Exploiting Linguistic Cues to Classify Rhetorical Relations

Caroline Sporleder and **Alex Lascarides**
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
{csporled,alex}@inf.ed.ac.uk

Abstract

We propose a method for automatically identifying rhetorical relations. We use supervised machine learning but exploit cue phrases to automatically extract and label training data. Our models draw on a variety of linguistic cues to distinguish between the relations. We show that these feature-rich models outperform the previously suggested bigram models by more than 20%, at least for small training sets. Our approach is therefore better suited to deal with relations for which it is difficult to automatically label a lot of training data because they are rarely signalled by unambiguous cue phrases (e.g., CONTINUATION).

1 Introduction

Clauses in a text relate to each other via rhetorical relations such as CONTRAST, EXPLANATION or RESULT (see, e.g., (Mann & Thompson 87)). For example, (1b) relates to (1a) with RESULT:

(1) a. A train hit a car on a level crossing.
b. It derailed.

Many NLP applications would benefit from a method which automatically identifies such relations. Question-answering and information extraction systems, for instance, could use them to answer complex queries about the cause or result of an event. Rhetorical relations have also been shown to be useful for automatic text summarisation (Marcu 98).

While rhetorical relations are sometimes signalled by cue phrases (also known as *discourse connectives*) such as *but*, *since* or *consequently*, these are often ambiguous. For example, *since* can indicate either a temporal or an explanation relation (examples (2a) and (2b), respectively). Furthermore, cue phrases are often missing (as in (1) above). Hence, it is not possible to rely on cue phrases alone.

(2) a. She has worked in retail since she moved to Britain.
b. I don't believe he's here since his car isn't parked outside.

In this paper, we present a machine learning method which uses a variety of (relatively shallow) linguistic and textual features, such as word stems, part-of-speech tags or tense information, to determine the rhetorical relation between two adjacent text spans (sentences or clauses) *in the absence* of a cue phrase. We employ a supervised machine learning technique based on decision trees and boosting (Schapire & Singer 00). However, to avoid manual annotation of large amounts of training data, we train on automatically labelled examples, building on earlier work by (Marcu & Echihabi 02), who extracted examples from large text corpora and used cue phrases to label them with the correct rhetorical relation. The cue phrases were then removed before the classifiers were trained.

This approach works because there is often a certain amount of redundancy between the cue phrase and the general linguistic context. For example, the two clauses in example (3a) are in a CONTRAST relation signalled by *but*. However, this relation can also be inferred if no cue phrase is present (see (3b)).

(3) a. She doesn't make bookings but she fills notebooks with itinerary recommendations.
b. She doesn't make bookings; she fills notebooks with itinerary recommendations.

(Hobbs *et al.* 93) and (Asher & Lascarides 03) propose a *logical* approach to inferring relations, which in this case would rely on the linguistic cues of a negation in the first span, syntactic parallelism of the two spans, and the fact that they both have the same subject. We intend to explore whether such cues can also be exploited as features in a statistical model for recognising rhetorical relations.

Thus, the main difference between our research and the earlier work by (Marcu & Echihabi 02) is that their models rely on word co-occurrence statistics alone while we use a variety of linguistic features, similar to those used by (Lapata & Lascarides 04) and inspired by symbolic approaches to the task (Hobbs *et al.* 93; Corston-Oliver 98). We also use a different set of relations.

2 Related Research

(Marcu & Echihabi 02) present a machine learning approach to automatically identify four rhetorical relations (CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION and ELABORATION) from the inventory of relations described in (Mann & Thompson 87). Two types of non-relations (NO-RELATION-SAME-TEXT, NO-RELATION-DIFFERENT-TEXTS) are also included. The training data are extracted automatically from a large text corpus (around 40 million sentences) using manually constructed extraction patterns containing cue phrases which typically signal one of these relations. For example, if a sentence begins with the word *but*, it is extracted together with the immediately preceding sentence and labelled with the relation CONTRAST. Examples of non-relations are created artificially by selecting non-adjacent text spans (from the same or different texts). Because the text spans are non-adjacent and randomly selected, it is relatively unlikely that a relation holds between them. Using this method, the authors obtain between 900,000 and 4 million examples per relation.

