Natural Language Engineering **14** (3): 369–416. © 2006 Cambridge University Press
doi:10.1017/S1351324906004451 Printed in the United Kingdom

Using automatically labelled examples to classify rhetorical relations: an assessment

CAROLINE SPORLEDER
ILK/Language and Information Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: csporled@uvt.nl

ALEX LASCARIDES
School of Informatics, The University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK
e-mail: alex@inf.ed.ac.uk

(*Received 26 August 2005; revised 25 August 2006; first published online 19 December 2006*)

Abstract

Being able to identify which rhetorical relations (e.g., CONTRAST or EXPLANATION) hold between spans of text is important for many natural language processing applications. Using machine learning to obtain a classifier which can distinguish between different relations typically depends on the availability of manually labelled training data, which is very time-consuming to create. However, rhetorical relations are sometimes lexically marked, i.e., signalled by discourse markers (e.g., *because*, *but*, *consequently* etc.), and it has been suggested (Marcu and Echihabi, 2002) that the presence of these cues in some examples can be exploited to label them automatically with the corresponding relation. The discourse markers are then removed and the automatically labelled data are used to train a classifier to determine relations even when no discourse marker is present (based on other linguistic cues such as word co-occurrences). In this paper, we investigate empirically how feasible this approach is. In particular, we test whether automatically labelled, lexically marked examples are really suitable training material for classifiers that are then applied to unmarked examples. Our results suggest that training on this type of data may not be such a good strategy, as models trained in this way do not seem to generalise very well to unmarked data. Furthermore, we found some evidence that this behaviour is largely independent of the classifiers used and seems to lie in the data itself (e.g., marked and unmarked examples may be too dissimilar linguistically and removing unambiguous markers in the automatic labelling process may lead to a meaning shift in the examples).

1 Introduction

To interpret a text it is necessary to understand how its sentences and clauses are semantically related to each other. In other words, one needs to know the *discourse structure* of the text. Various theories of discourse structure have been proposed, for example *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987), *Discourse Representation Theory* (DRT) (Kamp and Reyle, 1993), *Segmented Discourse Representation Theory* (SDRT) (Asher and Lascarides, 2003) and *Discourse*

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

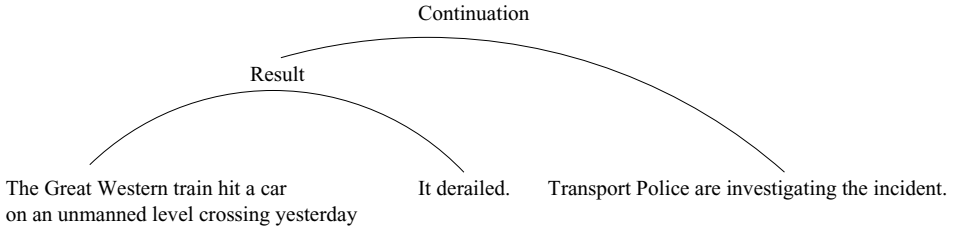


Fig. 1. Discourse Structure for Example 1.

Lexicalised Tree Adjoining Grammar (DLTAG) (Webber *et al*., 2003). Several of these theories analyse discourse as a hierarchical structure in which smaller discourse units are linked by **rhetorical relations** (also known as *discourse relations*), such as CONTRAST, EXPLANATION or RESULT, to form larger units which in turn can be arguments to discourse relations (hence the *hierarchical* discourse structure).

For instance, in SDRT (Asher and Lascarides, 2003) the logical form of discourse consists of a set of labels (which label the content of clauses or of text spans), and a mapping of those labels to logical forms. The logical forms can consist of rhetorical relations which take labels as arguments, thereby creating a hierarchical structure over the labels, and allowing rhetorical relations to relate the contents of individual clauses or of extended text spans. For instance, SDRT's logical form of the text in example (1) is shown in (1′), where we have assumed that π_1, π_2 and π_3 label the content of clauses (1a), (1b) and (1c) respectively, and for the sake of simplicity we have glossed the logical forms of these clauses as K_{π_1}, K_{π_2} and K_{π_3}:

(1) a. The high-speed Great Western train hit a car on an unmanned level crossing yesterday. [π_1]
b. It derailed. [π_2]
c. Transport Police are investigating the incident. [π_3]

(1′) $\pi_0 : \text{CONTINUATION}(\pi, \pi_3)$
$\pi : \text{RESULT}(\pi_1, \pi_2)$
$\pi_1 : K_{\pi_1}$
$\pi_2 : K_{\pi_2}$
$\pi_3 : K_{\pi_3}$

In words, this logical form stipulates that the contents of (1a) and (1b) are linked by a RESULT relation, and the label of this content is in turn the first argument to a CONTINUATION relation with the content of (1c). This analysis is represented graphically in Figure 1. The logical form for the text (1′) is assigned a precise dynamic, semantic interpretation (Asher and Lascarides, 2003). Roughly, this dynamic semantic interpretation entails (i) K_{π_1}, K_{π_2} and K_{π_3} are all true (i.e., the sentences (1a), (1b) and (1c) are all true); (ii) the eventuality in K_{π_1} (i.e., the train hitting a car) caused that in K_{π_2} (i.e., the train derailing); and (iii) the propositions labelled by π and π_3 (i.e., the contents of the text span (1a–b) and the sentence (1c)) have a contingent, common topic (e.g., both these spans are about the train accident). Furthermore, this discourse structure predicts that *the incident* in (1c) denotes the car hitting the train *and* the derailing, rather than just the derailing.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Being able to derive these rhetorical structures automatically would benefit many applications. Question-answering and information extraction systems, for example, could use them to answer complex queries about the content of a discourse which goes beyond the content of its individual clauses. For example, as we already mentioned, the RESULT relation between (1a) and (1b) entails that the train hitting the car *caused* it to derail; this causal information follows from the semantics of the discourse structure, but not from the individual sentences on their own. Rhetorical relations have also been shown to be useful for automatic text summarisation (Marcu, 1998). As a consequence, research on representing rhetorical relations and recognising them during discourse interpretation has received a lot of attention recently (see Section 2).

An important sub-task of full-scale discourse parsing is to identify which rhetorical relations hold between adjacent sentences and clauses. While rhetorical relations are sometimes signalled lexically by **discourse markers** (also known as *discourse connectives*) such as *but*, *since* or *consequently*, these are often ambiguous, either between two or more discourse relations or between discourse and non-discourse usage. For example, *since* can indicate either a temporal or an explanation relation (examples (2a) and (2b), respectively), while *yet* can signal a CONTRAST relation (example (3a)) but can also be used as a synonym of *so far* (example (3b)), with no function with respect to discourse structure whatsoever.

(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.

(3) a. Science has come to some definitive conclusions on what certain portions of the brain are used for. **Yet**, there are vast areas whose function remains a mystery.
b. While there have been plans to extend the airport, nothing has been decided **yet**.

Furthermore, discourse markers are often missing, as in (1) above where neither of the two relations was signalled by a discourse marker. In fact, roughly half the sentences in the British National Corpus lack a discourse marker entirely.

Determining the correct rhetorical relation is trivial if an example contains a discourse marker which unambiguously signals one rhetorical relation. If an example contains a (potential) marker which is ambiguous between two or more relations or between discourse and non-discourse usage, the marker has to be disambiguated.[1] Whether a given marker is considered to be ambiguous between different rhetorical relations depends to some extent on which inventory of rhetorical relations is used. Thus, different discourse theories vary with respect to which markers are considered ambiguous (see Section 3.1). The most difficult case arises if an example does not contain an explicit discourse marker, as in example (1) above. In this situation, the rhetorical relation has to be inferred solely from the linguistic context and from

[1] While there has been a lot of research on the automatic disambiguation of discourse markers, most of this involves distinguishing between discourse and non-discourse usages of a potential marker (Litman, 1996).

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

world knowledge. Hence, to determine which rhetorical relation holds between two text spans, it is not possible to rely on discourse markers alone; what is needed is a model which can classify rhetorical relations *in the absence* of an explicit discourse marker. Using supervised machine learning techniques for this task normally requires manually annotating hundreds or thousands of training examples. Given the inherent difficulty of discourse annotation, this can be very time-consuming.

As a way around this problem, Marcu and Echihabi (2002) propose a method for creating a training set automatically by labelling examples which contain an unambiguous discourse marker with the corresponding relation (i.e., the relation signalled by the marker). The discourse marker is then removed and a Naive Bayes classifier is trained on the automatically labelled data. This classifier then learns to exploit linguistic cues other than discourse markers (e.g., word co-occurrences) to determine the rhetorical relation even when no unambiguous discourse markers are present. Hence, it can, in theory, be applied to examples which naturally occur without a discourse marker. Henceforth, we will call examples that naturally occur with an unambiguous discourse marker *marked examples*, and examples that naturally occur with no discourse marker at all or with an ambiguous discourse marker *unmarked examples*.

Training on marked examples alone will work only if two conditions are fulfilled: First, there has to be a certain amount of redundancy between the discourse marker and the general linguistic context, i.e., removing the discourse marker should still leave enough residual information for the classifier to learn how to distinguish different relations. If this condition is not fulfilled a classifier which is trained on automatically labelled training data will perform badly, even if applied to unseen test data of a similar type (i.e., marked examples, from which the unambiguous discourse marker has been removed). Second, marked and unmarked examples have to be sufficiently similar that a classifier trained on the former generalises to the latter. In particular, properties which are predictive of a given relation in unmarked examples should also be predictive of the same relation in marked examples. If marked and unmarked examples vary substantially in their linguistic properties, a classifier that has been trained on automatically labelled marked examples (from which the discourse marker has been removed) may not perform very well on unmarked examples.

There is some empirical evidence that suggests that the first condition (i.e., redundancy) holds in at least some cases, as it is sometimes possible to remove the discourse marker from marked examples and still infer the relation. For instance, the two sentences in example (4a), taken from the American Newstext corpus, are in a RESULT relation signalled by the discourse marker *consequently*. But this relation can also be inferred when this discourse marker is removed (see (4b)).

(4) a. But Chater said she and Thompson "both understand the sensitivity and concern that has been expressed regarding the fact the award was paid by (Social Security). **Consequently**, based on this concern and the basic issues raised by many regarding the fairness and consistency of the award with current regulations governing their use – not to mention

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

how it might adversely affect the future payment of performance awards to federal employees – Dr. Thompson has voluntarily decided to return the entire performance award."

b. But Chater said she and Thompson "both understand the sensitivity and concern that has been expressed regarding the fact the award was paid by (Social Security). Based on this concern and the basic issues raised by many regarding the fairness and consistency of the award with current regulations governing their use – not to mention how it might adversely affect the future payment of performance awards to federal employees – Dr. Thompson has voluntarily decided to return the entire performance award."

Similarly, adding the discourse marker *consequently* at the beginning of the second sentence of example (1) yields a perfectly acceptable text, even though the RESULT relation is inferable without this marker, thanks to world knowledge and information about the lexical semantics of *train*, *hit* and *derail*.

A principle that writers produce text which is designed to minimise the extent to which readers perceive ambiguity seems plausible, and in line with Gricean maxims of conversation (Grice, 1975). However, the fact that a significant portion of sentences in any corpus lack unambiguous discourse markers is evidence that minimising perceived ambiguity is not the only principle that writers adhere to. It seems plausible to assume that they also tend to avoid unnecessary redundancy, and this would prompt them to use an ambiguous discourse marker or none at all in contexts where the content of the two spans is sufficient for the reader to infer their rhetorical connection. Thus it is unlikely that unambiguous discourse markers are completely redundant in all the examples in which they occur. Indeed, observe how removing the discourse marker *consequently* from example (5a) (which is taken from the American Newstext corpus), thereby forming (5b), results in a different rhetorical relation in interpretation: instead of RESULT one infers CONTINUATION.

(5) a. The film strip badges were replaced in 1970 with thermoluminescent dosimeters, which are crystal chips that can be read by computers and, **consequently**, are more accurate.

b. The film strip badges were replaced in 1970 with thermoluminescent dosimeters, which are crystal chips that can be read by computers and are more accurate.

It is clearly an empirical matter as to whether marked examples contain sufficient redundancy to be used as training data for a classifier. Soria and Ferrari (1998) investigated (for Italian) how far the removal of discourse markers in marked examples affects human judgements about which relation holds. While they found that the ability of their human subjects to determine the relations decreased when cue phrases were removed, the relation could still be inferred with an accuracy that was significantly above chance. These results lend support to the hypothesis that there is a certain amount of redundancy between the discourse marker and the linguistic context.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Marcu and Echihabi's (2002) research points in the same direction. Their classifier, which was trained on automatically labelled, marked examples, achieved an accuracy well above the random baseline when distinguishing four relations and two non-relations in an unseen test set of similar type (i.e., consisting of marked examples from which the cue phrases had been removed). In previous research (Sporleder and Lascarides, 2005), we obtained similar results for a different set of relations when we used automatically labelled data to train a more sophisticated classifier. Like Marcu and Echihabi, we tested our five-way classifier on marked test data from which the cue phrases had been removed.

Thus there is some empirical evidence that the first condition for building a rhetorical relations classifier from an automatically created training corpus is fulfilled: redundancy between the discourse marker and its context for inferring the rhetorical relation seems to be sufficient.

But there is currently virtually no empirical evidence for or against the second condition for building the classifier by training on unambiguously marked examples: i.e., are unambiguously marked examples sufficiently similar to unmarked examples that a classifier trained on the former generalises to the latter? We did not test our classifier on unmarked examples (Sporleder and Lascarides, 2005). And while Marcu and Echihabi (2002) did test how well their model can distinguish relations in unmarked data, their evaluation was fairly small scale. They only tested two-way classifiers distinguishing between ELABORATION and one other relation, such as CONTRAST. Furthermore, they only report the recall values for the non-ELABORATION relation. This means that it is very difficult to assess how well the model is really doing on this type of data, as a classifier which always predicts the non-ELABORATION relation would achieve 100% recall on this relation, but would not be very useful in practice.

