4

Features for automatic discourse analysis of paragraphs

Daphne Theijssen, Hans van Halteren, Suzan Verberne and Lou Boves
Department of Linguistics, Radboud University Nijmegen

Abstract

In this paper, we investigate which information is useful for the detection of rhetorical (RST) relations between (Multi-) Sentential Discourse Units ((M-)SDUs) – text spans consisting of one or more sentences – within the same paragraph. In order to do so, we simplified the task of discourse parsing to a decision problem in which we decided whether an (M-)SDU is rhetorically related to either a preceding or a following (M-)SDU. Employing the RST Treebank (Carlson et al. 2003), we offered this choice to machine learning algorithms together with syntactic, lexical, referential, discourse and surface features. Next, we determined which of the features were most useful for predicting the direction of the relation by ranking them on the basis of three different metrics. Highly ranked features that predict the presence of a rhetorical relation are syntactic similarity, word overlap, word similarity, continuous punctuation and many reference features. Other highly ranked features predict the absence of a relations (i.e. are used to introduce new topics or arguments): time references, proper nouns, definite articles, the word *further* and the verb *bring*.

Proceedings of the 18th Meeting of Computational Linguistics in the Netherlands, pp. 53–68
Edited by: Suzan Verberne, Hans van Halteren, Peter-Arno Coppen.

4.1 Introduction

In the field of language and speech technology, the analysis of discourse structures in texts receives much attention. A commonly used model for discourse analysis is Rhetorical Structure Theory (RST), which was developed by Mann and Thompson (1988). RST is based on the idea that rhetorical relations exist between adjacent spans of text, of which one span, called the NUCLEUS, is more important for the purpose of the author than the other spans, called the SATELLITES. Sometimes spans are equally vital; then the relation is called multi-nuclear. The smallest text spans that can hold rhetorical relations are called Elementary Discourse Units (EDUS). The popularity of RST has led to the development of an RST Treebank of manually annotated English texts, which is available for training and testing purposes (Carlson et al. 2003). It consists of 385 Wall Street Journal articles from the Penn Treebank (Marcus et al. 1993) with a total of 176,383 words. In the RST Treebank, mono-nuclear relations are always binary (53 of 78 relation types), while multi-nuclear relations can also occur between more than two spans (the remaining 25 relation types).

The literature shows that various automatic RST parsers have been created. A state-of-the-art and publicly available system for automatic RST parsing of English texts is Sentence-level PArsing of DiscoursE (SPADE, Soricut and Marcu 2003). It produces an RST tree for every sentence in the input, but makes no attempt to find relations between sentences and at higher levels. In previous research by Marcu (Marcu 1999, Marcu 2000) and by others, the extraction of rhetorical relations at all text levels has been addressed. It has resulted in the discourse parsers RASTA (Rhetorical Structure Theory Analyzer, Corston-Oliver 1998) and DAS (Discourse Analyzing System, LeThanh 2004). However, RASTA nor DAS is generally available, nor is any of the systems reported by Marcu.

Because no system for automatic discourse (RST) analysis that is suitable for text analysis at all text levels is available, we decided to start the development of one. SPADE provides a first step towards such a system by splitting sentences into EDUS and providing RST trees for each sentence. A second step could be to find relations between text spans consisting of at least one sentence within the same paragraph. The goal of this paper is to investigate which information about the sentences may be useful for this second step. In other words, we attempt to answer the question: "Which features are most effective/powerful for predicting the presence of rhetorical (RST) relations between (Multi-)Sentential Discourse Units within paragraphs in English?" We introduce the term *(Multi-)Sentential Discourse Unit* ((M-)SDU) as a text span with a length of at least one sentence and at most one paragraph, forming a discourse unit in a text.

Following Soricut and Marcu (2003), we limited ourselves to binary relations[1]. We reduced discourse analysis to a classification task that we offered to various machine learning algorithms together with an inventory of potentially relevant features. Next, we ranked the features with the help of the classification algorithms

[1] In the RST Treebank, 99% of the rhetorical relations are binary (the remaining 1% being multi-nuclear relations between more than two spans).

and a feature ranking metric.

The organization of this paper is as follows: In Section 4.2, we introduce the classification task, describe the potentially relevant features and present the accuracies reached by the classification algorithms. The ranking of the features is described in Section 4.3. The final Section (8.4) contains our overall conclusion and gives recommendations for future research.