The cue phrases were then removed from the extracted data and a set of Naive Bayes classifiers was trained to distinguish between relations on the basis of co-occurrences between pairs of lexical items. (Marcu & Echihabi 02) report a test set accuracy of 49.7% for the six-way classifier.

(Lapata & Lascarides 04) present a method for inferring temporal connectives. They, too, extract training data automatically, using connectives such as *while* or *since*. But their task differs from ours and Marcu and Echihabi's, in that they aim to predict the original temporal connective (which was removed from the test set) rather than the underlying rhetorical relation. They thus tackle connectives which are ambiguous with respect to the rhetorical relations they signal, such as *since*, and they do not address how to disambiguate them. To achieve their task, they train simple probabilistic models based on nine types of linguistically motivated features. They report accuracies of up to 70.7%.

There have also been a variety of non-statistical approaches to the problem. (Corston-Oliver 98), for instance, presents a system which takes fully syntactically analysed sentences as input and determines rhetorical relations by applying heuristics which take a variety of linguistic cues into account, such as clausal status, anaphora and deixis. (Le Thanh *et al.* 04) use heuristics based on syntactic properties and cue phrases to split sentences into discourse spans and to determine which intra-sentential spans should be related. In a second step, they then combine several cues, such as syntactic properties, cue words and semantic information (e.g. synonyms) to determine which relations hold between these spans. Finally, they derive a discourse structure for the complete text by incrementally combining sub-trees into larger textual units.

3 Our Approach

3.1 Relations and Cue Phrase Selection

We chose a subset of rhetorical relations from SDRT's inventory (Asher & Lascarides 03), namely: CONTRAST, RESULT, EXPLANATION, SUMMARY and CONTINUATION. These relations were selected on the basis that for each of them, there are *unambiguous* cue phrases but these relations also frequently occur *without* a cue phrase; so it is beneficial to be able to determine them automatically if no cue phrase is present. This is in marked contrast to relations such as CONDITION, which always require a cue phrase (e.g., *if... then* or *suppose that ...*).

SDRT relations are defined purely on the basis of truth conditional semantics and therefore tend to be less fine-grained than those used in Rhetorical Structure Theory (RST) (Mann & Thompson 87) (see below). Let $R(a,b)$ denote the fact that a relation R connects two spans a and b. For each of the five relations it holds that $R(a,b)$ is true only if the the contents of a and b are true too. In addition, **contrast(a,b)** entails that a and b have parallel syntactic structures that induce contrasting themes, **result(a,b)** entails that a causes b, **summary(a,b)** entails that a and b are semantically equivalent, **continuation(a,b)** means that a and b have a contingent, common topic and **explanation(a,b)** means that b is an answer to the question *why a?* (cf. (Bromberger 62)).

To identify mappings from cue phrases to the SDRT relations they signal, and in particular to identify unambiguous cue phrases, we undertook an extensive corpus study, using 30 randomly selected examples for each cue phrase (i.e., around 2,000 examples in all), as well as linguistic introspection given SDRT's dynamic semantic interpretation. The differences between SDRT and RST mean that some cue phrases which are ambiguous in RST are unambiguous in SDRT. For example, *in other words* can signal either SUMMARY or RESTATEMENT in RST, but SDRT does not not distinguish these relations since the length of the related spans is irrelevant to SDRT's semantics. Similarly, SDRT does not distinguish EXPLANATION and EVIDENCE, and therefore, while *because* is ambiguous in RST, it is unambiguous in SDRT, signalling only EXPLANATION. SDRT also does not distinguish CONTRAST, ANTITHESIS and CONCESSION, making *but* unambiguous.

Sentences (4) to (8) below show one automatically extracted example for each relation (cue phrases which were used for the extraction and removed before training are underlined, and the two spans are indicated by square brackets).