Clearly, more empirical evidence is required to determine whether labelling training data automatically by exploiting the presence of unambiguous discourse markers in some examples is a useful strategy. In this paper, we investigate this issue in more detail. In particular we want to know whether a classifier trained on automatically labelled data performs well when applied to test data where no unambiguous discourse marker was originally present and whether there are classifier-specific differences. We compare two different models: one designed to be maximally simple and 'knowledge-lean' (Marcu and Echihabi, 2002); the other designed to be more sophisticated, both in terms of the model itself and in terms of the features used (Sporleder and Lascarides, 2005).

We also compare the performance achieved by training on automatically labelled data to that obtained by training on a small set of manually labelled examples which did not contain an unambiguous discourse marker, to determine whether there are situations in which it is better to invest development resources in manual labelling rather than in writing templates to extract and automatically label marked examples from corpora.

The next section gives an overview of related research in the area of discourse parsing and automatic identification of rhetorical relations. Section 3 describes our data and the automatic labelling process. Section 4 gives details of the two models

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

we used in the experiments. Section 5 describes our experiments; a more detailed discussion of our findings is provided in Section 6. Finally, we conclude in Section 7.

2 Related research

There is a wide range of research both on representing discourse structure and on constructing it during discourse interpretation (Polanyi, 1985; Mann and Thompson, 1987; Kamp and Reyle, 1993; Asher and Lascarides, 2003; Webber *et al*., 2003). With respect to representation, different theories vary widely on a number of issues. For example, while RST represents discourse structure as a structure over text directly (Mann and Thompson, 1987), others treat it as a syntactic structure (Polanyi, 1985; Webber *et al*., 2003) or a semantic structure (Kamp and Reyle, 1993; Asher and Lascarides, 2003). Furthermore, some theories include rhetorical relations in their representations (Mann and Thompson, 1987; Polanyi, 1985; Asher and Lascarides, 2003), while others structure the discourse by different means (e.g., by logical constants (Kamp and Reyle, 1993), or by discourse connectives (Webber *et al*., 2003)). For those theories which include an inventory of rhetorical relations, the number of relations vary because they are discriminated on the basis of different factors. The types of arguments to rhetorical relations also differ. For example, RST assumes that rhetorical relations relate text spans, and it differentiates rhetorical relations on the basis of cognitive effects and truth conditional and non-truth conditional content. In contrast, SDRT assumes that rhetorical relations relate labels of the content of token utterances – either propositions, questions or requests – and the rhetorical relations are differentiated on the basis of truth conditional content alone (so the taxonomy of SDRT relations is smaller than that for RST). The smallest units in the discourse structure also vary across different theories: RST allows units smaller than a clause to be an argument to a rhetorical relation, while in SDRT the smallest units are typically individual clauses (since arguments to relations must label propositions, questions or requests). Finally, some theories (e.g., SDRT) allow for the possibility of more than one relation holding simultaneously between two spans, while other theories (such as RST) are more restrictive and require a unique relation.

Given the goals of this paper, we attempt to remain as theory-neutral as possible about how one represents discourse structure. For example, it does not matter for our purposes what types of arguments relations take, be they textual units, syntax trees, content, or labels of content. So we simply refer to arguments to relations as *spans*. However, given that we aim to build a classifier for identifying rhetorical relations, we do need to commit to what types of relations we will classify. Following SDRT, we will assume that rhetorical relations are differentiated on the basis of truth conditional content.[2] In particular, we will focus on classifying the following five relations from SDRT: CONTRAST, RESULT, SUMMARY, CONTINUATION and EXPLANATION; see Section 3 for the motivation for using these relations, for how it compares with

[2] Indeed, although Marcu and Echihabi (2002) used relations from RST rather than SDRT, they grouped relations in the classifier in ways that ignore non-truth conditional factors.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

the relations used in Marcu and Echihabi (2002), and for further details about what they mean.

This paper addresses a sub-problem of computing discourse structure on the basis of empirical evidence; i.e., identifying the rhetorical relations which connect two spans. Our machine learning approach to computing relations contrasts with most earlier work on computing discourse structure, which employs hand-crafted rules to derive discourse analyses. Examples are Hobbs *et al.* (1993) and Asher and Lascarides (2003), who propose a logical, non-monotonic approach in which interpretations for sentences in a discourse are built in a bottom-up fashion, starting with the syntactic analysis. This method, while theoretically elegant, has drawbacks for practical applications, as it relies on deep semantic analyses of the clauses and detailed representations of domain knowledge, making it brittle in the face of naturally occurring data and the complex domain reasoning that is required.

In contrast to the semantics and inference based approach, there are several models which rely heavily on syntactic processing and/or surface linguistic cues for computing discourse structure. For example, syntax plays an important role in Polanyi *et al.*'s (2004a; 2004b) approach, together with lexical cues such as hypernym relations or modal information. Corston-Oliver (1998) also discusses a system that takes fully syntactically analysed sentences as input and then determines their discourse structure by applying heuristics which take a variety of linguistic cues into account, such as clausal status, anaphora or deixis. More recently, Forbes *et al.* (2001) propose a discourse parser which works within the Lexicalised Tree Adjoining Grammar (LTAG) framework; it uses lexicalised tree fragments which treat discourse markers as heads to derive possible discourse structures. Le Thanh *et al.* (2004) use heuristics based on syntactic properties and discourse markers 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 for smaller textual units. Marcu (1997), in contrast, presents a surface-based discourse parser that does not require a syntactic pre-processing step. But like these previous parsers, his system relies on discourse markers or word co-occurrences if no discourse markers are present. Pardo *et al.* (2004) discuss another discourse analyser which does not presuppose syntactic parsing. Their system is geared towards scientific texts in Portuguese and relies solely on surface cues and pattern matching templates.

While all of these systems are more tractable and robust than those which rely on complex inferences over full semantic representations and domain knowledge (Hobbs *et al.* (1993); Asher and Lascarides (2003)), they do still rely on the presence of certain cues in the data, such as discourse connectives, punctuation, synonyms or hypernyms. But rhetorical relations are present in the absence of such surface cues, and so this type of approach is inevitably limited in its coverage. Moreover, those systems, such as Polanyi *et al.*'s (2004a) and Forbes *et al.*'s (2001), which require full syntactic parsing of the clauses lack robustness because of lack of coverage in

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

syntactic parsing. A further problem with such systems is that the heuristics for building discourse structure tend to be hand-crafted, rather than acquired on the basis of empirical evidence from real text.

In response to these limits with symbolic approaches to discourse parsing, the last few years have seen an emergence of probabilistic models for computing discourse structure. This has been made possible by the creation of discourse-annotated corpora, such as the *RST Discourse Treebank* (RST-DT) (Carlson *et al*., 2002), which can serve as training data for these models. For example, Marcu (1999) proposes a shift-reduce discourse parser which is trained on a small, manually annotated corpus; it incorporates a decision-tree learner which exploits information sources such as part-of-speech tags, cue words, punctuation and the presence of verbs. The best model achieves around 60% recall and 63% precision on retrieving both the right rhetorical relations and the right arguments to them. More recently, Soricut and Marcu (2003) present a model which determines rhetorical structures *within* sentences; this model is trained on the RST-DT (Carlson *et al*., 2002), and the best model achieves an F-score of 49% on identifying the right rhetorical relations (among 18 possibilities) and the right arguments to those relations. Finally, Baldridge and Lascarides (2005) use a probabilistic head-driven parser, which is largely inspired by models of sentential parsing (e.g., Collins (2003)), to compute discourse structures over dialogues from the scheduling domain. Their best model achieves an F-score of 43% on identifying the right rhetorical relations (among 31 possibilities) and the right arguments to those relations.

While using supervised machine learning to train discourse parsers looks promising, annotating texts with their discourse structure is time-consuming. Annotators also typically require a lot of training to produce consistent results, due to the inherent subjectivity of the task (Carlson *et al*., 2003). Consequently, there has also been research into how manual annotation of corpora can be avoided or reduced. Nomoto and Matsumoto (1999) train a decision tree learner to identify discourse relations between sentences and investigate the use of different active learning schemes to perform intelligent sampling, selecting the examples to annotate next that will be most informative for learning.

Going one step further, Marcu and Echihabi (2002) present an approach which does not require any manual annotation effort at all. They aim to identify rhetorical relations by exploiting the fact that some examples contain discourse markers which unambiguously signals a particular rhetorical relation. The basic idea is that such examples can be extracted and labelled with the appropriate relation automatically. Their relations are chosen from the inventory of relations described in Mann and Thompson (1987), namely CONTRAST, CAUSE-EXPLANATION-EVIDENCE, CONDITION and ELABORATION (where CONTRAST and ELABORATION are supertypes for more specific relations in RST). Two types of non-relations (NO-RELATION-SAME-TEXT and NO-RELATION-DIFFERENT-TEXTS) are also included. The training data are extracted automatically from a large corpus (around 40 million sentences) using manually constructed extraction patterns containing discourse markers which typically signal one of these relations unambiguously. For example, if a sentence begins with the word

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

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, although inevitably there is some noise in this training data. Using this method, the authors obtain between 900,000 and 4 million examples per relation. The discourse markers are then removed from the extracted data and Naive Bayes classifiers are trained to distinguish between different relations on the basis of co-occurrences between pairs of words, each word in the pair coming from each span to the rhetorical relation. Marcu and Echihabi (2002) report an accuracy of 49.7% for the six-way classifier when tested on a set of automatically labelled, unambiguously marked examples from which the discourse marker was removed. They also tested several binary classifiers (distinguishing between ELABORATION and another relation) on a set of *unmarked* examples. However, they do not report the accuracy for these experiments, only the recall on the non-ELABORATION relation, which lies between and 44.74% and 69.49%.

In previous work (Sporleder and Lascarides, 2005), we took up Marcu and Echihabi's (2002) idea of labelling training data automatically. However, we used a more sophisticated model which combined simple decision lists with boosting (Schapire and Singer, 2000) and made use of a variety of linguistic features, such as span length, part-of-speech tags, syntactic and semantic information. Using this model we obtained an accuracy of 57.6% for a five-way classifier when applied to a test set of automatically labelled, marked examples from which the discourse marker was removed. These results are not directly comparable to Marcu and Echihabi's (2002) results, as we used a different set of rhetorical relations and different data. We did not test our model on unmarked examples.

Lapata and Lascarides (2004) present a similar method for inferring temporal connectives. They, too, extract training data automatically, using temporal connectives (e.g., *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. Using an ensemble of different models, Lapata and Lascarides (2004) report accuracies of up to 70.7%.

While the models of Marcu and Echihabi (2002), Lapata and Lascarides (2004) and Sporleder and Lascarides (2005) perform well on the data where there was originally a discourse marker, it is not entirely clear that their models will accurately identify rhetorical relations on examples which occur naturally without an unambiguous discourse marker. But there are many such examples in any corpus, and a useful and robust model of discourse structure should identify the correct rhetorical relations in such cases as well. The purpose of this paper is to investigate to what extent models of rhetorical relations which are trained on automatically labelled data are able to do this.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

3 Data

3.1 Relations and discourse marker selection

As in Sporleder and Lascarides (2005), we chose a subset of the rhetorical relations from SDRT's inventory of relations (Asher and Lascarides, 2003): CONTRAST, RESULT, EXPLANATION, SUMMARY and CONTINUATION. We selected these relations on the basis that (i) for each of them there are discourse markers which *unambiguously* signal them, and (ii) these relations also frequently occur *without* a discourse marker, making it beneficial to be able to determine them automatically if no cue phrase is present. An example of a relation that lacks this second property is CONDITION, which always requires a discourse marker (e.g., *if…then* or *suppose that …*).

This set of five relations roughly overlaps with, but is not identical to, the ones used in Marcu and Echihabi (2002) – we say "roughly" because we use SDRT relations rather than RST ones, and so the semantics of individual relations do not exactly match. We did not use exactly the same set of relations as Marcu and Echihabi (2002) partly because we were interested to see how well their models performed on different classes, but also because some of the relations they investigated, such as CONDITION, always require a discourse marker. As we just mentioned, we are interested in relations which often occur without such phrases, since we want to investigate whether models which train on marked examples can identify rhetorical relations in unmarked examples: that is, examples that do not contain an unambiguous discourse marker.

SDRT relations are assigned a dynamic, truth conditional semantics and therefore tend to be less fine-grained than those used in Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). Let $R(a, b)$ denote the fact that a relation R connects two spans a and b (or, more accurately for SDRT, a and b would be labels, labelling the content of two spans). Asher and Lascarides (2003) offer a detailed, dynamic semantic truth definition for the five relations that we will model. Roughly speaking, these semantic interpretations amount to the following. First, for each of the five relations we investigate, it holds that $R(a, b)$ is true only if 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 (note that it does *not* entail that the text span a is larger than b), **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?*, as defined by the semantics of *why*-questions in Bromberger (1962) and Achinstein (1980).

