Acquiring the Meaning of Discourse Markers

Ben Hutchinson
School of Informatics
University of Edinburgh
B.Hutchinson@sms.ed.ac.uk

Abstract

This paper applies machine learning techniques to acquiring aspects of the meaning of discourse markers. Three subtasks of acquiring the meaning of a discourse marker are considered: learning its **polarity**, **veridicality**, and **type** (i.e. causal, temporal or additive). Accuracy of over 90% is achieved for all three tasks, well above the baselines.

1 Introduction

This paper is concerned with automatically acquiring the meaning of discourse markers. By considering the distributions of individual *tokens* of discourse markers, we classify discourse markers along three dimensions upon which there is substantial agreement in the literature: **polarity**, **veridicality** and **type**. This approach of classifying linguistic *types* by the distribution of linguistic *tokens* makes this research similar in spirit to that of Baldwin and Bond (2003) and Stevenson and Merlo (1999).

Discourse markers signal relations between discourse units. As such, discourse markers play an important role in the parsing of natural language discourse (Forbes et al., 2001; Marcu, 2000), and their correspondence with discourse relations can be exploited for the unsupervised learning of discourse relations (Marcu and Echihabi, 2002). In addition, generating natural language discourse requires the appropriate selection and placement of discourse markers (Moser and Moore, 1995; Grote and Stede, 1998). It follows that a detailed account of the semantics and pragmatics of discourse markers would be a useful resource for natural language processing.

Rather than looking at the finer subtleties in meaning of particular discourse markers (e.g. Bestgen et al. (2003)), this paper aims at a broad scale classification of a subclass of discourse markers: structural connectives. This breadth of coverage is of particular importance for discourse parsing, where a wide range of linguistic realisations must be catered for. This work can be seen as orthogonal to that of Di Eugenio et al. (1997), which addresses the problem of learning *if* and *where* discourse markers should be generated.

Unfortunately, the manual classification of large numbers of discourse markers has proven to be a difficult task, and no complete classification yet exists. For example, Knott (1996) presents a list of around 350 discourse markers, but his taxonomic classification, perhaps the largest classification in the literature, accounts for only around 150 of these. A general method of automatically classifying discourse markers would therefore be of great utility, both for English and for languages with fewer manually created resources. This paper constitutes a step in that direction. It attempts to classify discourse markers whose classes are already known, and this allows the classifier to be evaluated empirically.

The proposed task of learning automatically the meaning of discourse markers raises several questions which we hope to answer:

Q1. Difficulty How hard is it to acquire the meaning of discourse markers? Are some aspects of meaning harder to acquire than others?

Q2. Choice of features What features are useful for acquiring the meaning of discourse markers? Does the optimal choice of features depend on the aspect of meaning being learnt?

Q3. Classifiers Which machine learning algorithms work best for this task? Can the right choice of empirical features make the classification problems linearly separable?

Q4. Evidence Can corpus evidence be found for the existing classifications of discourse markers? Is there empirical evidence for a separate class of TEMPORAL markers?

We proceed by first introducing the classes of discourse markers that we use in our experiments. Section 3 discusses the database of discourse markers used as our corpus. In Section 4 we describe our experiments, including choice of features. The results are presented in Section 5. Finally, we conclude and discuss future work in Section 6.

2 Discourse markers

Discourse markers are lexical items (possibly multiword) that signal relations between propositions, events or speech acts. Examples of discourse markers are given in Tables 1, 2 and 3. In this paper we will focus on a subclass of discourse markers known as *structural connectives.* These markers, even though they may be multiword expressions, function syntactically as if they were coordinating or subordinating conjunctions (Webber et al., 2003).

The literature contains many different classifications of discourse markers, drawing upon a wide range of evidence including textual cohesion (Halliday and Hasan, 1976), hypotactic conjunctions (Martin, 1992), cognitive plausibility (Sanders et al., 1992), substitutability (Knott, 1996), and psycholinguistic experiments (Louwerse, 2001). Nevertheless there is also considerable agreement. Three dimensions of classification that recur, albeit under a variety of names, are **polarity**, **veridicality** and **type**. We now discuss each of these in turn.

2.1 Polarity

Many discourse markers signal a concession, a contrast or the denial of an expectation. These markers have been described as having the feature **polarity**=NEG-POL. An example is given in (1).