(4) [We can't win] [but we must keep trying.] (CONTRAST)

(5) [The ability to operate at these temperatures is advantageous,] [because the devices need less thermal insulation.] (EXPLANATION)

(6) [By the early eighteenth century in Scotland, the bulk of crops were housed in ricks,] [the barns were consequently small.] (RESULT)

(7) [The starfish is an ancient inhabitant of tropical oceans.] [In other words, the reef grew up in the presence of the starfish.] (SUMMARY)

(8) [First, only a handful of people have spent more than a few weeks in space.] [Secondly, it has been impractical or impossible to gather data beyond some blood and tissue samples.] (CONTINUATION)

3.2 Data

We used three corpora, mainly from the news domain, to extract our data set: the British National Corpus (BNC, 100 million words), the North American News Text Corpus (350 million words) and the English Gigaword Corpus (1.7 million words). We took care to remove duplicate texts. Since we were mainly interested in written texts, we also excluded all BNC files which are transcripts of speech.

Most of our corpora were not annotated with sentence boundaries, so we used a publicly available sentence splitter (Reynar & Ratnaparkhi 97), which was pre-trained on news texts, to automatically insert sentence boundaries.

The extraction happened in two steps. First, we processed the raw text corpora to extract *potential* training examples using manually written extraction patterns based on 55 (relatively unambiguous) cue phrases. All extracted examples were then parsed with the RASP parser (Carroll & Briscoe 02) and the parse trees were processed to (i) identify the two spans using simple heuristics (based on clause boundaries and the position of the cue phrases) and (ii) filter out any false positives that could not be filtered out using the raw texts alone.

An example of the latter is sentence (9), which was extracted as an example of a SUMMARY relation based on the apparent presence of the cue phrase *in short*. However, the parser correctly identified this string as part of the prepositional phrase *in short order* and the example was discarded. Examples which could not be parsed (or only partially parsed) were also discarded at this stage. For each of the extracted training examples, we also kept track of its position in the paragraph as we used this information in one of our features.

(9) In short order I was to fly with 'Deemy' on Friday morning.

Using this two step extraction method we were able to extract both intra- and inter-sentential relations (see (4) and (7) above, respectively). However, we limited the length of the extracted spans to one sentence as we specifically wanted to focus on relations between small units of text.

There are three potential sources of noise in the extraction process: (i) the two spans are not related, (ii) they are related but the wrong relation is hypothesised and (iii) the hypothesised span boundaries are wrong. The latter applies particularly to SUMMARY and RESULT, where either span can contain more than one sentence. In this case we would only extract the first (or last) sentence of the span. However, this will not cause any harm provided the partially extracted span already contains enough cues for our model to correctly learn the relation.

In our extraction method we went for high precision at the expense of recall. A small-scale evaluation using 100 randomly selected, hand-corrected examples (20 per relation) revealed 11 extraction errors overall. In no case was the wrong relation predicted. Three errors were due to hypothesising a relation where there was none. The remaining 8 errors were wrong boundary predictions (partly due to our "one sentence per span" limit, partly due to sentence-splitting errors). Hence we achieved an overall precision of 89% (97% if the less important boundary errors are excluded).

The number of training examples we could extract automatically differed for every relation: for CONTINUATION we obtained less than 2,000 examples whereas for the most frequently extracted relation, CONTRAST, we obtained around 50,000 examples. On the whole, our data set is much smaller than the one used by (Marcu & Echihabi 02), which contained around 10 million examples for six relations. Our task is thus more challenging in the sense that we are classifying rhetorical relations on the basis of a smaller training set.

3.3 Machine Learning

We used BoosTexter (Schapire & Singer 00) as our machine learning system. BoosTexter was originally developed for text categorisation. It combines a boosting algorithm with simple decision rules and allows a variety of feature types, such as nominal, numerical or text-based features. For the latter, BoosTexter applies n-gram models when forming classification hypotheses. We used BoosTexter's default settings in all experiments discussed below.

3.4 Features

We implemented a variety of linguistically motivated features (72 in total), roughly falling into 9 classes: positional features, length features, lexical features, part-of-speech features, temporal features, syntactic features and cohesion features.

Positional Features We defined three positional features. The first encodes whether the relation holds intra- or inter-sententially. The second and third encode whether the example occurs towards the beginning or end of a paragraph. The motivation for these features is that the likelihood of different relations varies with both their paragraph position and the position of sentence boundaries relative to span boundaries. For instance, CONTRAST is more likely to hold between two clauses within a sentence than CONTINUATION. And a SUMMARY relation is probably more frequent at the beginning or end of a paragraph than in the middle of it.