To create the mappings from discourse markers to the SDRT relations they signal, and in particular to identify unambiguous discourse markers, we undertook an extensive corpus study, using 10 randomly selected examples for each of 300 discourse markers listed in Oates (2000) (i.e., around 3,000 examples in all), as well as linguistic introspection given SDRT's dynamic semantic interpretation. The differences between the relations in SDRT vs. RST mean that some discourse markers which are ambiguous with respect to RST's taxonomy of rhetorical relations are unambiguous with respect to SDRT's. For example, *in other words* can signal either SUMMARY or RESTATEMENT

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

in RST but SDRT does not not distinguish these relations since the length of the related spans is irrelevant to SDRT's semantics. So in SDRT *in other words* signals SUMMARY only, with the rough truth conditions defined above. Similarly, SDRT does not distinguish EXPLANATION and EVIDENCE, and therefore, while *because* is ambiguous in RST, it is unambiguous in SDRT: if the subordinate clause is after the main clause (i.e,. the sentence is of the form *A because B*), then *because* unambiguously signals EXPLANATION; and if the subordinate clause is before the main clause (i.e., *Because A, B*), then *because* unambiguously signals RESULT, according to SDRT's definitions. SDRT also does not distinguish CONTRAST, ANTITHESIS and CONCESSION, making *but* unambiguous. From the list of discourse markers that we studied, we identified those that unambiguously signal one of the five rhetorical relations that we aim to model.

Sentences (6) to (10) below show one automatically extracted example for each relation (throughout this paper the discourse markers which were used for the extraction of the example and then removed before training are shown in bold face; spans, where relevant, are indicated by square brackets). A list of all 55 unambiguous discourse markers that we used for creating training data, together with some corresponding training examples, is given in Appendix A (Tables 9 to 13).

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

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

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

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

(10) [**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 Automatically extracted data

We used three corpora to extract training data: the British National Corpus (BNC, 100 million words), and two corpora from the news domain – the North American News Text Corpus (350 million words) and the English Gigaword Corpus (1.7 billion words).[3] We took care to remove duplicate texts, e.g., newspaper articles that occur in both the North American News Texts and the Gigaword Corpus were included only once.[4] We also removed all Wall Street Journal articles contained in the RST

[3] The same corpora were used in Sporleder and Lascarides (2005).
[4] This is relatively easy to do because the publication dates of the articles are known. So, where the same time period of the same newspaper was covered in both corpora, we only included one set.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Discourse Treebank (Carlson *et al*., 2002) as we used this treebank to obtain the manually labelled data (see Section 3.3). Since we were mainly interested in written texts, we also excluded all BNC files which are transcripts of speech. For texts which were not annotated with sentence boundaries (all from the news domain), we used a publicly available sentence splitter (Reynar and Ratnaparkhi, 1997), which was pre-trained on newstexts, to automatically insert sentence boundaries.

Like Marcu and Echihabi (2002), we extracted both intra- and inter-sentential examples (see examples (6) to (10) above). Extracting training examples comprises two steps: (i) identifying potential examples based on the presence of discourse markers and (ii) determining the span boundaries. To identify examples of our five relations, we manually wrote extraction patterns based on the 55 unambiguous discourse markers. In general, we aimed for high precision and extracted conservatively, using the linguistic context to weed out false positives wherever possible. For example, we used the phrase *in short* as one of our cues to identify the SUMMARY relation, but we only extracted examples in which *in short* occurred at the beginning of a sentence or after a punctuation mark, such as a semi-colon or a dash, as in example (11a). If *in short* occurs in other positions, as in example (11b), it frequently does not function as a discourse marker.

(11) a. Of highest priority is driving the consistent transformation of South Africa into a non-racial, non-sexist society – **in short**, the molding of a united nation.
b. Flu vaccines are currently **in short** supply.
c. **In short** order I was to fly with 'Deemy' on Friday morning.

However, even the restriction to sentence or clause initial occurrences of *in short* does not avoid false positives entirely. For example, in (11c), the phrase *in short* occurs sentence initially but it is not a discourse connective. Rather, it is part of the larger prepositional phrase *in short order*. One could avoid these examples by extracting only those sentences in which *in short* is followed by a comma. However, we found that this is overly restrictive as commas are frequently omitted after discourse markers. Another possibility would be to use a parser, as we have done in previous work (Sporleder and Lascarides, 2005). Relying on a parser is problematic though, as it can result in a significant reduction in training data, because there is always a certain percentage of sentences that cannot be parsed. We found that this lack of coverage greatly outweighs any gains obtained through better filtering of false positives, as a small proportion of false positives in the training data does not do much harm to the performance of the models whereas a significant reduction in training data due to low coverage of the parser can cause noticeable damage to the accuracy of the models. Consequently, we decided not to use a parser and instead relied on our surface-based extraction rules, even if this means that the extracted data will contain a small number of false positives. (Details about the accuracy of our extraction methods are given shortly.)

The second extraction step involves determining the spans. The task consists of (i) determining the boundary of the span containing the discourse marker and

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

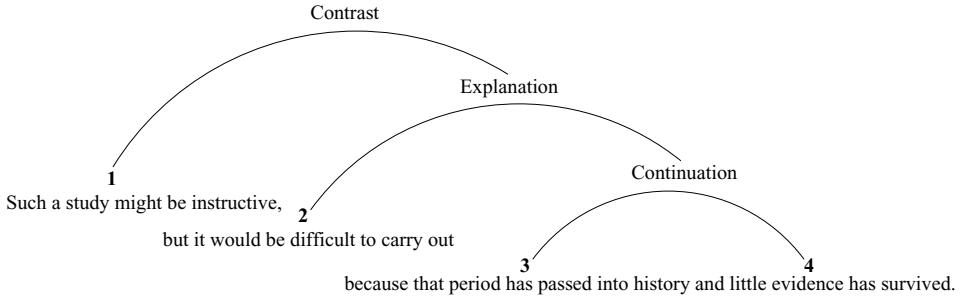


Fig. 2. Discourse tree for Example 12.

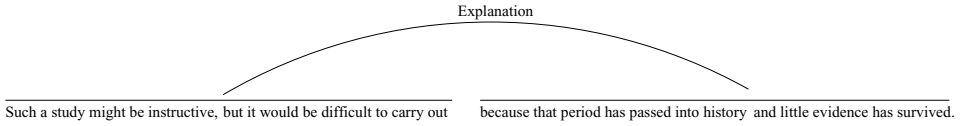


Fig. 3. Over-estimating the left span.

(ii) identifying the second span involved in the relation signalled by the discourse marker.

A sentence can contain several spans. For example, the sentence in example (12) consists of four elementary discourse units (EDUS) that can function as the spans of rhetorical relations (see the discourse tree in Figure 2): the two rightmost EDUS (3 and 4) are related by CONTINUATION,[5] the resulting, larger span is then related to the second EDU by EXPLANATION (as signalled by the discourse marker *because*) and finally the result is related to the first EDU by CONTRAST. Ideally an extraction method for the EXPLANATION relation in (12) should return EDU 2 (*but it would be difficult to carry out*) as the left span and EDUS 3 to 4 (*because that period has passed into history and little evidence has survived*) as the right span.[6]

(12) Such a study might be instructive, but it would be difficult to carry out **because** that period has passed into history and little evidence has survived.

There are two potential sources of error. First, the extension of a span can be under- or over-estimated, i.e., the wrong span boundaries are assumed. For example, an extraction method that assumed that there were two EDUS in the sentence and that the second started with the discourse marker *because* (see Figure 3) would over-estimate the extent of the first span. Second, the wrong spans can be selected, i.e., the relation does not hold between the pair of spans that are identified by the

[5] Note that we would not extract the continuation relation in this example because it is marked by *and* which is not used in the extraction templates, as this discourse marker is not unambiguous (it can signal CONTRAST, NARRATION and several other relations in addition to CONTINUATION).

[6] For the CONTRAST relation signalled by *but*, the extraction method should return EDU 1 (*Such a study might be instructive*) as the left span and 2 to 4 (*but it would be difficult to carry out because that period has passed into history and little evidence has survived*) as the right span.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

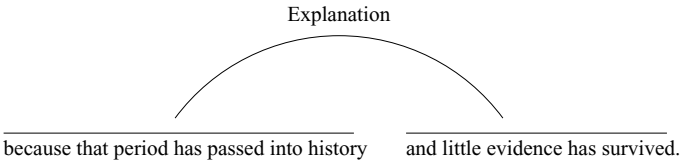


Fig. 4. Choosing the wrong spans.

model. For example, the span detection method might wrongly hypothesise that the relation signalled by *because* in (12) holds between the spans *that period has passed into history* and *and little evidence has survived* (see Figure 4).

Identifying the correct spans automatically in cases like this is not trivial and would normally involve full discourse parsing, i.e., knowledge of the discourse tree structure. For our purposes this is obviously not an option as our aim is to automatically extract examples so as to bootstrap a tool for identifying rhetorical relations, which could then be further developed into a full discourse parser. We therefore need to develop an approximate solution to the span detection problem.

One could use a syntactic parser to help with span detection. But for reasons outlined above we decided not to rely on parsing. Another possibility would be to use an external tool for segmenting discourse into spans. Soricut and Marcu (2003) and Sporleder and Lapata (2005) propose two such tools, both of which utilise supervised machine learning and were trained on the RST-DT. However, we found that the span detection heuristics outlined in the next paragraph work just as well for our purposes.

In cases where the relation is signalled by a *pair* of discourse markers, one for each span, span detection is relatively easy as the span boundaries are indicated by the discourse markers. This applies to many examples of CONTINUATION, which is often marked by a pair of cues, such as *first … second*. In all other cases we relied on the following information to determine the spans:

- the position of the discourse marker in the sentence
- linguistic background knowledge (e.g., the possible positions of spans that can potentially be related by the discourse marker)
- punctuation (i.e., we assumed that the spans of an intra-sentence relation were marked by punctuation)

We further made the following two simplifying assumptions about spans, which are also, albeit implicitly, made by Marcu and Echihabi (2002). First, we assumed that the spans involved in inter-sentential relations (e.g., in cases where *but* is sentence-initial) are *sentences*, rather than multi-sentence units or parts of sentences. This is not always true as it is conceivable that a relation (especially SUMMARY) holds between a multi-sentence span and a sentence, or even between a sentence and part of an adjacent sentence, though this latter case is rare in the RST-DT (around 5% according to Soricut and Marcu (2003)). Second, we assume that there are at most two spans in a sentence. Hence for intra-sentential relations we try to determine the border between the two spans but we do not try to identify the border between each span and a possible third or fourth span in the sentence. In the majority

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

of cases these assumptions are justified. However, in a few cases we will over- or under-estimate the extension of a span. For instance, in sentence (12), we would assume that there are two spans, with a span boundary in front of *because*; we would thus over-estimate the extent of the left span. Note, that this type of error does not necessarily pose a big problem for our machine learning approach, because an example with slightly wrong span boundaries might still contain enough cues for our models to correctly learn the relation.

To determine how reliable our extraction method is, we manually corrected a small sample of automatically extracted examples (50 examples for each relation, 250 in total). There are four potential sources of error: (i) the wrong relation is predicted, (ii) a relation is predicted where there is none, (iii) one of the hypothesised spans is wrong, and (iv) the hypothesised spans are either over or under-estimated (i.e., the span boundaries are wrong). There were five errors overall in our sample. There was no case in which a wrong relation had been hypothesised. This suggests that the discourse markers which we identified as being unambiguous do indeed unambiguously signal one relation. In one case a relation had been predicted where there was none (see example (13), the hypothesised spans are indicated by square brackets).[7] In two cases the spans were wrong. Example (14) illustrates one of these errors. The correct analysis is shown in (14a): the sentence consists of three spans and the last two are related by RESULT (since the *because*-clause precedes the clause it is related to; see Section 3.1). The example extracted by our method is shown in (14b); here the boundary between the last two spans is missed and hence it is assumed that an EXPLANATION relation holds between the first span and the rest of the sentence. The final two errors were due to over-estimating the extension of one of the spans (cf. the discussion for example (12)). The overall accuracy on the sample was thus 98%; hence our extraction method is fairly reliable.

(13) [During the nineteenth century, with improvements in communications, these traditions were modified by the] [**consequently** increased sensitivity to changing preferences in architectural styles.]

(14) a. [The cost of one wall was saved and] [**because** the buildings housing animals were close to the barn,] [they could be supplied more easily with straw.]

b. [The cost of one wall was saved and] [**because** the buildings housing animals were close to the barn, they could be supplied more easily with straw.]

Writing the templates and code for the automatic extraction and labelling process took around a week. Using this extraction method, we were able to extract 8.3 million examples in total from our three corpora. The number of training examples that we could extract automatically for different relations varied considerably, due to the different frequencies of the discourse markers that we used in the extraction

[7] The reason why we would not identify a RESULT relation in example (13) is that RESULT takes propositions as its arguments and it is not clear that the argument to *consequently* is a proposition.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 1. *Number of automatically extracted examples per relation*

	CONTRAST	EXPLANATION	RESULT	SUMMARY	CONTINUATION
examples	6,753,104	1,490,274	14,978	16,718	8,495

process. Taking all corpora into account, the least examples (around 8,500) were extracted for CONTINUATION, which is a common relation but rarely signalled by unambiguous discourse markers. The most examples were extracted for CONTRAST (nearly seven million, largely thanks to the relatively frequent discourse marker *but*). EXPLANATION was also fairly frequent, with over one million examples. For RESULT just under 15,000 examples were extracted; for SUMMARY just under 17,000 (see Table 1).

Before we trained the classifiers on the automatically extracted and labelled data, we removed from each example the unambiguous discourse markers which formed the basis for the example's extraction. Note that the reason for removing the unambiguous marker was purely technical: if we had not removed it, the classifiers would have relied on it as a predictive cue for the correct relation, at the expense of other, less predictive features. However, we wanted the classifier to learn precisely from these other features, to ensure that rhetorical relations could also be determined if no unambiguous marker was present.

