specific years, months and the like. These markers affect the temporal categories of the clauses, as can be seen in (17) from M&S, where the presence of tomorrow shifts the tense from Present Progressive to the Futurate(non-modal future).

(17) He is leaving (tomorrow).

4.2 Since

For *since*, the only features used were the ones describe above, and the accuracy of the classifier in the various experiments run is shown in Table 6. From the results we can infer that these features aid in distinguishing the *temporal* from the *causal* sense. But it also shows that instances where both interpretations are licensed (*temporal/causal*) pattern with instances of *temporal* interpretation. To get a bet-

Experiment	Accuracy
(T,C,T/C)	75.5% (53.6%)
({T,T/C}, C)	90.1% (53.6%)
(T,{C,T/C})	74.2% (65.6%)
(T,C)	89.5% (60.9%)

Table 6: Average accuracy of sense disambiguation in 10-fold cross validation for *since*. T stands for Temporal, C for Causal, and T/C for Temporal/Causal. Accuracy of the baseline(predict most frequent sense) is parathesized.

ter understanding of how the features patterned with the senses, we computed the co-occurence of various configurations of the tense feature with the senses (Table 7). An examination of the *temporal/causal* instances with a perfective Arg1 revealed that Arg2 for these instances usually had an explicit temporal marker. This suggests that when Arg2 presents an alternate way to temporally ground the start of the consequent state in Arg1, the possibility of a *causal* interpretation might be entertained.

Feature	T	T/C	C
Arg1 Perfective	65.6%	26.2%	8.2%
Arg2 Simple Past	61.5%	28.8%	9.7%
Arg1 Simple Present	10%	2.5%	87.5%
Arg2 Simple Present	0%	0%	100%

Table 7: Cooccurence of a feature with a sense for *since*

4.3 While

In addition to the features described above, a few additional features were specific to *while*. The first was the relative position of Arg2 to Arg1.[2] This could be **preposed** as in (18), **postposed** as in (19) or **interposed** as in (20). These examples were annotated as having an *opposition*, *temporal*, and *concessive* senses, respectively, and we wanted to examine any correlation of position with sense.

(18) While **it is possible that the Big Green initiative will be ruled unconstitutional**, *it is of course conceivable that in modern California it could slide through*.

(19) *A nurse contracted the virus* while **injecting an AIDS patient**.

(20) *The basket product*, while **it has got off to a slow start**, *is being supported by some big brokerage firms*.

[2]We tried this feature for *since* and *when*, but they were detrimental to performance.

Table 9 shows such a correlation: the interposed position correlated with *concessive*, while the preposed position correlated with one of the two *contrastive* senses.

The two other features we used were targeted at distinguishing between the *comparative* and *concessive* senses, as in (21) and (22). The first feature checked if the same verb was used in both arguments, and the second checked if the adverb *not* was present in the head verb phrase of a single argument.

(21) *The benchmark 11 3/4% Treasury bond due 2003/2007 rose 1/8 to 111 21/32 to yield 10.11%* while **the 12% issue of 1995 rose 3/32 to 103 23/32 to yield 11.01%**.

(22) While **the third-quarter figures may appear relatively bullish**, *it would take a significantly stronger figure to alter market perceptions that the economy is softening.*

The accuracy of the classifier in the various experiments run is shown in Table 8. While the distinction between the *temporal* and *non-temporal* senses (line 3) is strong, the distinction among *non-temporal* senses (line 4) stands to improve. The features used in these experiments, namely the presence of the same head verb in Arg1 and Arg2 and the presence of *not* in one of the arguments, give us possible directions for future inquiry. Specifically, it appears that improved lexical knowledge can aid in making better distinctions. Once a larger scale annotation of these senses are available, the use of resources like Wordnet [5], Verbnet [8], VerbOcean [3] and kernel-based tree similarity metrics [4] will be investigated.

Experiment	Accuracy
(T,Con,Comp,Opp)	71.8% (47.4%)
(T, Con, {Comp,Opp})	80.8% (62.8%)
(T,{Con,Comp,Opp})	89.7% (79.1%)
(Con,Comp,Opp)	71.9% (58.7%)

Table 8: Average accuracy of sense disambiguation in 10-fold cross validation for *while*. T stands for Temporal, Con for Concessive, Opp for Opposition, and Comp for Compare.Accuracy of the baseline(predict most frequent sense) is parathesized.