Length Features Information about the length of the spans might be equally useful. For example, it is possible that the average span length for CONTINUATION is longer than for CONTRAST.

Lexical Features Lexical information is also likely to provide useful cues for identifying the correct relation (cf. (Marcu & Echihabi 02)). For example, word overlap may be evidence for a SUMMARY relation. Furthermore, while we do not use cue phrases as our model features (as they provide the basis on which the data is labelled), there may be words not in our cue phrase inventory which hint at the presence of a particular relation. For instance, *still* often occurs in contrasts.

We incorporated a variety of lexical features. For each of the spans, we included the string of lemmas and stems of all words as a text-based feature. We also separately included the lemmas of all content words. Encoding lexical items as text-based features allows BoosTexter to automatically identify n-grams that may be good cues for a particular relation. Note that BoosTexter will only consider n-grams that form a continuous string. Hence bigrams in BoosTexter are different from the (non-adjacent) word-pairs used in (Marcu & Echihabi 02).

As a further feature, we calculated the overlap between the spans, i.e., what proportion of stems, lemmas, and content-word lemmas occurs in both, and added this as a numerical feature.

Part-of-Speech Features We encoded the string of part-of-speech tags for both spans as a text-based feature as it is possible that certain part-of-speech tags (e.g., certain pronouns) are more likely for some relations than for others. Following (Lapata & Lascarides 04), we also encoded specific information about the verbs, nouns and adjectives in the spans. In particular, we included the string of verb (noun, adjective) lemmas contained in each span as text-based features. For instance, the strings of verb lemmas in example (5), repeated as (10) below, are "*operate be*" (left span) and "*need*" (right span).

(10) The ability to operate at these temperatures is advantageous because the devices need less thermal insulation.

We also mapped the lemmas to their most general WordNet (Fellbaum 98) class (e.g., verb-of-cognition or verb-of-change for verbs, event or substance for nouns etc.). Ambiguous lemmas which belong to more than one class, were mapped to the class of their most frequent sense. If a lemma was not in WordNet, the lemma itself was used. Finally, we also calculated the overlaps between lemmas and between WordNet classes for each part-of-speech class and included these as numerical features.

Temporal Features Tense and aspect provide clues about temporal relations among events and may also influence the probabilities of different rhetorical relations. We therefore included temporal features in the model. To do so, we first extracted all verbal complexes from the parse trees and then used simple heuristics to classify each of them in terms of finiteness, modality, aspect, voice and negation (Lapata & Lascarides 04). For example, *need* in example (10) maps to: present, 0, imperfective, active, positive. We also introduced an additional feature where we only encoded this information for the main verbal complex in each span.

Syntactic Features It is likely that some relations (e.g., SUMMARY) have syntactically less complex spans than others (e.g., CONTINUATION). To estimate syntactic complexity we determined the number of NPs, VPs, PPs, ADJPs, and ADVPs contained in each span. Information about the argument structure of a clause may serve as another measure of syntactic complexity. We therefore encoded several aspects of argument structure as well, e.g., whether a verb has a direct or indirect object or whether it is modified by an adverbial. This information can be easily extracted from the RASP parse trees. We also included information about the subjects, i.e., their part-of-speech tags, whether they have a negative aspect (e.g. *nobody*, *nowhere*) and the WordNet classes to which they map (see above).

Cohesion Features The degree of cohesion between two spans may be another informative feature. To estimate it we looked at the distribution of pronouns and at the presence or absence of ellipses (cf. (Hutchinson 04)). For the former, we kept track of the number of first, second and third person pronouns in each span. We also used simple heuristics to identify whether either span ends in a VP ellipsis and included this information as a feature.

4 Experiments

We conducted three main experiments. First we assessed how well humans can determine rhetorical relations in the absence of cue phrases. This gives a measure of the difficulty of the task. We then determined the performance of our machine learning models and compared it to two baselines. Finally, we looked at which features are particularly useful for predicting the correct relation.