Sometimes an unambiguous marker is accompanied by another, ambiguous discourse cue. For instance, in example (15) the expression *it follows that* unambiguously signals the RESULT relation. It is preceded by *so* which can also indicate RESULT, but not unambiguously; it can also signal SUMMARY, for example, as in (16). As *so* does not unambiguously signal RESULT, there is no technical reason for removing it. If it is kept, the classifier might learn that the occurrence of the word *so* at the beginning of a span is a good predictor for RESULT, but it is unlikely to rely exclusively on this cue, as the training set probably also contains examples in which *so* indicates a SUMMARY relation. Hence, the classifier also has to exploit other features. Keeping ambiguous discourse markers in the training examples thus does not impede a classifier's ability to determine relations in unmarked or not unambiguously marked examples. On the contrary, keeping ambiguous cues provides the classifiers with a valuable source of evidence for or against a given relation, in cases where such an ambiguous cue is present. For this reason, while we did remove the *unambiguous* discourse markers which were used to extract an example, we did not remove any other, *ambiguous* markers that might also have been present. Similarly, the manually labelled data which we tested our models on did not contain any unambiguous discourse markers, but some of them did contain ambiguous markers (see Sections 3.3 and 5.1.2).

(15) [These pieces are very recognizable, but there is obviously a market for them somewhere] [So **it follows that** they are going out of the country.]

(16) [John ate a fantastic salmon followed by a very nice dessert.] [So he had a great meal.]

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

3.3 Manually labelled data

The aim of this paper is to test the utility of automatically labelling training data for the task of classifying rhetorical relations. To determine how well a classifier that was trained on automatically labelled (unambiguously marked) examples performs on examples in which the rhetorical relation is not signalled unambiguously by a discourse marker, we need to label a set of those examples manually with the correct rhetorical relation.

Currently, the only publicly available English corpus that is annotated with discourse relations is the RST Discourse Treebank (RST-DT) (Carlson *et al*., 2002).[8] We could not use the RST-DT directly as it is annotated in the framework of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) whereas the classes for our models are a subset of SDRT's taxonomy of relations (Asher and Lascarides, 2003). RST and SDRT have different inventories of rhetorical relations and there is often no one-to-one mapping between the individual relations in these taxonomies (see the discussion in Section 3.1).

On the other hand, labelling new examples from scratch is a lot of work, especially since most examples will not contain any of our five relations, making it necessary to label quite a large data set to obtain a reasonably large number of examples for each of our relations. However, we could use the RST-DT as the starting point for automatically finding *potential* examples, which are then manually labelled with one of our five relations from SDRT or identified as not containing any of the relations. To this end, we first extracted from the RST-DT all examples which fulfilled the following three conditions: (i) they involved a relation between two adjacent sentences[9] or between clauses within a sentence, (ii) they were labelled with an RST relation which could potentially be mapped to one of our SDRT relations, and (iii) they did not contain any of the 55 unambiguous discourse markers we used to create the automatically labelled training set. Note that the examples were allowed to contain discourse cues which were not in our set of unambiguous markers, for example, ambiguous markers such as *so* or *and*. The motivation for allowing these was that they may also be present in the automatically labelled examples on which the models were trained. As we mentioned in the previous section, only unambiguous markers were removed from the automatically labelled data.

We identified RST relations that could potentially map to one of the five SDRT relations by using a mixture of factors: (a) a comparison of the definitions of the RST relation (Mann and Thompson, 1987) with the dynamic semantic analysis of the relations in SDRT (Asher and Lascarides, 2003); and (b) a comparison of an existing mapping from discourse markers to RST (Oates, 2000) with the mapping from discourse markers to relations in SDRT that we mentioned earlier.

Having extracted examples from the RST-DT which match conditions (i)–(iii) above, we then randomly selected around 300 examples per (hypothesised) SDRT relation for

[8] The corpus can be obtained from the Linguistic Data Consortium. Work is currently under way on a second discourse annotated corpus, the Penn Discourse TreeBank (see `http://www.cis.upenn.edu/~pdtb/`).

[9] We only classified sentences as adjacent if there was no intervening paragraph boundary.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 2. *Intra- and inter-annotator agreement for manual labelling of relations*

	Accuracy	Kappa
intra-annotator agreement	79.47%	.679
inter-annotator agreement	71.86%	.592

manual labelling. The only exception to this was SUMMARY; the potential examples we identified for this relation via conditions (i)–(iii) were rare in the RST-DT; in fact, only around 50 such potential examples were extracted and we therefore manually labelled all of them. The manual labelling was done by an independent, experienced annotator, who saw for each example the two spans (as taken from the RST-DT annotation) together with two preceding and two following sentences as context. We also retained all paragraph boundary markings. The annotator was given a set of instructions to annotate the data with one of the five relations, or with OTHER for those cases where he judged that none of the five relations held. As we mentioned above (Section 2), in SDRT it is possible that two relations hold simultaneously. However, we instructed our annotator to always assign a unique relation. Situations where multiple relations held simultaneously were resolved as follows. If one of our five relations held together with a relation not in our set, the relation in our set was assigned rather than OTHER. A situation where two of our five relations hold together can only occur with CONTINUATION, which sometimes holds together with CONTRAST, EXPLANATION or RESULT. In these cases the annotator was instructed *not* to assign CONTINUATION.

The instructions included semantic definitions for each of the five relations, together with both manually constructed and naturally occurring examples which exhibit the rhetorical relation. To avoid biasing the annotator, he did not have access to the original RST-DT relation nor to the SDRT relation that was hypothesised in the automatic mapping process.

Our final manually labelled data set contained 1,051 examples: 260 examples of CONTINUATION, 266 examples of RESULT, 44 examples of SUMMARY, 268 examples of EXPLANATION, and 213 examples of CONTRAST. The examples labelled OTHER were discarded since our models were not trained to classify this. Three examples of our manually labelled data for each relation are listed in Appendix B.

To test the reliability of the manual labelling, we computed intra- and inter-annotator agreement. For the former, we asked our original annotator to re-annotate 200 randomly chosen examples six months after the original labelling was done. To compute inter-annotator agreement, we asked a second annotator to annotate the same 200 examples. Table 2 shows the agreement figures, measured in terms of accuracy and the Kappa coefficient (Siegel and Castellan, 1988). While the figures are not very high, they are in line with the agreement figures reported in the literature for similar discourse annotation tasks. For example, Carlson *et al.* (2003) measured the agreement achieved for relation assignment while building the RST-DT and report accuracies between 60% and 79%, where the higher figure was only achieved after training the annotators for several months.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

We also kept track of how long it took our annotator to label the 1,250 examples. By coincidence, annotating the data took about the same time as writing the templates for the automatic extraction and labelling process, i.e., around a week. On top of this, we spent around one day on writing the annotation instructions and training the annotator. However, because we did not label examples from scratch but started with the RST-DT, this is an optimistic estimate of the time required for labelling examples manually.

4 Statistical modelling

To explore whether classifiers differ with respect to how much use they can make of automatically labelled training data, we implemented two different models: a relatively simple Naive Bayes model which uses word co-occurrence to hypothesise which relation holds between two spans, as proposed by Marcu and Echihabi (2002); and a more complex model which combines decision rules with boosting, as implemented in the BoosTexter system (Schapire and Singer, 2000), and uses a variety of shallow linguistic features (Sporleder and Lascarides, 2005). We introduce the two models in the following sections.

4.1 The Naive Bayes word pair model

The Naive Bayes classifier was implemented along the lines of Marcu and Echihabi (2002). This model assumes that the relation that holds between two spans can be determined on the basis of co-occurrences between words.

Let r_i be the rhetorical relation that holds between two spans W_1 and W_2. The word pair model assumes that r_i can be determined on the basis of the word pairs in the Cartesian product over the words in the two spans: $(w_i, w_j) \in W_1 \times W_2$. The model is derived as follows: Given the assumption in the word pair model, the most likely relation is given by $\underset{r_i}{\text{argmax}}\, P(r_i|W_1 \times W_2)$. According to Bayes rule:

$$P(r_i|W_1 \times W_2) = \frac{P(W_1 \times W_2|r_i)P(r_i)}{P(W_1 \times W_2)} \tag{17}$$

Since for any given example $P(W_1 \times W_2)$ is fixed, the following holds:

$$\underset{r_i}{\text{argmax}}\, P(r_i|W1 \times W_2) = \underset{r_i}{\text{argmax}}\, P(W_1 \times W_2|r_i)P(r_i) \tag{18}$$

We estimate $P(r_i)$ via maximum likelihood on the training set. And to estimate $P(W_1 \times W_2|r_i)$ we assume that all word pairs in the Cartesian product are independent. I.e.:

$$P(W_1 \times W_2|r_i) \approx \prod_{(w_i,w_j)\in W_1 \times W_2} P((w_i, w_j)|r_i) \tag{19}$$

To estimate the probability of a word pair (w_i, w_j) given a relation r_i, we use maximum likelihood estimation and Laplace smoothing. We converted all words in the spans to lower case but – to stay faithful to Marcu and Echihabi (2002) – we did not apply any other pre-processing, such as stemming.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

4.2 The multi-feature BoosTexter model

While the Naive Bayes model uses only word pairs as features, our second model employs a variety of linguistically motivated features which are retrievable from shallow processing. To combine the linguistic-based features into a classifier we used BoosTexter (Schapire and Singer, 2000). 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-valued features. Text-valued features can, for instance, encode sequences of words or parts-of-speech. BoosTexter applies *n*-gram models when forming classification hypotheses for these features (i.e., it tries to detect *n*-grams in the sequence which are good predictors for a given class label).

We implemented 41 linguistically motivated features, roughly falling into six classes: positional features, length features, lexical features, part-of-speech features, temporal features, and cohesion features. These are described (and motivated) below. Some of our features make use of word stems, word lemmas and part-of-speech tags. To obtain the word stems we applied the Porter stemmer to our data (Porter, 1980); to obtain the lemmas and part-of-speech tags, we used the RASP toolkit[10] (Minnen *et al.*, 2001).

Positional Features We implemented 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, respectively. The motivation for these features is that the likelihood of different relations may vary with both their paragraph position and the position of sentence boundaries relative to span boundaries. For instance, SUMMARY frequently holds between sentences, while EXPLANATION typically links two spans within a sentence. Similarly, 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, the average span length for some relations (e.g., EXPLANATION) may be longer than for others (e.g., SUMMARY). We therefore also encoded the lengths of the two spans in terms of the number of words contained in them.

Lexical Features Information about lexical items is also likely to provide useful cues for identifying the correct relation (cf. Marcu and Echihabi (2002)). For example, word overlap may be evidence for a SUMMARY relation. Furthermore, while we removed the unambiguous discourse marker on whose basis a given example was labelled with a particular relation, it is possible that the example contains other cue words for the relation, which can and should be exploited by the model (see Section 3.3). For instance, in the automatically labelled example (20), *but* would be removed as it was used to label the example as CONTRAST, however, *still* would be

[10] See `http://www.informatics.susx.ac.uk/research/nlp/rasp/`.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

retained. The presence of *still* is an important cue for a CONTRAST relation, though it does not unambiguously signal CONTRAST, since it can also indicate a purely temporal relation.

(20) It's a long book, **but** still very entertaining.

We incorporated a variety of lexical features. For each of the spans, we included the sequence of lemmas and stems of all words as a text-valued feature. We also included the lemmas of all content words. Including lexical items as text-based features allows BoosTexter to automatically identify n-grams that may be good cues for a particular relation. We also calculated the overlap between the spans, i.e., what proportion of stems, lemmas, and content-word lemmas occurs in both, and added these overlap figures as numerical features.

Part-of-Speech Features We encoded the sequence of part-of-speech tags for both spans as a text-valued feature. It is possible that certain part-of-speech tags (e.g., certain pronouns) are more likely for some relations than for others. Following Lapata and Lascarides (2004), we also kept specific information about the verbs, nouns and adjectives in the spans, since their models show that these linguistic cues can be highly informative for predicting temporal relations. 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 (7), repeated as (21) below, are "*operate be*" (left span) and "*need*" (right span).

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

Furthermore, to ameliorate sparse data that would arise from using only the word lemmas themselves, we mapped the lemmas to their most general WordNet (Fellbaum, 1998) class (e.g., verb-of-cognition or verb-of-change for verbs, event or substance for nouns etc.), and used those as features as well. 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 the lemmas of both spans and between WordNet classes for each part-of-speech class (of both spans) and included these as numerical features.

Temporal Features The motivation for including temporal features in the model was that tense and aspect provide clues about temporal relations among events and may also influence the probabilities of different rhetorical relations (e.g., distinguishing EXPLANATION from RESULT). We extracted simple temporal features from verbs. First, we parsed all our examples using Charniak's parser (Charniak, 2000), and extracted all verbal complexes from the parse trees. We then used simple heuristics to classify each of them in terms of finiteness, modality, aspect, voice and negation (Lapata and Lascarides, 2004). For example, *need* in example (21) maps to: present, 0, imperfective, active, affirmative.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

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 (as proposed by Hutchinson (2004) for the task of classifying discourse connectives). For the former, we encoded 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.

5 Experiments

The aim of the experiments is to test how well models that are trained on automatically labelled, originally marked data perform on data in which the rhetorical relation is not unambiguously marked. We will investigate the performance of the two models on this type of data in Section 5.1.2. However, for comparison it is also interesting to know how well the models perform on automatically labelled test data (see Section 5.1.1), i.e., data in which the relation was originally signalled by an unambiguous discourse marker but – as with the automatically labelled training data – the discourse marker was removed before testing. A good performance of the models on this type of data indicates that there is a certain amount of redundancy between the cue phrase and the general linguistic context which allows the models to learn to distinguish between relations even if the discourse marker is removed. This is one condition that has to hold to make training on automatically labelled data a useful strategy.