4.2 Discourse analysis as a classification task

In order to determine which features may be relevant for automatic discourse analysis of paragraphs, we simplified the problem of RST discourse analysis to a task that can easily be performed by machine learning algorithms. We ignored the type and direction of the rhetorical relations to prevent data sparseness.

In a binary RST tree, we considered each *triple* of three adjacent (M-)SDUs x – y – z in the same paragraph in which either x – y or y – z are rhetorically related. In such triples, two (M-)SDUs are fixed, namely those between which the rhetorical relation holds. The third one is only restricted by three conditions: (1) it should be an (M-)SDU, thus represented by a separate node in the RST tree, (2) it should be adjacent to one of the two fixed (M-)SDUs, and (3) it should be in the same paragraph. The same relation can thus be present in more than one triple in our data set.

We employed classification algorithms to classify the triples according to the position of the relation in the triple: on the *left* (x – y) or on the *right* (y – z). Note that x, y and z may consist of more than one sentence and may thus contain relations between (M-)SDUs themselves. This means that our approach covers all relations between (M-)SDUs in the paragraph, as the eventual object of discourse parsing of paragraphs requires.

With the help of a Perl script, we automatically extracted 2136 triples (1196 *right*, 940 *left*) from 942 different paragraphs in 246 different Wall Street Journal texts in the RST Treebank.

4.2.1 Features

Machine learning algorithms need information about the triples to be able to classify them. We followed two strategies to establish an inventory of potentially relevant features: (1) by considering the literature on approaches previously taken by Corston-Oliver (1998), Marcu (1999), Marcu (2000) and LeThanh (2004), and (2) by studying a sample of the RST Corpus[2]. The result is a list of features that we subdivided into surface features, syntactic features, lexical features, reference features and discourse features. All features values were determined automatically with the help of a Perl script.

[2]We randomly selected 30 texts with a length of at least 5 sentences in the RST Treebank. Next, we extracted all binary relations between (M-)SDUs within the same paragraph. We established the proportion of SDU – SDU, SDU – M-SDU, M-SDU – SDU and M-SDU – M-SDU relations in the treebank and randomly selected 200 relations with the same proportions.

Surface features

Marcu (1999) used the presence of words and part-of-speech (POS) tags as features in his machine learning approach. We included all lemmas and POS tags present in the data. For the purpose of lemmatization we employed the CELEX lexicon (Baayen et al. 1995), and we took the Part-of-Speech tags from the Penn Treebank. We also used trigrams containing either the word token or the POS tag in each slot (three adjacent words are thus represented by eight different triples). The (M-)SDU lengths (in sentences and in words) were taken into account as well.

Since each lemma, POS tag and trigram was considered a separate feature, the number of surface features was too large (over 18,000) to be computationally feasible. We therefore chose the 1,000 most useful surface features according to the feature selection algorithm Relief (Kononenko 1994). Only these features have been applied in the experiments and analyses[3].

Syntactic features

In Corston-Oliver (1998), the syntactic features *tense* (e.g. past), *aspect* (e.g. progressive) and *polarity* (e.g. negative) are introduced. We have used similar information by counting the (relative) number of modals, infinitives, gerunds, past forms and present forms in each (M-)SDU, and by checking the clauses for negation.

A potentially relevant feature we discovered in the sample of the RST Treebank is *syntactic similarity*, as exemplified in Table 4.1. Existing metrics to establish syntactic similarity were not suitable for our purpose: parser evaluation metrics such as *Parseval* (Black et al. 1991) require that the two compared structures describe the same sentence, and methods such as *document fingerprinting* (Bernstein and Zobel 2005) establish the similarity of larger texts, not of small units such as (M-)SDUs. We have developed a simple metric which determines the syntactic similarity of two (M-)SDUs by comparing their clause structures (Theijssen 2007). The result is a continuous value between 0 and 1.

Lexical features

The example illustrating syntactic similarity also indicated the relevance of cue phrases such as *but*, *for this reason*, *in short*, etc. This has also been argued by Corston-Oliver (1998), Marcu (1999), Marcu (2000) and LeThanh (2004). We have included all 207 cue phrases that LeThanh considers relevant above clause level[4]. LeThanh (2004) also introduced noun phrase (NP) and verb phrase (VP) cues such as *goal* (NP), *purpose* (NP and VP) and *result from* (VP). We included all her 41 NP and 56 VP cues in our experiments[5].