4.1 Experiment 1: Human Agreement

As we mentioned earlier, automatically extracting and labelling training data for a supervised machine learning paradigm in the way suggested in this paper and in earlier work (Marcu & Echihabi 02) relies on the existence of a certain amount of redundancy between the cue phrase and other linguistic features in signalling which rhetorical relation holds. If cue phrases were only used in cases where a relation cannot be inferred from the linguistic context alone, any approach which aims to train a classifier on automatically extracted examples from which the cue phrases have been removed would fail.

The presence of redundancy in some cases is evident from examples like (3), where CONTRAST can be inferred even when the cue phrase is removed. However, there may be other cases where this is more difficult. To assess the difficulty of determining the rhetorical relation in examples from which the cue phrase has been removed, we conducted a small pilot study with human subjects.

We used our extraction patterns to automatically extract examples for the four rhetorical relations CONTRAST, EXPLANATION, RESULT and SUMMARY (CONTINUATION was added after the pilot study). We then manually checked the extracted examples to filter out false positives and randomly selected 10 examples per relation from which we then removed the cue phrases. We also semi-automatically selected 10 examples of adjacent sentences or clauses which were not related by any of the four relations. For each example, we also included the two preceding and following sentences as context, keeping track of any paragraph markings. We then asked three subjects who were trained in discourse annotation to classify each of the 50 examples as one of the four relations or as NONE. All subjects were aware that cue phrases had been removed from the examples but did not know the location of the removed cue phrase. We evaluated the annotations against the gold standard and calculated the average accuracy. To estimate inter-annotator agreement, we also determined the Kappa coefficient (Siegel & Castellan 88). The results are shown in Table 1.

Avg. Accuracy	Kappa (pairwise, avg.)
71.25	.61

Table 1: Human performance

While the agreement is far from perfect, it is relatively high for a discourse annotation task. Hence it seems that the task of predicting the correct relation for sentences from which the cue phrase has been removed is feasible for humans. However, the accuracy was not equally high for all relations: RESULT (90%), CONTRAST (83%) and EXPLANATION (75%) seem to be relatively easy, while SUMMARY (57%) is more difficult, and the accuracy was lowest for the NONE class (50%).

Interestingly, our findings regarding the relative ease with which a given relation can be inferred if the original cue phrase is removed, deviate from those obtained by (Soria & Ferrari 98), who conducted a similar experiment for Italian. They found that "additive relations" (like SUMMARY) are easiest to infer, followed by "consequential relations" (e.g., RESULT and EXPLANATION) and "contrastive relations" (e.g., CONTRAST), which were found to be the most difficult by far. Without further research it is difficult to say where the difference between our and Soria & Ferrari's findings stem from. They could be language-specific (i.e., English vs. Italian), domain-specific (mainly news texts vs. mixed genres) or due to the different taxonomies of relations.

4.2 Experiment 2: Probabilistic Modelling

Our machine learning experiments involved five relations: CONTRAST, EXPLANATION, RESULT, SUMMARY and CONTINUATION. The automatic extraction method yielded very different amounts of training data for each of them (see section 3.2). However, machine learning from skewed data is highly problematic as it often leads to classifiers which always predict the majority class (Japkowicz 00). To avoid this problem, we decided to create uniform training (and test) sets which contained an equal number of examples for each relation. The number of examples for the least frequent relation (CONTINUATION) was 1,732 and we randomly selected the same number of examples for each of the other relations. We used 90% of this data set for training (7,795 examples) and 10% for testing (865 examples), making sure that the distribution of the relations was uniform in both data sets, and evaluated BoosTexter's performance using 10-fold cross-validation.

For comparison, we also used two baselines. For the first, a relation was predicted at random. As there are five relations and all are equally frequent in the test set, the average accuracy achieved by this strategy will be 20%. For the second baseline, we implemented a bigram model along the lines of (Marcu & Echihabi 02). Table 2 shows the average accuracies of the three classifiers for all relations and also for each individual relation.

It can be seen that our feature-rich BoosTexter model performs notably better than either of the other two classifiers. It outperforms the random baseline by nearly 40% and the bigram model by more than 20%. This difference is statistically significant (χ^2 = 208.12, DoF = 1, p <= 0.01). Furthermore, the performance gain achieved by our model holds for every relation with the exception of EXPLANATION where the bigram model performs better.