But, even if this condition is fulfilled, it could be that automatically labelled, unambiguously marked examples are just not similar enough to examples that are not unambiguously marked for a classifier trained on the former type of data to generalise to the latter (see the discussion in Section 1). We investigate this in Section 5.1.2, where we apply the classifiers to examples that are not unambiguously marked.

Finally, in Section 5.2, we train our classifiers on manually labelled data in which the relation was not signalled by an unambiguous discourse marker, to determine how much labelling effort would be required to obtain a performance that is similar to training on automatically labelled data.

5.1 Training on automatically labelled data

In the first set of experiments, we trained the two classifiers on the automatically extracted and labelled data. We then tested the trained classifiers on both, automatically labelled, marked data (Section 5.1.1) and manually labelled, unmarked data (Section 5.1.2).

The extracted data set is highly imbalanced with 85% of the examples being labelled as CONTRAST, 14.5% being labelled as EXPLANATION and the remaining three relations together making up just around 0.5% of the data (see Section 3.2). Learning from data that is so skewed is problematic as it will bias most learners in favour of the majority class(es). The class imbalance problem has recently gained a lot of

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

attention in the machine learning community and several different solutions have been proposed (see Chawla *et al.* (2002) for an overview). We opted for random undersampling, i.e., randomly selecting a subset of examples from the majority classes for training.

Note that the distribution of different relations in the automatically extracted data set is to some extent artificial; it does not reflect the true distributions of the relations in the corpus, as the amount of extracted examples depends only on the (relative) frequencies of unambiguous discourse markers that signal the relations. Hence the relative frequency of a relation in our data set is not a good predictor of the relative frequency of that relation in general. In particular, it is not a good predictor of the relative frequency of that relation in the set of unmarked or ambiguously marked examples, i.e., our ultimate test set. Some relations are more likely to be signalled by an unambiguous discourse marker than others. For example, while CONTRAST is frequently marked and thus very prevalent in the automatically extracted data set, CONTINUATION is frequently *not* marked by an (unambiguous) discourse marker and thus relatively rare in our data. In general, however, CONTINUATION is a very frequent relation, and the relative likelihoods of CONTINUATION vs. CONTRAST in the set of unmarked or ambiguously marked examples will therefore be quite different.

Since we do not know the true distribution of relations in the set of unmarked or ambiguously marked examples, we decided to make our training set as uniform as possible, while not discarding too many examples of the minority classes. Hence, we used the full set of examples for CONTINUATION (8,542), RESULT (14,978) and SUMMARY (16,718), and randomly selected 16,750 examples each for EXPLANATION and CONTRAST. This amounts to around 72,000 training examples overall.

BoosTexter has two parameters which need to be optimised (the number of training iterations and the maximal length of *n*-grams for text-valued features), so we set aside 250 randomly selected examples for each relation as a development set for optimisation. We found that 200 iterations and $n = 2$ works best and we kept this set-up in all following experiments.[11]

5.1.1 Testing on unambiguously marked data

First, we tested our models on a test set of lexically unambiguously marked examples from which the discourse marker had been removed (i.e., data that is similar to the automatically labelled training data). We used 10-fold cross-validation and computed the average precision (Prec.), recall (Rec.) and F-score for each relation and for all relations together, as well as the average overall accuracy (Acc.).

The results for the Naive Bayes word pair model are shown in Table 3. The overall accuracy obtained by this model is fairly low at 42.34%, though it is significantly better than the 20% average accuracy that would be obtained by randomly guessing a relation ($\chi^2 = 2258.98, DoF = 1, p <= 0.01$). Marcu and Echihabi (2002) report

[11] We were only able to try $n = 1$ and $n = 2$; higher order *n*-grams required more than the 2GB RAM we had available. This relatively high memory requirement is probably due to the fact that our model contained a fairly large number of text-valued features.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 3. *Applying the Naive Bayes word pair model to unambiguously marked data, 10-fold cross-validation*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	23.54	62.36	34.17
result	n/a	52.07	27.41	35.90
summary	n/a	56.49	32.79	41.46
explanation	n/a	47.56	71.32	57.05
contrast	n/a	50.31	26.06	34.29
all	42.34	45.99	43.99	40.57

Table 4. *Applying the BoosTexter model to unambiguously marked data, 10-fold cross-validation*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	53.37	54.90	54.11
result	n/a	56.33	47.08	51.26
summary	n/a	61.41	60.98	61.16
explanation	n/a	67.75	79.35	73.05
contrast	n/a	59.20	57.85	58.42
all	60.88	59.61	60.03	59.60

an accuracy of 49.7% for their 6-way Naive Bayes word pair classifier. The results are not entirely comparable though, as they model a different set of relations and, crucially, also train on a much larger data set of just under 10 million examples compared to our 72,000 examples. Given that the word pair model is geared towards large data sets, it is likely that its performance would increase if we trained it on more data.

Looking at the individual F-scores, it can be observed that there is some variation regarding how well the model performs for a given relation. The relatively low F-score for CONTINUATION and RESULT might be due to the fact that, even after random undersampling of the more frequent relations EXPLANATION and CONTRAST, these are (still) the least frequent relations in our training set. However, the performance on CONTRAST is also fairly low. We experimented with adding additional training examples for CONTRAST but found that this leads to a degradation in overall performance of the model: while the F-score for CONTRAST goes up slightly if more examples are added, this is more than outweighed by the fact that the F-scores for the other relations go down.

Table 4 shows the results of applying the BoosTexter model. It can be seen that this model generally performs better than the word pair model. This difference in overall accuracy is significant ($\chi^2 = 1019.82, DoF = 1, p <= 0.01$). Hence, for relatively small data sets, this model seems to work better than the word pair model.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 5. *Applying the Naive Bayes word pair model to data that is not unambiguously marked, averaged over 10 training runs*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	26.62	62.85	37.40
result	n/a	24.87	8.12	12.24
summary	n/a	5.47	8.41	6.63
explanation	n/a	31.55	25.15	27.97
contrast	n/a	23.40	7.65	11.53
all	25.92	22.38	22.44	19.15

As with the word pair model, EXPLANATION and SUMMARY are predicted most reliably, the other three relations seem to be slightly more difficult.

As both models perform significantly better than the random baseline on the test set, it seems that there is a certain amount of redundancy between the discourse marker and the general linguistic context in (unambiguously) marked examples. Hence, rhetorical relations are learnable – at least to some extent – from automatically labelled data. In the next section, we investigate whether these results carry over to unmarked or ambiguously marked test data.

5.1.2 Testing on Unmarked data

Testing the models on unambiguously marked data from which the discourse markers had been removed led to reasonably good results. But we are interested in whether a labeller that is trained on automatically labelled data of this type can also identify rhetorical relations in examples that are not unambiguously marked. In other words, do the semantic clues for rhetorical relations that were learnt by the classifier from unambiguously *marked* examples approximate those that one needs to determine the rhetorical relations in *unmarked* (or ambiguously marked) examples?

As in the previous experiments, the models were trained on the automatically labelled data. We could not employ proper cross-validation due to the need of keeping the test and training data disjoint (due to the fact that the former were manually labelled and the latter were automatically labelled). In order to still be able to generalise away from the potential idiosyncrasies of any given training set, we split the training data randomly into 10 equal sized, disjoint subsets and trained 10 different models on it, each time leaving out one subset and using the remaining nine for training, thus ensuring that the models were trained on the same amount of data as those in the previous experiments. Each of the ten models was then applied to the complete set of manually labelled data (1,051 instances) and the results of the 10 models were averaged.

Table 5 shows the results for the Naive Bayes word pair model. It can be seen that the performance drops quite markedly when compared to testing on data that was originally unambiguously marked: the accuracy drops by more than 38% (from 42.34% to 25.92%) and the F-score falls by more than 52% (from 40.57% to

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 6. *Applying the BoosTexter model to unmarked data, averaged over 10 training runs*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	36.70	20.35	26.17
result	n/a	25.08	19.74	22.08
summary	n/a	9.32	45.91	15.49
explanation	n/a	37.51	37.13	37.30
contrast	n/a	21.38	21.60	21.47
all	25.80	26.00	28.94	24.50

19.15%). The decrease in performance is particularly noticeable for SUMMARY (from 41.46% F-score to 6.63%). However, in spite of the big drop in performance, the model still performs significantly above the 20% baseline of choosing one of the five relations randomly ($\chi^2 = 18.88, DoF = 1, p <= 0.01$).

Table 6 shows the results of applying the BoosTexter model. While this model outperformed the Naive Bayes word pair model when tested on examples where the relation was originally unambiguously marked by a discourse marker, it drops to a performance level that is similar to that of the word pair model when tested on naturally unmarked or ambiguously marked data. Because BoosTexter starts from a higher level, the fall in performance is even more pronounced: both accuracy and F-score decrease by around 58% (from 60.88% to 25.80% for accuracy and from 59.60% to 24.50% for F-score). As with the word pair model, the decrease is particularly stark for the SUMMARY relation, where the F-score drops from 61.16% to 15.49%, i.e., by more than 74%. However, unlike the word pair models, the BoosTexter models obtain a relatively high recall for SUMMARY. This might indicate that the bad performance for this relation has something to do with the difference in its relative frequency in the training and test sets; the distribution of relations is relatively uniform in the training set by design, whereas in the test set, SUMMARY is relatively rare (44 instances out of 1051, see Section 3.3). The uniform distribution in the training data leads BoosTexter to predict all relations with more or less the same frequency, which for SUMMARY results in a high recall but low precision.

Overall the decline in performance when testing on unmarked or ambiguously marked data shows that our models, which were trained on examples where an unambiguous discourse marker was originally present, do not generalise well to examples which occur naturally without unambiguous discourse markers. However, the performance is still higher than the baseline of selecting a relation randomly.

The fact that both models show a significant drop in performance to just above chance suggests that this behaviour may be largely independent of the type of classifier used (i.e. knowledge-lean vs. complex). Instead it looks like the problem might stem from the data itself, i.e., marked and unmarked data could just be too dissimilar linguistically to allow a classifier to generalise from one to the other. We investigate this further in Section 6. In the next section, however, we explore how the performance achieved by training on automatically labelled data compares to that

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 7. *Training and testing on manually labelled data, Naive Bayes word pair model, 5 times 2-fold cross-validation*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	27.27	12.00	16.48
result	n/a	27.65	9.70	13.41
summary	n/a	2.44	29.09	4.50
explanation	n/a	29.85	5.97	9.89
contrast	n/a	19.43	23.28	20.54
all	12.88	21.33	16.01	12.96

achieved by training on manually labelled, not unambiguously marked examples. In particular, we want to know how many examples one would have to annotate to obtain a performance similar to training on automatically labelled data.

5.2 Training on manually labelled (unmarked) data

The set of manually labelled data consists of 1,051 examples (see Section 3.3). To train and then test the models, we split this set in half, ensuring that each half contained similar proportions of each rhetorical relation, and then trained the models on one half and tested on the other. We then swapped the training and test set and repeated the procedure five times, each time with a different split of the data set (i.e., effectively using five times 2-fold cross-validation).

Table 7 shows the results of applying the Naive Bayes word pair model. The average accuracy is 12.88% and the average F-score is 12.96%. This is noticeably lower than the performance achieved by training on automatically labelled examples and testing on manually labelled ones (25.92% accuracy and 19.15% F-score). The difference in accuracy is significant ($\chi^2 = 45.83, DoF = 1, p <= 0.01$). Furthermore, randomly guessing a relation would actually lead to a higher accuracy (i.e., 20%) on average. Again this difference is significant ($\chi^2 = 16.19, DoF = 1, p <= 0.01$). The bad performance of the word pair model when trained on the manually labelled data set can be explained by the fact that this model requires a large training set to achieve reasonable performance levels. This explains why the model's performance when trained on over 72,000 automatically labelled examples is much better than when trained on 526 manually labelled examples. Indeed, the need for a large data set appears to outweigh the disadvantages of using a different type of data for training vs. testing; i.e., it outweighs any linguistic differences that there might be between the automatically labelled examples and the manually labelled ones, even though our prior experiments suggest that such differences are real and affecting performance.

Table 8 shows the results of training and testing the BoosTexter model on manually labelled examples. For this model the situation is different: training on a small set of manually labelled examples leads to a 14.50% higher accuracy and a 9.19% higher F-score than training on a large set of automatically labelled examples. The

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 8. *Training and testing on manually labelled data, BoosTexter model, 5 times 2-fold cross-validation*

Relation	Avg. Acc	Avg. Prec	Avg. Rec	Avg. F-Score
continuation	n/a	36.78	36.85	36.77
result	n/a	38.53	46.32	41.99
summary	n/a	13.75	3.64	5.63
explanation	n/a	49.80	50.15	49.85
contrast	n/a	36.70	32.21	34.19
all	40.30	35.11	33.83	33.69

difference in accuracy is significant ($\chi^2 = 57.05, DoF = 1, p <= 0.01$). The results might be slightly optimistic as the manually labelled examples all come from the set of 385 Wall Street Journal articles which make up the RST-DT (Carlson *et al.*, 2002), whereas the automatically labelled examples come from a variety of sources, though mainly in the news domain.[12] However, we only used a fraction of the examples in the RST-DT for labelling and we randomised them beforehand (see Section 3.3), so one would expect a fairly wide spread across the 385 articles.[13] Furthermore, the difference in performance between training on automatically extracted examples and training on manually labelled examples is so large that it can hardly be explained by the relative homogeneity of our manually labelled examples alone. Hence, for this model, it looks like training on even a relatively small set of manually labelled examples leads to a significantly better performance than training on a large set of automatically extracted examples. This is in stark contrast to the word pair model, where having a large training set outweighs any reductions in performance that are caused by training and then testing on different types of examples.