(1) Suzy's part-time, **but** she does more work than the rest of us put together. (Taken from Knott (1996, p. 185))

This sentence is true if and only if Suzy both is part-time and does more work than the rest of them put together. In addition, it has the additional effect of signalling that the fact Suzy does more work is surprising — it denies an expectation. A similar effect can be obtained by using the connective *and* and adding more context, as in (2)

(2) Suzy's efficiency is astounding. She's part-time, **and** she does more work than the rest of us put together.

The difference is that although it is possible for *and* to co-occur with a negative polarity discourse relation, it need not. Discourse markers like *and* are said to have the feature **polarity**=POS-POL. [1] On the other hand, a NEG-POL discourse marker like *but* always co-occurs with a negative polarity discourse relation.

The gold standard classes of POS-POL and NEG-POL discourse markers used in the learning experiments are shown in Table 1. The gold standards for all three experiments were compiled by consulting a range of previous classifications (Knott, 1996; Knott and Dale, 1994; Louwerse, 2001). [2]

POS-POL	NEG-POL
after, and, as, as soon as, because, before, considering that, ever since, for, given that, if, in case, in order that, in that, insofar as, now, now that, on the grounds that, once, seeing as, since, so, so that, the instant, the moment, then, to the extent that, when, whenever	although, but, even if, even though, even when, only if, only when, or, or else, though, unless, until, whereas, yet

Table 1: Discourse markers used in the **polarity** experiment

2.2 Veridicality

A discourse relation is *veridical* if it implies the truth of both its arguments (Asher and Lascarides, 2003), otherwise it is not. For example, in (3) it is not necessarily true either that David can stay up or that he promises, or will promise, to be quiet. For this reason we will say *if* has the feature **veridicality**=NON-VERIDICAL.

(3) David can stay up **if** he promises to be quiet.

The disjunctive discourse marker *or* is also NON-VERIDICAL, because it does not imply that both of its arguments are true. On the other hand, *and* does imply this, and so has the feature **veridicality**=VERIDICAL.

The VERIDICAL and NON-VERIDICAL discourse markers used in the learning experiments are shown in Table 2. Note that the **polarity** and **veridicality** are independent, for example *even if* is both NEG-POL and NON-VERIDICAL.

2.3 Type

Discourse markers like *because* signal a CAUSAL relation, for example in (4).

[1] An alternative view is that discourse markers like *and* are underspecified with respect to polarity (Knott, 1996). In this account, discourse markers have positive polarity only if they can never be paraphrased using a discourse marker with negative polarity. Interpreted in these terms, our experiment aims to distinguish negative polarity discourse markers from all others.

[2] An effort was made to exclude discourse markers whose classification could be contentious, as well as ones which showed ambiguity across classes. Some level of judgement was therefore exercised by the author.

VERIDICAL	NON-VERIDICAL
after, although, and, as, as soon as, because, but, considering that, even though, even when, ever since, for, given that, in order that, in that, insofar as, now, now that, on the grounds that, once, only when, seeing as, since, so, so that, the instant, the moment, then, though, to the extent that, until, when, whenever, whereas, while, yet	assuming that, even if, if, if ever, if only, in case, on condition that, on the assumption that, only if, or, or else, supposing that, unless

Table 2: Discourse markers used in the **veridicality** experiment

(4) The tension in the boardroom rose sharply **because** the chairman arrived.

As a result, *because* has the feature **type**=CAUSAL. Other discourse markers that express a temporal relation, such as *after*, have the feature **type**=TEMPORAL. Just as a POS-POL discourse marker can occur with a negative polarity discourse relation, the context can also supply a causal relation even when a TEMPORAL discourse marker is used, as in (5).

(5) The tension in the boardroom rose sharply **after** the chairman arrived.

If the relation a discourse marker signals is neither CAUSAL or TEMPORAL it has the feature **type**=ADDITIVE.

The need for a distinct class of TEMPORAL discourse relations is disputed in the literature. On the one hand, it has been suggested that TEMPORAL relations are a subclass of ADDITIVE ones on the grounds that the temporal reference inherent in the marking of tense and aspect “more or less” fixes the temporal ordering of events (Sanders et al., 1992). This contrasts with arguments that resolving discourse relations and temporal order occur as distinct but inter-related processes (Lascarides and Asher, 1993). On the other hand, several of the discourse markers we count as TEMPORAL, such as *as soon as*, might be described as CAUSAL (Oberlander and Knott, 1995). One of the results of the experiments described below is that corpus evidence suggests ADDITIVE, TEMPORAL and CAUSAL discourse markers have distinct distributions.