	Avg. Accuracy		
Relation	random	bigrams	BT
contrast	20.00	33.11	43.64
explanation	20.00	75.39	64.45
result	20.00	16.21	47.86
summary	20.00	19.34	48.44
continuation	20.00	25.48	83.35
all	**20.00**	**33.96**	**57.55**

Table 2: Results for BoosTexter (BT) and two baselines (10-fold cross-validation)

The comparison with the bigram model is not entirely fair as this method is geared towards large training sets. For example, (Marcu & Echihabi 02) use it on a data set of nearly 10 million examples, and their 6-way classifier achieves

49.7% compared with the 5-way classifier reported here with 33.96% accuracy. However, while it is possible that the bigram model outperforms our feature-rich BoosTexter model on large training sets, obtaining large amounts of training data is not always feasible, even if these are extracted automatically. As we have mentioned, some relations occur relatively infrequently. Others may appear more often but usually without an unambiguous cue phrase signalling the relation. In these cases even very large text corpora may not be big enough to extract sufficient training data for a bigram model to perform well. In our experiments, this case arose with the CONTINUATION relation, for which less than 2,000 examples could be extracted from a text corpus of 450 million words. For such relations, our approach seems a better choice than the bigram model proposed by (Marcu & Echihabi 02).

It is interesting that our model and the bigram model differ with respect to which relations are identified most reliably. Our model achieves the highest accuracy for CONTINUATION and the lowest for CONTRAST, while the bigram model achieves the highest accuracy for EXPLANATION and the lowest for RESULT. This suggests that it might be possible to achieve even better results by combining both models, for example, by incorporating the bigram model as a feature in our BoosTexter model.

Since our model already achieves fairly good results for the relation for which we could extract the fewest training examples (CONTINUATION), but less good results for relations for which we could extract a larger set of training examples, such as CONTRAST, it may also be possible to further improve performance by including more training data for the latter.

4.3 Experiment 3: Feature Exploration

To determine which features are particularly useful for the task, we conducted a further experiment in which we trained an individual BoosTexter model for each of our features. We then tested these one-feature classifiers on an unseen test set (again using 10-fold cross-validation) and calculated the accuracies. Table 3 shows the 10 best performing features and their average accuracies.

This suggests that lexical features (stems, words, lemmas) are the most useful features. Table 4 shows some of the words chosen by BoosTexter as being particularly predictive of a given relation. Most of the choices seem fairly intuitive. For instance, an EXPLANATION relation is often signalled by tentatively qualifying adverbs such as *perhaps* or *probably*, while SUMMARY and CONTINUATION relations frequently contain pronouns and CONTRAST can be signalled by words such as *other*, *still* or *not* etc. Of course the predictive power of such words may be to some extent domain dependent. Our examples came largely from the news domain and the situation may be slightly different for other domains.

Feature	Avg. Accuracy
left stems	42.51
left words	41.79
intra/inter	39.18
left pos-tags	34.62
right words	32.82
right stems	32.58
right pos-tags	31.72
left content words	29.78
left noun lemmas	28.30
right span length	28.12

Table 3: Best features (10-fold cross-validation)

Table 3 also suggests that the lexical items in the left span are more important than those in the right span. For example, the feature *left stems* is 10% more accurate then the feature *right stems.* This makes sense from a processing perspective: if the relation is already signalled in the left span the sentence will be easier to process than if the signalling is delayed until the right span is read.

Relation	Predictive Words
contrast	other, still, not, ...
explanation	perhaps, probably, mainly, ...
result	undoubtedly, so, indeed, ...
summary	their, this, yet ...
continuation	you, it, there ...

Table 4: Words chosen as cues for a relation

Another feature which proves very useful is *intra/inter*, which encodes whether the relation is intra- or inter-sentential. BoosTexter predicts CONTINUATION if the relation is inter-sentential and EXPLANATION otherwise. This decision rule is probably responsible for the high accuracy achieved for CONTINUATION as most CONTINUATION relations are indeed inter-sentential (though there are exceptions).