Given that training on even a small set of manually labelled examples leads to significant improvements over training on automatically labelled data alone, we wanted to know how little manually labelled data are necessary to beat training on automatically labelled data. Therefore, we conducted a learning curve experiment for training on manually labelled data. We only tested the BoosTexter model as the

[12] Under ideal circumstances, one would want both data sets, the manually labelled data and the automatically labelled data, to come from similar corpora or even the same corpus. This was infeasible, however. Automatic extraction of training data requires an extremely big extraction corpus – in our case a concatenation of three large corpora – otherwise there will not be enough unambiguously marked relations to extract a sufficient number of training examples. For example, if we had used the RST-DT to extract unambiguously marked examples, we would only have extracted one, two and three training examples for CONTINUATION, SUMMARY, and RESULT, respectively. On the other hand, for the manually labelled data we had to start with an existing discourse annotated corpus to quickly identify *potential* examples of our five relations (see Section 3.3). Hence, we had to use the RST-DT to obtain manually labelled examples. To keep the data sets still fairly comparable, we made sure that the examples came mostly from the same domain, i.e., news texts.

[13] For SUMMARY we used all available examples, because this relation is so rare (only 44 cases in the RST-DT) but, here too, we would expect a fairly even spread across the articles, i.e., it is very unlikely that the examples all came from only a small set of articles.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

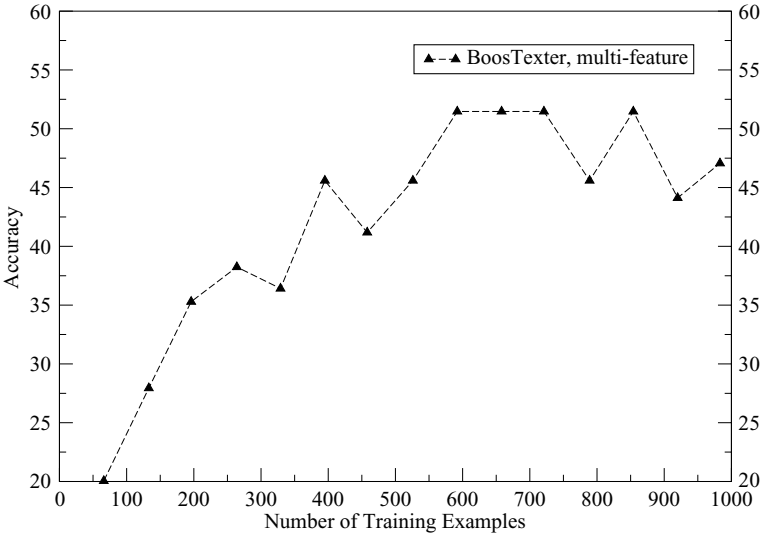


Fig. 5. Learning curve for training and testing on manually labelled, unmarked data.

Naive Bayes word pair model was found not to perform very well on small amounts of training data. For the learning curve, we randomly split the manually labelled data into 16 subsets of similar size and with similar proportions of examples per relation. One of these was set aside as a test set and the remaining 15 were used to create 15 progressively increasing training sets, with the smallest containing just one subset and the largest containing all 15. We then trained a model on each of the training sets and tested it on the test set.

Figure 5 shows the results. While there is some variation for individual training sets (due to the fact that they were randomly selected), it can be seen that all training sets, except for the smallest, lead to an accuracy of above 25%, i.e., the accuracy achieved by training a BoosTexter model on automatically labelled data. In other words, around 140 manually labelled examples are enough to beat training on automatically labelled data. This estimate might again be slightly optimistic due to the relative homogeneity of our manually labelled data. However, given that automatic data labelling is not entirely cost-neutral either, i.e., some human effort is required for writing the extraction rules, it seems that manually labelling some examples may sometimes be a better option than training on automatically labelled examples alone. In our case, writing the extraction rules took around one week, while our annotator labelled more than 220 examples per day. However, this is an optimistic estimate of the time involved in manually labelling examples, as we did not start from scratch but used the RST-DT to pre-select examples. Also, large scale labelling would also require some time for training the annotators and, because rhetorical relations are not distributed uniformly, it might be necessary to label quite a large amount of data to ensure that there are sufficient training examples for the rarer relations.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

6 Discussion

The fact that models which have been trained on automatically labelled examples perform well on a test set consisting of similar data (i.e., unambiguously marked examples from which the markers have been removed) demonstrates that spans linked by unambiguous discourse markers tend to provide sufficient evidence *on their own* to identify the rhetorical relation between them. In other words, there is a certain amount of semantic redundancy between an unambiguous discourse marker and its context, at least as far as inferring the rhetorical relation is concerned. On the other hand, the above experiments also suggest that classifiers trained on automatically labelled examples are not much use for determining which rhetorical relations hold in examples that are not unambiguously marked.

While one cannot formally prove that there is no model which, when trained on automatically labelled marked examples, can successfully identify rhetorical relations in unmarked examples, these experiments are quite suggestive, as they represent two very different models and sets of features (knowledge-lean vs. knowledge-rich), and the same behaviour (relatively good performance on marked examples, performance just above chance on unmarked examples) was observed for both of them. This provides circumstantial evidence that the problem may lie with the data not with the models. In fact, some error analysis reveals potential reasons why we got the results we did, and these reasons would apply to any model. We discuss these here.

First, the poor performance could be due to labelling errors in the automatically extracted data. We did a small scale evaluation of our extraction method in Section 3.2 and found it to be fairly reliable (with an overall accuracy of around 98%). However, a closer inspection of a new sample of 50 randomly selected automatically labelled examples for each relation revealed that a small proportion (around 15%) of examples labelled as SUMMARY seemed to be false positives. An example is given in (22):

(22) a. Along with its size, the Portfolio's claimed PC compatibility is its most eye-catching feature.
b. However, the machine has a basic memory of only 128K, a not-very-standard keyboard and screen, no disk drive, and an operating system called DOS, which stands for DIP Operating System, after the machine's British designers.
c. **In short**, it seems inevitable that most PC software will need some kind of re-write to run on the Portfolio, and some of it will never fit at all.
d. It's true that Atari is offering a card drive for desktop PCs, and battery-backed add-in memory cards, but these cost over 1 per kilobyte.

This example would be labelled RESULT by an annotator; observe that one can replace *in short* with a discourse marker like *therefore* and preserve the meaning of the discourse overall. Indeed, the semantics of RESULT better reflects the intuitive interpretation of this discourse than SUMMARY does, since RESULT entails that (22b) causes (22c), and SUMMARY does not entail this information.

The source of the error for this example lies in our decision that the cue phrase *in short* unambiguously signals the relation SUMMARY. The example above shows

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

that it sometimes signals RESULT rather than SUMMARY. This error was made in spite of using corpus evidence and linguistic introspection to decide which cue phrases signalled which rhetorical relations, as discussed in Section 3.1. The wrong classification of *in short* is the only error of this type we found among the examples we inspected, however. This suggests that there are still relatively few false positives, but they may have some effect on the classifier's ability to determine relations in unmarked examples, e.g. making it more likely that a RESULT relation is misclassified as SUMMARY. SUMMARY is indeed the relation which is recognised least reliably by both classifiers in the unmarked test set, with an F-score of 6.63% (Naive Bayes) and 15.49% (BoosTexter) compared to an average F-score for all relations of 19.15% and 24.50%. However, incorrect labelling of the marked examples cannot be the whole story, because the classifiers do not perform particularly well on the relations for which we found no incorrect labels. It therefore seems likely that there are other reasons why the classifiers do not generalise very well to examples that are not unambiguously marked.

One possibility could be that unambiguously marked examples are simply not representative of unmarked (or ambiguously marked) examples. For example, it could be that there are features which are highly predictive of a particular relation in the automatically labelled training data but which are not as predictive in the unmarked or ambiguously marked test data (due to syntactic and semantic differences between these two types of examples). A model which puts high confidence in such a feature at the expense of other features will perform well on marked data but less well on unmarked data (which is exactly the behaviour we observe with the two classifiers). Conversely, features which are highly informative of a given relation in the unmarked data may not register at all during training on automatically labelled data.

There is some evidence that this may indeed play some role. For example, the BoosTexter model which was trained on automatically labelled data contains a decision rule which assigns CONTINUATION to inter-sentential relations (and EXPLANATION otherwise). This rule makes sense for the automatically labelled, marked data, because CONTINUATION relations tend to be inter-sentential if they are explicitly marked by an unambiguous discourse marker, such as *first … second* (see example (23)).[14] However, CONTINUATION relations that are not unambiguously marked are just as likely to hold between two clauses within a sentence, as in example (24) (observe the ambiguous discourse marker *and* in this example). This means that the accuracy of the decision rule decreases when it is applied to examples that are not unambiguously marked, and including this rule might actually harm the model's performance on this type of data.

(23) [**First**, patients have to distinguish between a doctor's professional fees and the private hospital charges.] [**Second**, the patient should always be provided with an itemised bill to enable him to know what he has been charged for.]

[14] In the few cases where CONTINUATION occurred intra-sententially in the automatically labelled data, the two clauses were mostly separated by a semi-colon rather than a comma.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

(24) [A unit of DPC Acquisition Partners launched a $10-a-share tender offer for the shares outstanding of Dataproducts Corp.,] [and said it would seek to liquidate the computer-printer maker "as soon as possible" even if a merger isn't consummated.]

It is impossible to quantify the extent to which this kind of phenomenon is hurting performance because one cannot interpret the decision rules in a BoosTexter model in a piecemeal fashion. Nevertheless, it is clearly the case that features such as span length and part-of-speech tags may well be sufficiently different across the two types of data to contribute to the poor performance of the model on data that is not unambiguously marked. For example, we found a significant[15] difference in average span length for the RESULT relation, with the spans in the automatically labelled examples being on average six tokens longer for the left span and four tokens longer for the right span. For CONTRAST, EXPLANATION and RESULT, we also found differences in part-of-speech tags. For example, conjunctions are more likely at the beginning of the right span in the manually labelled examples than they are in the automatically labelled ones. Typically, these conjunctions are ambiguous discourse markers, such as *and*. While these can also occur in the automatically labelled examples (see Section 3.3), they occur less frequently because the automatically labelled examples already contain an unambiguous discourse marker. Furthermore, personal pronouns are very frequent at the beginning of the right span of EXPLANATION in the automatically labelled examples. In 44% of the examples, the right span started with a personal pronoun. However, in the manually labelled examples, personal pronouns are rare at the beginning of the right span and occur in only 4% of the cases. We believe that this may be due to the fact that in the unambiguously marked examples EXPLANATION tends to be intra-sentential, which makes the use of pronouns in the second span more likely. In the not unambiguously marked examples, on the other hand, EXPLANATION is frequently inter-sentential.

Another factor which may play a role in the poor performance of the models is, that removing an unambiguous discourse marker may change the meaning of the training example, and yet this change in meaning is not registered in the automatically generated label. This is a distinct factor from having false positives in the training data, but potentially just as damaging for the model. This is because, if such a shift in meaning takes place when preparing an unambiguously marked example for training, then the labelled example consists of spans whose interpretation signals a different rhetorical relation from the one with which the example is labelled. This 'mismatch' between interpretation and the label is not normally present in the manually labelled, unmarked (or ambiguously marked) examples (provided these are labelled correctly), because these examples are not modified during the labelling process (i.e., no discourse markers are removed) and so their meanings do not shift. Thus the correlation between the label and the meanings of the spans may be sufficiently different for the marked vs. unmarked examples that the model performs poorly when trained on one and tested on the other.

[15] Wilcoxon two sample test, $p < 0.01$.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

There is in fact some evidence that this phenomenon of 'meaning shift' occurs in the training set of marked examples in two ways. Firstly, removing an unambiguous discourse marker can lead to a discourse which sounds anomalous or incoherent. For example, consider (25), taken from our training set of marked examples labelled with CONTRAST:

(25) Manually adding best-fit curves to data plots can be laborious and prone to error. **But** Don Bradbury reviews TableCurve and finds it has all the right lines.

Observe how removing the discourse marker *but*, which happens as part of the preparation of the training set, results in a text which sounds quite odd. The relation CONTRAST is supported in the interpretation of (25) on the grounds that the truth of the first span would normally lead one to infer that the second span is false; i.e., this is an example of CONTRAST that is grounded in violation of expectation. Asher and Lascarides (2003) suggest that this type of CONTRAST requires the relation to be linguistically signalled in an explicit way, with a (perhaps ambiguous) discourse marker or with intonation. Thus "*John likes sport but he hates football*" is acceptable, whereas "*John likes sport. He hates football*" sounds odd (unless it is spoken with a marked intonation). Example (25) also supports their claim.

On the other hand, for CONTRAST examples where there are no logical relationships among the truth conditions of the spans, but rather the CONTRAST is present simply because the difference in content between two spans is salient, no discourse markers are necessary: e.g., "*John has green eyes. Mary has blue eyes.*" Moreover, if other (perhaps ambiguous) markers had been present to signal CONTRAST in examples like (25), then the incoherence of removing the unambiguous marker would have been ameliorated. If we add *then again*, which can signal CONTRAST, PARALLEL or NARRATION among others, to (25) to form (26), then removing *but* (as shown in (27)) produces an acceptable text with no overall change in meaning:

(26) Manually adding best-fit curves to data plots can be laborious and prone to error. But then again, Don Bradbury reviews TableCurve and still finds it has all the right lines.