4.4 When

The features used for *when* were the same as those used for *since*, namely the tense vector from each argument, and the explicit time feature. The results for the experiments run are show in Table 10. With these features the classifier was able to make some distinction between the *temporal* and *conditional* senses, but it failed quite badly on distinguishing between the *temporal* and *temporal/causal* senses.

Feature	T	Con	Comp	Opp
Preposed	0.1%	37.4%	0%	62.5%
Interposed	0%	75%	0%	25%
Arg2 Non-Finite Participal	73.3%	6.7%	0%	20%
Same verb	2.5%	0%	62.5%	25%
Single *not* Arg	0%	62.5%	0%	27.5%

Table 9: Cooccurence of a feature with a sense for *while*

Experiment	Accuracy
(T,T/C,Cond)	61.6% (47.6%)
(T,{T/C,Cond})	50% (52.3%)
({T,T/C},Cond)	82.6% (69.1%)

Table 10: Average accuracy of sense disambiguation in 10-fold cross validation for *when*. T stands for Temporal, T/C for Temporal/Causal, and Cond for Conditional. Accuracy of the baseline(predict most frequent sense) is parathesized.

Table 11 shows the cooccurence of feature patterns with sense, and it can be seen that the *temporal/causal* sense tends to exhibit the same patterns as the *temporal* sense. This is in parallel with the results for *since* in Table 6. The patterning of the *conditional* sense of *when* with tense is also worth further investigation.

5 Conclusions and future work

We have identified several features that helped in disambiguating the three connectives in the study. As we carry out more sense annotation of connectives in the PDTB, we will develop a better understanding of their specificity to these connectives or their general applicability. The features used in this study may or may not be applicable across genres, as an informal (single-annotator) study of fiction from DAVIES (`http://view.byu.edu`) shows a very different distribution of senses for the connectives *while* and *when*.

Feature	T	C	Cond
Arg1 Simple Past	54.1%	40.5%	5.4%
Arg2 Simple Past	54.3%	42.9%	2.8%
Arg1 Simple Present	30%	0%	70%
Arg2 Simple Present	33.3%	0%	66.7%

Table 11: Cooccurence of a feature with a sense for *when*

Even though there was a relatively small number of instances of annotated connectives, an improvement of 15-20% over the baseline was seen across the board. This suggests that one could hope to disambiguate between the senses of connectives to a reasonable degree given the current state-of-the-art, and that the annotation of senses provided by the PDTB will be a very useful resource.

References

[1] James Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM*, volume 6, pages 832–843, 1983.

[2] Diane Blakemore. *Semantic Constraints on Relevance*. Blackwell, Oxford, 1987.

[3] Timothy Chklovski and Patrick Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of EMNLP*, 2004.

[4] Michael Collins and Nigel Duffy. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *ACL*, 2002.

[5] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT publications, 1988.

[6] Jacques Jayez and Corinne Rossari. Pragmatic connectives as predicates. In Patrick Saint-Dizier, editor, *Predicative Structures in Natural Language and Lexical Knowledge Bases*, pages 306–340. Kluwer Academic Press, Dordrecht, 1998.

[7] Andrew Kehler. *Formalizing the Dynamics of Information*, chapter Resolving Temporal Relations using Tense Meaning and Discourse Interpretation. CSLI Publications, 2000.

[8] Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence*, 2000.

[9] R. Lakoff. *Studies in Linguistic Semantics*, chapter If's, And's, and But's about conjunction, pages 114–149. Holt, Rinehart, Winston, 1971.

[10] Andrew K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[11] Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. The Penn Discourse Treebank. In *4th International Conference on Language Rescourses and Evaluation (LREC 2004) Lisbon*, 2004.

[12] Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14:15–28, June 1988.

[13] Bonnie Webber and Aravind Joshi. Anchoring a lexicalized tree adjoining grammar for discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal*, pages 8–92. Montreal, Canada, 1998.