5 Conclusion

We have presented a machine learning method for automatically classifying discourse relations in the absence of cue phrases. Our method uses feature-rich models which combine a wide variety of linguistic features. We employed supervised machine learning techniques to train these models but extracted and labelled our training data automatically using predefined extraction patterns. Consequently no annotation effort is required.

We tested our method on five rhetorical relations and compared the performance of our models to that achieved by a bigram model. We found that our feature-rich models significantly outperform the simpler bigram models, at least on relatively small training sets. This means that our method is particularly suitable for relations which are rarely signalled by (unambiguous) cue phrases (e.g., CONTINUATION). In such cases, it is difficult to obtain sufficiently large training sets that a bigram model will perform well, even if the training set is obtained automatically from very large text corpora (manually constructing sufficiently large training sets is, of course, equally problematic).

In future research, we plan to conduct classification experiments with the most frequent relations to investigate whether our models are indeed outperformed by bigram models on large training sets and if so at what point this happens.

So far we have only tested our method on examples from which the cue phrases had been removed and not on examples which occur naturally without a cue phrase. However, these are exactly the types of examples at which our method is aimed. So we also intend to create a small, manually labelled, test corpus containing naturally occurring examples without cue phrases and test our method on this to determine whether our results carry over to that data type; the RST Discourse Treebank (Carlson *et al.* 02) could be used as a starting point for this (cf. (Marcu & Echihabi 02)).

Acknowledgements

The research in this paper was supported by EPSRC grant number GR/R40036/01. We would like to thank Ben Hutchinson and Mirella Lapata for interesting discussions, participating in our labelling experiment and letting us use some of their Perl scripts. We are also grateful to the reviewers for their helpful comments.

References

(Asher & Lascarides 03) Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

(Bromberger 62) Sylvain Bromberger. An approach to explanation. In Ronald J. Butler, editor, *Analytical Philosophy*, pages 75–105. Oxford University Press, 1962.

(Carlson *et al.* 02) Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. RST Discourse Treebank. Linguistic Data Consortium, 2002.

(Carroll & Briscoe 02) John Carroll and Edward Briscoe. High precision extraction of grammatical relations. In *Proceedings of COLING-02*, pages 134–140, 2002.

(Corston-Oliver 98) Simon H. Corston-Oliver. Identifying the linguistic correlates of rhetorical relations. In *Proceedings of the ACL Workshop on Discourse Relations and Discourse Markers*, pages 8–14, 1998.

(Fellbaum 98) Christiane Fellbaum, editor. *WordNet: An Electronic Database.* MIT Press, Cambridge, MA, 1998.

(Hobbs *et al.* 93) Jerry R. Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142, 1993.

(Hutchinson 04) Ben Hutchinson. Acquiring the meaning of discourse markers. In *Proceedings of ACL-04*, pages 685–692, 2004.

(Japkowicz 00) Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of IJCAI-00*, pages 111–117, 2000.

(Lapata & Lascarides 04) Mirella Lapata and Alex Lascarides. Inferring sentence-internal temporal relations. In *Proceedings of NAACL-04*, pages 153–160, 2004.

(Le Thanh *et al.* 04) Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. Generating discourse structures for written text. In *Proceedings of COLING-04*, pages 329–335, 2004.

(Mann & Thompson 87) William C. Mann and Sandra A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, ISI, Los Angeles, CA, 1987.

(Marcu & Echihabi 02) Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL-02*, pages 368–375, 2002.

(Marcu 98) Daniel Marcu. Improving summarization through rhetorical parsing tuning. In *The 6th Workshop on Very Large Corpora*, pages 206–215, 1998.

(Reynar & Ratnaparkhi 97) Jeffrey C. Reynar and Adwait Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of ANLP-97*, pages 16–19, 1997.

(Schapire & Singer 00) Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

(Siegel & Castellan 88) Sidney Siegel and N. John Castellan. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York, 1988.

(Soria & Ferrari 98) Claudia Soria and Giacomo Ferrari. Lexical marking of discourse relations – some experimental findings. In *Proceedings of the ACL-98 Workshop on Discourse Relations and Discourse Markers*, 1998.