(27) Manually adding best-fit curves to data plots can be laborious and prone to error. Then again, Don Bradbury reviews TableCurve and still finds it has all the right lines.

These examples show that in certain circumstances, removing discourse markers can result in an incoherent text, and the extent to which this occurs in our training data is an empirical matter. We inspected 200 marked examples of CONTRAST, and found that 30% of them sounded incoherent when the unambiguous discourse marker was removed (and syntax modified to make the spans grammatical).[16] Consequently, it might be the case that the model correlates the relation CONTRAST with *discourse incoherence*. This may make it difficult to identify CONTRAST in examples which are

[16] A similar inspection of 100 examples featuring the other rhetorical relations revealed that 3% of them became incoherent when the discourse marker was removed.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

not unambiguously marked, since they are all coherent; a violation of expectation is in these cases typically signalled in a linguistically explicit way with an *ambiguous* discourse marker such as *then again*, or *while*. For example:

(28) [**While** some analysts say the dollar eventually could test support at 1.75 marks and 135 yen,] [Mr. Scalfaro and others don't see the currency decisively sliding under support at 1.80 marks and 140 yen soon.]

The second way in which the method for automatically labelling unambiguously marked examples results in a shift in their meaning, occurs for examples where in fact *two* rhetorical relations hold between the spans simultaneously, and removing the unambiguous discourse marker removes one of these relations in the interpretation. Note that allowing more than one relation to hold between two spans is supported in SDRT (see Section 2), so as to reflect the fact that a given utterance can make more than one illocutionary contribution to the discourse. This contrasts with RST, which assumes that there is at most a unique relation between two given spans, but Moore and Pollack (1992) argue persuasively that this uniqueness assumption is problematic. For our set of relations, this plurality of relations holding can happen with CONTINUATION, which can hold together with CONTRAST, EXPLANATION, or RESULT.[17]

For instance, it can be argued that in example (29), both CONTRAST and CONTINUATION hold.

(29) **Although** the electronics industry has changed greatly, possibly the greatest change is that very little component level manufacture is done in this country.

The cue phrase *although* in (29) (unambiguously) indicates CONTRAST (and it is the presence of this discourse marker that allows us to classify this example as an unambiguously marked example of CONTRAST), but there is also convincing evidence that the spans are related by CONTINUATION as well. The first piece of evidence comes from the semantic definitions of the relations themselves, and how they capture the intuitive interpretation of (29). In particular, the semantics of CONTINUATION given in (Asher and Lascarides, 2003) entails that the definite description *the greatest change* is one of the changes in the electronics industry and that the eventuality in the second span is temporally included in the one for the first span. This temporal information is *not* a logical consequence of CONTRAST, but it is part of the intuitive interpretation of (29).

Further evidence that CONTINUATION is part of the interpretation of (29) comes from observing that this relation still holds when *although* is removed; see (30).

(30) The electronics industry has changed greatly. Possibly the greatest change is that very little component level manufacture is done in this country.

[17] Note that this *plurality* of relations is different from *ambiguity*. Plurality of relations means that relations are actually holding simultaneously.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

In fact, removing the cue phrase renders CONTINUATION the dominant relation. However, our method for automatically extracting labelled training data means that (30) would be labelled as CONTRAST and not as CONTINUATION, as CONTRAST is the relation signalled by the (now removed) cue phrase *although*. This could potentially be a big problem for classifiers that are trained on automatically labelled data, though all the marked examples that we inspected for RESULT and EXPLANATION were ones where removing the cue phrase did not affect interpretation – RESULT and EXPLANATION were preserved as part of their meaning.

There is one further element that might make training on unambiguously marked examples and testing on unmarked or ambiguously marked ones problematic. We have already observed how removing a discourse marker can change the meaning of a text. This leads to training examples where the intuitive interpretation of the spans (with the marker removed) does not match its label. The converse situation is also problematic for the design of the model: *adding* an unambiguous discourse marker to an *unmarked* example can result in a perfectly coherent text, but with a different meaning from the original form. For example, one could conceivably add *because* to an unmarked example of CONTINUATION, and in doing so produce a perfectly acceptable text whose interpretation is one of EXPLANATION. This could prove problematic, because, given the way the training data of unambiguously marked examples was created, our models are essentially performing the task of assessing which unambiguous discourse marker was removed.

To see to what extent the set of unmarked examples has this feature, we manually inspected 100 examples from the unmarked test set which were erroneously labelled with the Boostexter model. For each of these examples, we added an unambiguous discourse marker which signals the relation that the model (incorrectly) predicted, to assess whether the resulting text was acceptable and signalled that relation. We found that for 7 of the 100 examples, this was the case. For example, observe the difference in meaning between the original unmarked example (31a) (which was labelled CONTINUATION but incorrectly predicted to be RESULT), and the meaning of the perfectly acceptable (31b), where the discourse marker *consequently*, which unambiguously signals RESULT, has been added:

(31) a. [Among those companies expected to have a down quarter are Hewlett-Packard Co., Amdahl Corp. and Sun Microsystems Inc., generally solid performers in the past.]
[International Business Machines Corp. also is expected to report disappointing results.]

b. [Among those companies expected to have a down quarter are Hewlett-Packard Co., Amdahl Corp. and Sun Microsystems Inc., generally solid performers in the past.]
[**Consequently**, International Business Machines Corp. also is expected to report disappointing results.]

A further factor that might contribute to poor performance of the model when it is tested on examples that are not unambiguously marked is the quality of the labels on the manually labelled test examples. We have observed that intra- and

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

inter-annotator agreement scores show the task of identifying rhetorical relations to be relatively well-defined; indeed, the fact that training and testing on manually labelled data lead to results that were significantly above the baseline suggests that the manually labelled data has been labelled in a relatively consistent manner. However, the annotator agreement scores also indicate the difficulty and subjectivity of the task, and manually labelling a data set with rhetorical relations is inevitably error prone.

To see to what extent this is a problem, two annotators, who were different from the annotator that prepared the set of unmarked examples, inspected 100 unmarked examples. They were allowed to confer about the labels, and assess together if the original label was correct. They agreed that 11% of them were labelled incorrectly. For example, (32) was erroneously marked as SUMMARY rather than EXPLANATION, and (33) was erroneously labelled CONTINUATION instead of CONTRAST:

(32) ["It will be very expensive'," the spokesman warned.] ["The price cannot be less than $7,000".]

(33) [While lawyers arranged individual tie-ups before,] [the formal network of court reporters should make things easier and cheaper.]

We have now identified four factors which potentially hurt the performance of a model which is trained on unambiguously marked examples and tested on unmarked or ambiguously marked ones. First, the marked examples may not be representative of unmarked ones, at least with respect to the features we have investigated here. Secondly, the training data can be 'faulty' in that the intuitive interpretation of the example with its unambiguous discourse marker removed does not match the rhetorical relation with which it is labelled. This can be caused by false positives, or by the very act of removing the discourse marker, which is a necessary step in preparing the training set. Thirdly, there is noise in the manually labelled test examples as a consequence of incorrect manual labels. And finally, the model can also misclassify unmarked examples if adding an unambiguous discourse marker to it produces an acceptable text, but with a changed meaning.

Since BoosTexter models defy any precise interpretation or a transparent link between the model and the influence of different features on the decision rules that are constructed, it is impossible to quantify the extent to which these four factors contribute to poor performance on the task; nor can we assess which of these four factors is dominant. However, this small manual inspection of the data provides anecdotal evidence about the differences between unambiguously marked examples and those that are not, and the possible factors that lead to detrimental effects on performance on the task of classifying rhetorical relations.

7 Conclusion

Obtaining enough training material is a common problem for supervised machine learning. This is especially true in an area like discourse processing where manual annotation of data is particularly time-consuming, due to the inherent subjectivity of the task and the careful training of the annotators that is required. In this paper,

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

we investigated whether it is feasible to bootstrap a rhetorical relations classifier by exploiting the presence of unambiguous discourse markers in some examples to label them automatically with the correct relation. A model trained on this data after the unambiguous markers have been removed, should, in theory, also be able to determine rhetorical relations in examples which occur naturally without unambiguous relation markers.

To test this hypothesis, we implemented two models: a Naive Bayes model which uses word co-occurrences to predict a relation; and a BoosTexter-based model which exploits a variety of linguistic cues. The first model is relatively knowledge-lean while the second is more complex, both in the feature set and in the underlying machine learning method. We then tested the performance of these models under three conditions: (i) when trained on automatically labelled data and tested on unseen data of a similar type (i.e., examples from which the relation-signalling discourse markers had been removed), (ii) when trained on automatically labelled examples and tested on examples which naturally occurred without an unambiguous discourse marker, and (iii) when trained and tested on manually labelled, unmarked or ambiguously marked data.

Both models performed relatively well under the first condition, which suggests that rhetorical relations can, in principle, be learnt from automatically labelled data. However, for both models, the performance dropped significantly (to an accuracy of around 25%, i.e., only about 5% higher than random guessing) when they were applied to examples which occur naturally without an unambiguous discourse marker. In other words, while the models learnt something from the automatically labelled data, they did not generalise to unmarked examples.

This performance degradation seems to be largely independent of the type of classifier used, as we observed it with both the simple, knowledge-lean Naive Bayes classifier and with the more complex BoosTexter model. Instead, it may be that the problem stems from the data itself. For example, unmarked and marked examples may just be too dissimilar linguistically to allow a classifier to generalise from one to the other. In particular, some properties which are predictive of a given relation in unmarked examples may not be predictive of the same relation in marked examples. We found evidence that this is indeed the case for some features. Furthermore, we also found evidence that the labelling process itself might have a detrimental effect and that the very act of removing the unambiguous discourse marker from the training data may lead to a meaning change, which may mean that the assigned relation label is no longer correct. Factors like these would have a negative effect on any model.

We also compared training on automatically labelled data to training on manually labelled data. We trained both models on a small set of manually labelled examples and tested them on unmarked data. For the Naive Bayes word pair model, the performance dropped compared to training on automatically labelled data, due to the fact that this model needs a fairly large amount of training data for a good performance; hence training on a large amount of automatically labelled data still leads to a better performance than training on a small amount of manually labelled, unmarked examples. However, for the BoosTexter model training on a small set manually labelled data led to a much higher performance than training on a large set

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

of automatically labelled data. Furthermore, only a very small amount of manually labelled examples (around 140) is required to obtain an accuracy comparable to that achieved by training on the automatically labelled data.

To sum up, it seems that training on automatically labelled, unambiguously marked data is not necessarily a good strategy when the aim is to develop a rhetorical relations classifier for those examples in which the relation is not unambiguously signalled. We found evidence that classifiers trained in this way do not generalise very well to unmarked data. Furthermore, the problem seems to be largely independent of the models and instead stem from the data itself. Training on even a very small amount of manually labelled data seems to lead to a much higher performance than training on a large automatically labelled data set, unless the model is one which, like the Naive Bayes model, requires a lot of training data. In that case, data quantity seems to be more important than quality. However, the highest accuracy we were able to obtain for the Naive Bayes classifier on unmarked data was around 25%, i.e., just above chance for our classification task. For most applications, this will be much too low. Hence, it may be better to invest resources in manually labelling even a relatively small set of data and use it to train a classifier that, like the BoosTexter model, does not require a huge amount of training data.

In future work, we plan to investigate in more detail why classifiers trained on automatically labelled data do not generalise to examples that are not unambiguously marked. This kind of knowledge is potentially useful as it could point to other ways in which automatically labelled data could be used. For instance, if only some automatically labelled data are problematic, it might be possible to use automatic example selection to identify unproblematic examples. It might then be possible to combine these with manually labelled examples to boost a small training set. Alternatively, if the problem arises mainly from the fact that features which are predictive of a particular relation in marked examples are not predictive of the same relation in unmarked examples, it might be possible to use automatic feature selection to avoid problematic features and employ only those which generalise across different types of data.

Acknowledgements

The research reported in this paper was carried out while the first author was at the University of Edinburgh. It was supported by EPSRC grant number GR/R40036/01. We would like to thank Vasilis Karaiskos and Max Leadley-Brown for doing the manual annotations and Mirella Lapata for helpful suggestions and discussions. We are also grateful to four anonymous reviewers for their comments and suggestions.

A List of discourse markers used in example labelling

Tables 9 to 13 below list the discourse markers we used to extract the automatically labelled data, together with one or two extracted examples for each. In the examples, discourse markers are shown in bold face and the hypothesised spans are indicated by square brackets. Note that using the cues on their own will overgenerate, hence we used them together with quite elaborated extraction rules, which are not shown.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 9. *Discourse Markers signalling* CONTRAST

Discourse Marker	Example	CONTRAST
although	[These abstract entities cannot be seen, smelt, touched or heard,] [**although** the effects which they produce on our physical beings can be observed in these ways.]	
but	[Research grants worth £420 million have been announced by the EC, revealing its plans for the third set of Esprit projects.] [**But** it may not be enough according to some industry commentators.]	
by comparison	[An analogue display on an electronic wristwatch would show a typical watch face with minute and second hands.] [Computers, **by comparison**, respond to signals that are in binary form, which means they are transmitted at two levels only.]	
(in|by) contrast	[The proposed toughening of the rules of engagement comes amid calls by NATO members for more effective air strikes against Bosnian Serb forces.] [UN officials, **in contrast**, express reluctance on rule changes that might alter the peacekeepers neutral role in the conflict and draw them deeper into the Balkans conflict.]	
conversely	[Kozyrev said in Paris Wednesday that Russia was not prepared to make its political and military actions subject to NATO,] [and **conversely** did not claim the right to veto NATO activities.]	
despite the fact	[China imported 10.8 million tonnes of steel in the first half of 1994,] [**despite the fact** that it had yet to use up the 30 million tonnes it imported last year.]	
even (so|though)	[Chretien offered no new policy initiatives,] [**even though** the convention is supposedly a policy-making convention.]	
however	[India, which exploded a nuclear device in 1974, says it does not have a bomb.] [In June, **however**, Prime Minister Rao announced that India would keep its nuclear options open.]	
in contrast	[Spain produced the fireworks this weekend when Real Madrid beat Atletico in an explosive derby match that saw six goals in the first half and a red card in the second.] [Italy, **in contrast**, produced floods, lashing rain and a set of damp Serie A results while France's headlines focused on Parisian crowd trouble.]	