The ADDITIVE, TEMPORAL and CAUSAL discourse markers used in the learning experiments are shown in Table 3. These features are independent of the previous ones, for example *even though* is CAUSAL, VERIDICAL and NEG-POL.

ADDITIVE	TEMPORAL	CAUSAL
and, but, whereas	after, as soon as, before, ever since, now, now that, once, until, when, whenever	although, because, even though, for, given that, if, if ever, in case, on condition that, on the assumption that, on the grounds that, provided that, providing that, so, so that, supposing that, though, unless

Table 3: Discourse markers used in the **type** experiment

3 Corpus

The data for the experiments comes from a database of sentences collected automatically from the British National Corpus and the world wide web (Hutchinson, 2004). The database contains example sentences for each of 140 discourse structural connectives.

Many discourse markers have surface forms with other usages, e.g. *before* in the phrase *before noon.* The following procedure was therefore used to select sentences for inclusion in the database. First, sentences containing a string matching the surface form of a structural connective were extracted. These sentences were then parsed using a statistical parser (Charniak, 2000). Potential structural connectives were then classified on the basis of their syntactic context, in particular their proximity to S nodes. Figure 1 shows example syntactic contexts which were used to identify discourse markers.

```
(S ...) (CC and) (S...)
(SBAR (IN after) (S...))
(PP (IN after) (S...))
(PP (VBN given) (SBAR (IN that) (S...)))
(NP (DT the) (NN moment) (SBAR...))
(ADVP (RB as) (RB long)
      (SBAR (IN as) (S...)))
(PP (IN in) (SBAR (IN that) (S...)))
```

Figure 1: Identifying structural connectives

It is because structural connectives are easy to identify in this manner that the experiments use only this subclass of discourse markers. Due to both parser errors, and the fact that the syntactic heuristics are not foolproof, the database contains noise. Manual analysis of a sample of 500 sentences revealed about 12% of sentences do not contain the discourse marker they are supposed to.

Of the discourse markers used in the experiments, their frequencies in the database ranged from 270 for *the instant* to 331,701 for *and*. The mean number of instances was 32,770, while the median was 4,948.

4 Experiments

This section presents three machine learning experiments into automatically classifying discourse markers according to their **polarity**, **veridicality** and **type**. We begin in Section 4.1 by describing the features we extract for each discourse marker token. Then in Section 4.2 we describe the different classifiers we use. The results are presented in Section 4.3.

4.1 Features used

We only used structural connectives in the experiments. This meant that the clauses linked syntactically were also related at the discourse level (Webber et al., 2003). Two types of features were extracted from the conjoined clauses. Firstly, we used lexical co-occurrences with words of various parts of speech. Secondly, we used a range of linguistically motivated syntactic, semantic, and discourse features.

4.1.1 Lexical co-occurrences

Lexical co-occurrences have previously been shown to be useful for discourse level learning tasks (Lapata and Lascarides, 2004; Marcu and Echihabi, 2002). For each discourse marker, the words occurring in their superordinate (main) and subordinate clauses were recorded,[3] along with their parts of speech. We manually clustered the Penn Treebank parts of speech together to obtain coarser grained syntactic categories, as shown in Table 4.

New label	Penn Treebank labels
vb	vb vbd vbg vbn vbp vbz
nn	nn nns nnp
jj	jj jjr jjs
rb	rb rbr rbs
aux	aux auxg md
prp	prp prp$
in	in

Table 4: Clustering of POS labels

We then lemmatised each word and excluded all lemmas with a frequency of less than 1000 per million in the BNC. Finally, words were attached a prefix of either SUB_ or SUPER_ according to whether they occurred in the sub- or superordinate clause linked by the marker. This distinguished, for example, between occurrences of *then* in the antecedent (subordinate) and consequent (main) clauses linked by *if*.

We also recorded the presence of other discourse markers in the two clauses, as these had previously been found to be useful on a related classification task (Hutchinson, 2003). The discourse markers used for this are based on the list of 350 markers given by Knott (1996), and include multiword expressions. Due to the sparser nature of discourse markers, compared to verbs for example, no frequency cutoffs were used.