(cont'd)

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Discourse Marker	Example	CONTRAST *(cont'd)*
in spite of	["The moment we realized the concept of the tournament is based on the contract between the IIHF and the NHL,] [**in spite of** that, we tried to do everything we could to play our best.]	
much as	[He added that he sympathized with "the desires of the heirs to find a place that is more philosophically in tune . . .] [**much as** we feel bad about it going."]	
(never|none)theless	[Moi is constitutionally barred from running again this year;] [the opposition is **nevertheless** determined to try and beat whoever is chosen to run for KANU.]	
notwithstanding	["He (Mugabe) is amending the electoral law himself,] [**notwithstanding** that he is a candidate himself," Hondora said.]	
on the other hand	[As well as losing its overall majority, the CDU was deprived of its junior coalition partner, the Free Democrats, who failed to make the five percent necessary to win a seat.] [The Greens **on the other hand** just made the mark.]	
regardless of	no example extracted	
that said	["I don't think that team can ever be duplicated," said David Robinson, a member of this year's squad as well as the 1992 and 1988 US lineups.] ["**That said**, I do feel this has the potential to be a very special team."]	
having said that	["The mayor has been able to put in significant cuts so far without any dramatic change in social tensions or even a significant downturn in the delivery of services.] [**Having said that**, it must be recognized that a work-force reduction of 15,000 people has occurred."]	
then again	["They're one of the most prolific offenses ever, but it's still always embarrassing giving up six touchdowns.] [**then again**, no one stopped them all year."]	
whereas	["This house (with its bleached wood floors and stark white walls) is so simple and uncluttered,] [**whereas** my life is very complicated and cluttered."]	
yet	["The overwhelming majority of patients want information about their care . . .] **yet** physicians seriously underestimate the amount of information patients want."]	

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 10. *Discourse Markers signalling* CONTINUATION

Discourse Marker	Example CONTINUATION
left cue . . . right cue[a]	[**First** the United States insisted on renewal and expansion of so-called "voluntary" purchases plans of auto parts announced by individual manufacturers, which are tantamount to purchasing quotas.] [**Second**, the United States also insisted on commitments on the future numbers of dealerships offering foreign brands.] ["**For one thing**, there are too many general contractors.] [**Also** it will be specialised construction groups which may benefit.]

[a] *Left cue* can be any of: *for one thing*, *for a start*, *first(ly)*, *second(ly)*, *third(ly)*, . . . , *tenth(ly)*. *Right cue* can be any of: *for another (thing)*, *further(more)*, *also*, *in addition*, *moreover*, *next*, *second(ly)*, *third(ly)*, . . . , *tenth(ly)*, *next*, *finally*, *last(ly)*.

Table 11. *Discourse Markers signalling* EXPLANATION

Discourse Marker	Example EXPLANATION
because[a]	[The Japanese justice ministry refused Maradona a visa] [**because** it said he had been implicated in drug cases more than once.]
for the reason that	["We cannot move to border areas] [**for the reason that** border areas are full of peril for our lives," it added.]
on the grounds that	[Choi said he decided to defect to the South after he was summoned back to Pyongyang from Yanji, in northeast China, near the border with North Korea,] [**on the grounds that** he had mixed too much with South Korean businessmen.]
(which|this) is why	[Many displaced people have said they would like to return home but lack transport,] [**which is why** UNAMIR has organised an evacuation.]

[a] This is for *A because B*. If the spans occur the other way round (*Because A, B*) *because* signals RESULT (according to the SDRT definitions) rather than EXPLANATION.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 12. *Discourse Markers signalling* RESULT

Discourse Marker	Example RESULT
as a consequence	[In-house personnel are being utilized on a full-time basis to carry out the above activities and] [**as a consequence**, internal data collection capabilities have increased.]
consequently	[It was noted at that time that there were low level information technology utilisation in the departments,] [**consequently** the cost of implementation would be a major factor when considering the design of either system.]
for this/that reason	[Of these, the reindeer, reintroduced in Scotland in 1952, is extremely limited in range;] [probably **for this reason** it is absent from the main text.]
in so doing	["The developed countries must do much more to foster economic growth in the developing world,] [and **in so doing**, help to combat poverty, inequality and exclusion," he said.]
in this way	[They wanted to recover a wider, more comprehensive vision of the church,] [and **in this way** to overcome the deeply entrenched divisions between the separated confessions and denominations.]
it follows that	["These pieces are very recognizable, but there is obviously a market for them somewhere," he said.] ["So **it follows that** they are going out of the country – perhaps to former Eastern bloc countries, or the Far East."]
it may be concluded	[The majority of actual examples investigated by Johnson revealed profiles of the second type,] [so that **it may be concluded** that much of the material forming the bar is eroded from the sea floor.]

Table 13. *Discourse Markers signalling* SUMMARY

Discourse Marker	Example SUMMARY
in other words	[The committee was to meet late Tuesday to decide whether Justice Minister Alfredo Biondi's decree is constitutional,] [**in other words** whether there was an emergency situation allowing the law to be introduced as a decree, rather than going through the usual parliamentary procedure.]
in short	[Both Seoul and Washington, which has 37,000 troops in South Korea, have said they want to see North Korea make a "soft landing"–] [**in short** avoid a chaotic collapse and a war.]

(cont'd)

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Discourse Marker	Example SUMMARY *(cont'd)*
put another way	[That $420 annuity payment would still be $420 in 33 years, even though it would have the buying power of only about $135 in today's money;] [**put another way**, you would need $1,307 in 33 years to buy what $420 buys today, given a 3.5 percent annual inflation rate.]
summing up	[Second, China has obtained great achievements in its reform and opening up which started in rural areas and spread across the country.] [So, **summing up** the experience in the past two decades is of important significance for the Party to persist in the policies adopted since the Third Plenary Session of the 11th CPC National Congress and to expand China's reforms and development.]
to sum up	[To counter that momentum, the industry has developed a host of arguments in defense of the cell.] [**To sum up**: There are plenty of distractions in any car at any given time, from putting on makeup to glancing at a newspaper.]
to summari(s|z)e	[What these types signify is that an individual's blood has or lacks a particular antigen, a protein.] [**To summarize**, group A has the A antigen but lacks B; B has B but lacks A; AB has both A and B; O lacks both A and B.]

B Examples of manually labelled relations

Tables 14 to 18 list three unmarked examples for each relation. Square brackets indicate the gold standard span boundaries as taken from the RST-DT (Carlson *et al.*, 2002).

Table 14. *Unmarked examples of* CONTRAST

[The executive said any buy-out would be led by the current board, whose chairman is Maurice Saatchi and whose strategic guiding force is believed to be Charles Saatchi.]
[Mr. Spielvogel isn't part of the board, nor are any of the other heads of Saatchi's big U.S.-based ad agencies.]

[Speaking through its Dutch lawyers, ASKO also disclosed it holds a 15% stake in Ahold.]
[It was previously thought ASKO held a 13.6% stake that was accumulated since July.]

[Prices closed lower in Sydney, Singapore and Wellington,]
[were mixed in Hong Kong and higher in Taipei, Manila, Paris, Brussels and Seoul.]

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 15. *Unmarked examples of* CONTINUATION

[Already, British Aerospace and French government-controlled Thomson-CSF collaborate on a British missile contract and on an air-traffic control radar system.]
[Just last week they announced they may make a joint bid to buy Ferranti International Signal PLC, a smaller British defense contractor rocked by alleged accounting fraud at a U.S. unit.]
In the new guidelines, the Justice Department says that in attempting to freeze disputed assets before trial,
["the government will not seek to disrupt the normal, legitimate business activities of the defendant"]
[and "will not seek . . . to take from third parties assets legitimately transferred to them."]
[The five astronauts returned to Earth about three hours early because high winds had been predicted at the landing site.]
[Fog shrouded the base before touchdown.]

Table 16. *Unmarked examples of* EXPLANATION

[Wyse Technology, for instance, is considered a candidate to sell its troubled operation.]
["Wyse has done well establishing a distribution business, but they haven't delivered products that sell," said Kimball Brown, an analyst at Prudential-Bache Securities.]
[Still, he adds: "We can't have this kind of thing happen very often.]
[When the little guy gets frightened, the big guys hurt badly.]
[The venture's importance for Thomson is great.]
[Thomson feels the future of its defense business depends on building cooperation with other Europeans.]

Table 17. *Unmarked examples of* RESULT

[Given the structure of most credit programs,]
[it is surprising that default rates are not even higher.]
[Unfortunately, the comment was buried in another article,]
[so it could not stand out in an education context.]
[A broker may have to approach as many as 20 underwriters who insure the endeavors on behalf of the syndicates.]
[It could take six months for a claim to be paid.]

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Table 18. *Unmarked examples of* SUMMARY

[Many agencies roll over their debt, paying off delinquent loans by issuing new loans, or converting defaulted loan guarantees into direct loans.] [In any case, they avoid having to write off the loans.]
["It will be very expensive," the spokesman warned.] ["The price cannot be less than $7,000."]
[As the Chinese have shown and the Soviets are learning, family farms thrive where collectives fail.] [Ownership, it seems, is the best fertilizer.]

References

Achinstein, P. (1980). *The Nature of Explanation*. Oxford University Press.

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

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

Bromberger, S. (1962). An approach to explanation. In R. J. Butler (Ed.), *Analytical Philosophy*, pp. 75–105. Oxford University Press.

Carlson, L., D. Marcu, and M. E. Okurowski (2002). RST Discourse Treebank. Linguistic Data Consortium.

Carlson, L., D. Marcu, and M. E. Okurowski (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt and R. Smith (Eds.), *Current Directions in Discourse and Dialogue*, pp. 85–112. Kluwer Academic Publishers.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Assocation for Computational Linguistics*, Seattle, WA, pp. 132–139.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics 29*(4), 589–638.

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

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Database*. Cambridge, MA: MIT Press.

Forbes, K., E. Miltsakaki, R. Prasad, A. Sarkar, A. Joshi, and B. Webber (2001). D-LTAG System – discourse parsing with a lexicalized tree adjoining grammar. In *Proceedings of the ESSLLI-01 Workshop on Information Structure, Discourse Structure and Discourse Semantics*.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Synax and Semantics Volume 3: Speech Acts*, pp. 41–58. Academic Press.

Hobbs, J. R., M. Stickel, D. Appelt, and P. Martin (1993). Interpretation as abduction. *Artificial Intelligence 63*(1–2), 69–142.

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

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Kamp, H. and U. Reyle (1993). *From Discourse To Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lapata, M. and A. Lascarides (2004). Inferring sentence-internal temporal relations. In *Proceedings of NAACL-04*, pp. 153–160.

Le Thanh, H., G. Abeysinghe, and C. Huyck (2004). Generation discourse structures for written text. In *Proceedings of COLING-04*, pp. 329–335.

Litman, D. J. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research 5*, 53–94.

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

Marcu, D. (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph. D. thesis, Department of Computer Science, University of Toronto.

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

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 365–372.

Marcu, D. and A. Echihabi (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL-02*, pp. 368–375.

Minnen, G., J. Carroll, and D. Pearce (2001). Applied morphological processing of English. *Natural Language Engineering 7*(3), 207–223.

Moore, J. D. and M. E. Pollack (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics 18*(4), 537–544.

Nomoto, T. and Y. Matsumoto (1999). Learning discourse relations with active data selection. In *Proceedings of EMNLP-99*.

Oates, S. L. (2000). Multiple discourse marker occurrence: Creating hierarchies for natural language generation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 41–45.

Pardo, T. A. S., M. das Graşas Volpe Nunes, and L. H. M. Rino (2004). DiZer: An automatic discourse analyzer for brazilian portuguese. In *Proceedings of the 17th Brasilian Symposium on Artificial Intelligence (SBIA)*.

Polanyi, L. (1985). A theory of discourse structure and discourse coherence. In P. D. K. W. H. Eilfort and K. L. Peterson (Eds.), *Papers from the General Session at the 21st Regional Meeting of the Chicago Linguistics Society*.

Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn (2004a). A rule based approach to discourse parsing. In *Proceedings of the 5th SIGDIAL Workshop in Discourse and Dialogue*, pp. 108–117.

Polanyi, L., C. Culy, M. van den Berg, G. L. Thione, and D. Ahn (2004b). Sentential structure and discourse parsing. In *Proceedings of the ACL-04 Workshop on Discourse Annotation*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*, 130–137.

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

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

Siegel, S. and N. J. Castellan (1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

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

Soricut, R. and D. Marcu (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press

Sporleder, C. and M. Lapata (2005). Discourse chunking and its application to sentence compression. In *Proceedings of the 2005 Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-05)*.

Sporleder, C. and A. Lascarides (2005). Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-05)*.

Webber, B. L., A. Knott, M. Stone, and A. Joshi (2003). Anaphora and discourse structure. *Computational Linguistics 29*(4), 545–588.

https://doi.org/10.1017/S1351324906004451 Published online by Cambridge University Press