4.1.2 Linguistically motivated features

These included a range of one and two dimensional features representing more abstract linguistic information, and were extracted through automatic analysis of the parse trees.

One dimensional features

Two one dimensional features recorded the location of discourse markers. POSITION indicated whether a discourse marker occurred between the clauses it linked, or before both of them. It thus relates to information structuring. EMBEDDING indicated the level of embedding, in number of clauses, of the discourse marker beneath the sentence's highest level clause. We were interested to see if some types of discourse relations are more often deeply embedded.

The remaining features recorded the presence of linguistic features that are localised to a particular clause. Like the lexical co-occurrence features, these were indexed by the clause they occurred in: either SUPER or SUB.

We expected negation to correlate with negative polarity discourse markers, and approximated negation using four features. NEG-SUBJ and NEG-VERB indicated the presence of subject negation (e.g. *nothing*) or verbal negation (e.g. *n't*). We also recorded the occurrence of a set of negative polarity items (NPI), such as *any* and *ever*. The features NPI-AND-NEG and NPI-WO-NEG indicated whether an NPI occurred in a clause with or without verbal or subject negation.

Eventualities can be placed or ordered in time us-

[3] For coordinating conjunctions, the left clause was taken to be superordinate/main clause, the right, the subordinate clause.

ing not just discourse markers but also temporal expressions. The feature TEMPEX recorded the number of temporal expressions in each clause, as returned by a temporal expression tagger (Mani and Wilson, 2000).

If the main verb was an inflection of *to be* or *to do* we recorded this using the features BE and DO. Our motivation was to capture any correlation of these verbs with states and events respectively.

If the final verb was a modal auxiliary, this ellipsis was evidence of strong cohesion in the text (Halliday and Hasan, 1976). We recorded this with the feature VP-ELLIPSIS. Pronouns also indicate cohesion, and have been shown to correlate with subjectivity (Bestgen et al., 2003). A class of features PRONOUNS^X represented pronouns, with X denoting either 1st person, 2nd person, or 3rd person animate, inanimate or plural.

The syntactic structure of each clause was captured using two features, one finer grained and one coarser grained. STRUCTURAL-SKELETON identified the major constituents under the S or VP nodes, e.g. a simple double object construction gives "NP VB NP NP". ARGS identified whether the clause contained an (overt) object, an (overt) subject, or both, or neither.

The overall size of a clause was represented using four features. WORDS, NPS and PPS recorded the numbers of words, NPs and PPs in a clause (not counting embedded clauses). The feature CLAUSES counted the number of clauses embedded beneath a clause.

Two dimensional features

These features all recorded combinations of linguistic features across the two clauses linked by the discourse marker. For example the MOOD feature would take the value <DECL,IMP> for the sentence *John is coming, but don't tell anyone!*

These features were all determined automatically by analysing the auxiliary verbs and the main verbs' POS tags. The features and the possible values for each clause were as follows: MODALITY: one of FUTURE, ABILITY or NULL; MOOD: one of DECL, IMP or INTERR; PERFECT: either YES or NO; PROGRESSIVE: either YES or NO; TENSE: either PAST or PRESENT.

4.2 Classifier architectures

Two different classifiers, based on local and global methods of comparison, were used in the experiments. The first, 1 Nearest Neighbour (1NN), is an instance based classifier which assigns each marker to the same class as that of the marker nearest to it. For this, three different distance metrics were explored. The first metric was the Euclidean distance function L_2, shown in (6), applied to probability distributions.

$$L_2(p,q) = \sqrt{\sum_x (p(x) - q(x))^2} \qquad (6)$$

The second, KL_α, is a smoothed variant of the information theoretic Kullback-Leibner divergence (Lee, 2001, with $\alpha = 0.95$). Its definition is given in (7).

$$KL_\alpha(p,q) = \sum_x p(x) \log \frac{p(x)}{\alpha q(x) + (1-\alpha)p(x)} \qquad (7)$$

The third metric, $Jacc_t$, is a t-test weighted adaption of the Jaccard coefficient (Curran and Moens, 2002). In it basic form, the Jaccard coefficient is essentially a measure of how much two distributions overlap. The t-test variant weights co-occurrences by the strength of their collocation, using the following function:

$$wt(w_i, x) = \frac{p(w_i, x) - p(w_i)p(x)}{\sqrt{p(w_i)p(x)}}$$

This is then used define the weighted version of the Jaccard coefficient, as shown in (8). The words associated with distributions p and q are indicated by w_p and w_q, respectively.

$$Jacc_t(p,q) = \frac{\sum_x min(wt(w_p, x), wt(w_q, x))}{\sum_x max(wt(w_p, x), wt(w_q, x))} \qquad (8)$$

KL_α and $Jacc_t$ had previously been found to be the best metrics for other tasks involving lexical similarity. L_2 is included to indicate what can be achieved using a somewhat naive metric.

The second classifier used, Naive Bayes, takes the overall distribution of each class into account. It essentially defines a decision boundary in the form of a curved hyperplane. The Weka implementation (Witten and Frank, 2000) was used for the experiments, with 10-fold cross-validation.

4.3 Results

We began by comparing the performance of the 1NN classifier using the various lexical co-occurrence features against the gold standards. The results using all lexical co-occurrences are shown

		All POS			Best single POS			Best
Task	Baseline	L_2	KL_α	$Jacc_t$	L_2	KL_α	$Jacc_t$	subset
polarity	67.4	74.4	72.1	74.4	76.7 (rb)	83.7 (rb)	76.7 (rb)	83.7[a]
veridicality	73.5	81.6	85.7	75.5	83.7 (nn)	91.8 (vb)	87.8 (vb)	91.8[b]
type	58.1	74.2	64.5	81.8	74.2 (in)	74.2 (rb)	77.4 (jj)	87.8[c]

[a] Using KL_α and either rb or DMs+rb. [b] Using both KL_α and vb, and $Jaccard_t$ and vb+in. [c] Using KL_α and vb+aux+in

Table 5: Results using the 1NN classifier on lexical co-occurrences

Feature	Positively correlated discourse marker co-occurrences
POS-POL	*though*$_\top$, *but*$_\top$, *although*$_\top$, *assuming that*$_\top$
NEG-POL	*otherwise*$_\perp$, *still*$_\top$, *in truth*$_\perp$, *still*$_\perp$, *after that*$_\top$, *in this way*$_\top$, *granted that*$_\top$, *in contrast*$_\top$, *by then*$_\perp$, *in the event*$_\perp$
VERIDICAL	*obviously*$_\perp$, *now*$_\perp$, *even*$_\perp$, *indeed*$_\top$, *once more*$_\top$, *considering that*$_\top$, *even after*$_\top$, *once more*$_\perp$, *at first sight*$_\top$
NON-VERIDICAL	*or*$_\top$, *no doubt*$_\top$, *in turn*$_\top$, *then*$_\top$, *by all means*$_\top$, *before then*$_\perp$
ADDITIVE	*also*$_\perp$, *in addition*$_\perp$, *still*$_\perp$, *only*$_\perp$, *at the same time*$_\perp$, *clearly*$_\perp$, *naturally*$_\perp$, *now*$_\perp$, *of course*$_\perp$
TEMPORAL	*back*$_\top$, *once more*$_\top$, *like*$_\top$, *and*$_\top$, *once more*$_\perp$, *which was why*$_\top$, . . .
CAUSAL	*again*$_\top$,*altogether*$_\perp$,*back*$_\perp$,*finally*$_\perp$, *also*$_\top$, *thereby*$_\perp$, *at once*$_\perp$, *while*$_\top$, *clearly*$_\top$, . . .

Table 6: Most informative discourse marker co-occurrences in the super- ($\top$) and subordinate ($\perp$) clauses

in Table 5. The baseline was obtained by assigning discourse markers to the largest class, i.e. with the most *types*. The best results obtained using just a single POS class are also shown. The results across the different metrics suggest that adverbs and verbs are the best single predictors of **polarity** and **veridicality**, respectively.

We next applied the 1NN classifier to co-occurrences with discourse markers. The results are shown in Table 7. The results show that for each task 1NN with the weighted Jaccard coefficient performs at least as well as the other three classifiers.

	1NN with metric:			Naive
Task	L_2	KL_α	$Jacc_t$	Bayes
polarity	74.4	81.4	81.4	81.4
veridicality	83.7	79.6	83.7	73.5
type	74.2	80.1	80.1	58.1

Table 7: Results using co-occurrences with DMs

We also compared using the following combinations of different parts of speech: vb + aux, vb + in, vb + rb, nn + prp, vb + nn + prp, vb + aux + rb, vb + aux + in, vb + aux + nn + prp, nn + prp + in, DMs + rb, DMs + vb and DMs + rb + vb. The best results obtained using all combinations tried are shown in the last column of Table 5. For DMs + rb, DMs + vb and DMs + rb + vb we also tried weighting the co-occurrences so that the sums of the co-occurrences with each of verbs, adverbs and discourse markers were equal. However this did not lead to any better results.

One property that distinguishes $Jacc_t$ from the other metrics is that it weights features the strength of their collocation. We were therefore interested to see which co-occurrences were most informative. Using Weka's feature selection utility, we ranked discourse marker co-occurrences by their information gain when predicting **polarity**, **veridicality** and **type**. The most informative co-occurrences are listed in Table 6. For example, if *also* occurs in the subordinate clause then the discourse marker is more likely to be ADDITIVE.

The 1NN and Naive Bayes classifiers were then applied to co-occurrences with just the DMs that were most informative for each task. The results, shown in Table 8, indicate that the performance of 1NN drops when we restrict ourselves to this subset. [4] However Naive Bayes outperforms all previous 1NN classifiers.

	Base-	1NN with:		Naive
Task	line	L_2	KL_α	Bayes
polarity	67.4	72.1	69.8	**90.7**
veridicality	73.5	85.7	77.6	**91.8**
type	58.1	67.7	58.1	**93.5**

Table 8: Results using most informative DMs

[4] The $Jacc_t$ metric is omitted because it essentially already has its own method of factoring in informativity.

Feature	Positively correlated features
POS-POL	*No significantly informative predictors correlated positively*
NEG-POL	NEG-VERBAL$_{\top}$, NEG-SUBJ$_{\top}$, ARGS=NONE$_{\top}$, MODALITY=<ABILITY,ABILITY>
VERIDICAL	VERB=BE$_{\top}$, WORDS$_{\bot}$, WORDS$_{\top}$, MODALITY=<NULL,NULL>
NON-VERID	TEMPEX$_{\top}$, PRONOUN$^{pers=2}_{\top}$, PRONOUN$^{pers=2}_{\bot}$
ADDITIVE	WORDS$_{\bot}$, WORDS$_{\top}$, CLAUSES$_{\bot}$, MODALITY=<ABILITY,FUTURE>, MODALITY=<ABILITY,ABILITY>, NPS$_{\bot}$, MODALITY=<FUTURE,FUTURE>, MOOD=<DECLARATIVE,DECLARATIVE>
TEMPORAL	EMBEDDING=7, PRONOUN$^{pers=3anim}_{\top}$, MOOD=<INTERROGATIVE,DECLARATIVE>
CAUSAL	NEG-SUBJ$_{\bot}$, NEG-VERBAL$_{\bot}$, NPI-WO-NEG$_{\bot}$, NPI-AND-NEG$_{\bot}$, MODALITY=<NULL,FUTURE>

Table 9: The most informative linguistically motivated predictors for each class. The indices $\top$ and $\bot$ indicate that a one dimensional feature belongs to the superordinate or subordinate clause, respectively.

Weka's feature selection utility was also applied to all the linguistically motivated features described in Section 4.1.2. The most informative features are shown in Table 9. Naive Bayes was then applied using both all the linguistically motivated features, and just the most informative ones. The results are shown in Table 10.

Task	Baseline	All features	Most informative
polarity	67.4	74.4	72.1
veridicality	73.5	77.6	79.6
type	58.1	64.5	77.4

Table 10: Naive Bayes and linguistic features

5 Discussion

The results demonstrate that discourse markers can be classified along three different dimensions with an accuracy of over 90%. The best classifiers used a global algorithm (Naive Bayes), with co-occurrences with a subset of discourse markers as features. The success of Naive Bayes shows that with the right choice of features the classification task is highly separable. The high degree of accuracy attained on the **type** task suggests that there is empirical evidence for a distinct class of TEMPORAL markers.

The results also provide empirical evidence for the correlation between certain linguistic features and types of discourse relation. Here we restrict ourselves to making just five observations. Firstly, verbs and adverbs are the most informative parts of speech when classifying discourse markers. This is presumably because of their close relation to the main predicate of the clause. Secondly, Table 6 shows that the discourse marker DM in the structure *X, but/though/although Y DM Z* is more likely to be signalling a positive polarity discourse relation between Y and Z than a negative polarity one. This suggests that a negative polarity discourse relation is less likely to be embedded directly beneath another negative polarity discourse relation. Thirdly, negation correlates with the main clause of NEG-POL discourse markers, and it also correlates with subordinate clause of CAUSAL ones. Fourthly, NON-VERIDICAL correlates with second person pronouns, suggesting that a writer/speaker is less likely to make assertions about the reader/listener than about other entities. Lastly, the best results with knowledge poor features, i.e. lexical co-occurrences, were better than those with linguistically sophisticated ones. It may be that the sophisticated features are predictive of only certain subclasses of the classes we used, e.g. hypotheticals, or signallers of contrast.

6 Conclusions and future work

We have proposed corpus-based techniques for classifying discourse markers along three dimensions: **polarity**, **veridicality** and **type**. For these tasks we were able to classify with accuracy rates of 90.7%, 91.8% and 93.5% respectively. These equate to error reduction rates of 71.5%, 69.1% and 84.5% from the baseline error rates. In addition, we determined which features were most informative for the different classification tasks.

In future work we aim to extend our work in two directions. Firstly, we will consider finer-grained classification tasks, such as learning whether a causal discourse marker introduces a cause or a consequence, e.g. distinguishing *because* from *so*. Secondly, we would like to see how far our results can be extended to include adverbial discourse markers, such as *instead* or *for example*, by using just features of the clauses they occur in.

Acknowledgements

I would like to thank Mirella Lapata, Alex Lascarides, Bonnie Webber, and the three anonymous reviewers for their comments on drafts of this paper. This research was supported by EPSRC Grant GR/R40036/01 and a University of Sydney Travelling Scholarship.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Timothy Baldwin and Francis Bond. 2003. Learning the countability of English nouns from corpus data. In *Proceedings of ACL 2003*, pages 463–470.

Yves Bestgen, Liesbeth Degand, and Wilbert Spooren. 2003. On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: An exploratory study. In *Proceedings of the MAD'03 workshop on Multidisciplinary Approaches to Discourse*, October.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, Seattle, Washington, USA.

James R. Curran and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition*, pages 59–67, Philadelphia, PA, USA.

Barbara Di Eugenio, Johanna D. Moore, and Massimo Paolucci. 1997. Learning features that predict cue usage. In *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL97)*, Madrid, Spain, July.

Katherine Forbes, Eleni Miltsakaki, Rashmi Prasad, Anoop Sarkar, Aravind Joshi, and Bonnie Webber. 2001. D-LTAG system—discourse parsing with a lexicalised tree adjoining grammar. In *Proceedings of the ESSLI 2001 Workshop on Information Structure, Discourse Structure, and Discourse Semantics*, Helsinki, Finland.

Brigitte Grote and Manfred Stede. 1998. Discourse marker choice in sentence planning. In Eduard Hovy, editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 128–137. Association for Computational Linguistics, New Brunswick, New Jersey.

M. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman.

Ben Hutchinson. 2003. Automatic classification of discourse markers by their co-occurrences. In *Proceedings of the ESSLLI 2003 workshop on Discourse Particles: Meaning and Implementation*, Vienna, Austria.

Ben Hutchinson. 2004. Mining the web for discourse markers. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Alistair Knott and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

Alistair Knott. 1996. *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, University of Edinburgh.

Mirella Lapata and Alex Lascarides. 2004. Inferring sentence-internal temporal relations. In *In Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting*, Boston, MA.

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

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*, pages 65–72.

Max M Louwerse. 2001. An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12(3):291–315.

Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 69–76, New Brunswick, New Jersey.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.

Jim Martin. 1992. *English Text: System and Structure*. Benjamin, Amsterdam.

M. Moser and J. Moore. 1995. Using discourse analysis and automatic text generation to study discourse cue usage. In *Proceedings of the AAAI 1995 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 92–98.

Jon Oberlander and Alistair Knott. 1995. Issues in cue phrase implicature. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*.

Ted J. M. Sanders, W. P. M. Spooren, and L. G. M. Noordman. 1992. Towards a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Suzanne Stevenson and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the 9th Conference of the European Chapter of the ACL*, pages 45–52, Bergen, Norway.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.

Ian H. Witten and Eibe Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.