[3]This selection was done separately for each individual training set in the ten-fold cross-validation described in Section 4.2.2. In total, 7,828 unique surface features have been selected.
[4]Only 21 of them were found in our total data set of 2136 triples.
[5]In our total data set of 2136 triples, 20 NP and 43 VP cues were present.

Table 4.1: Example of syntactic similarity in wsj_0688

	SDU 1	SDU 2
adverb	-	*But*
PP	*For instance*	*on the West Coast, where profitable oil production is more likely than in the midcontinent region, the Bakersfield, Calif.*
NP subject	*employment in Denver*	*office staff of 130*
modal	*will*	*will*
lexical verb	*be reduced*	*grow*
PP	*to 105*	*by 175*
PP	*from 430*	*to 305*

Other lexical features we found in the literature and the data sample were *word overlap* and *word similarity*. We defined three types of word overlap, namely the relative number of overlapping tokens, lemmas and stems. Word similarity was measured by employing Extended Gloss Overlap in *WordNet::Similarity* (Pedersen et al. 2004) and by consulting Lin's (1998) *Dependency Thesaurus*. In the example below, similar words are marked.

> *The FDA has said it* ***presented evidence*** *it uncovered to the company indicating that Bolar* <u>*substituted*</u> *the brand-name product for its own to gain government approval to sell generic versions of Macrodantin. Bolar has* ***denied*** *that it* <u>*switched*</u> *the brand-name product for its own in such testing.*
> — wsj_2382

Seeing data instances such as that below, we expected that the *presence of time references* could also be a relevant feature:

> <u>*Until recently*</u>*, Adobe had a lock on the market for image software, but* <u>*last month*</u> *Apple, Adobe's biggest customer, and Microsoft rebelled.* <u>*Now*</u> *the two firms are collaborating on an alternative to Adobe's approach, and analysts say they are likely to carry IBM, the biggest seller of personal computers, along with them.*
> — wsj_2365

Reference features

We found that many of the rhetorically related (M-)SDUs in the sample of the RST Treebank contained references. Referring to previously mentioned items by using personal pronouns, definite articles, demonstrative pronouns and (wh-)determiners (e.g. *which*) was therefore represented in features indicating their presence and their relative frequency in the (M-)SDU. We also established a list of 31 reference adverbs and adjectives (e.g. *other*) that we included in our approach. The list was based on the words found in the sample, supplemented with synonyms taken from the thesaurus of Microsoft Word 2003[6].

Corston-Oliver's (1998) system also includes an anaphora resolver which automatically finds the antecedents of reference words. Since the system is not generally available, we employed the anaphora resolution tool GuiTAR (Poesio and Alexandrov-Kabadjov 2004) to check whether an anaphoric relation was present between two (M-)SDUs.

We here introduce a new feature *NP simplification*, being the lack of NP modifiers or NP heads in noun phrases that have been used previously in the text. Both types are illustrated below: the head *transaction(s)* in the first example, and the modifiers *Wall Street Journal's "American Way of Buying"* in the second example are missing in the second underlined phrase:

> *Grimm counted 16 transactions valued at $1 billion or more in the latest period, twice as many as a year earlier. The largest was the $12 billion merger creating Bristol-Myers Squibb Co.*
> — wsj_0645

> *When consumers have so many choices, brand loyalty is much harder to maintain. The Wall Street Journal's "American Way of Buying" survey found that 53% of today's car buyers tend to switch brands. For the survey, Peter D. Hart Research Associates and the Roper Organization each asked about 2,000 U.S. consumers about their buying habits.*
> — wsj_1377

Discourse features

The last type of features concerns information on the structure of the text. From what we saw in the sample of the RST Treebank, we expected that the presence of continuous punctuation is a helpful cue for the detection of rhetorical relations. In the example below, the second quotation part consists of more than one sentence. Moreover, both sentences are between (the same) parentheses:

> *("A turban," she specifies, "though it wasn't the time for that 14 years ago. But I loved turbans.")*
> — wsj_1367

Also, we included information on the position of the (M-)SDU in the text (paragraph number) and in the paragraph (sentence number) as a discrete feature. The internal (binary) discourse structure of the (M-)SDU was also taken into account. We represented this by the number of EDUs and the nuclearity (NUCLEUS or SATELLITE) of both spans in the highest rhetorical relation. For example, if the internal discourse structure of an (M-)SDU is N1-S3, it contains a relation between a NUCLEUS span of 1 EDU and a SATELLITE span of 3 EDUs.

[6] The English thesaurus of Microsoft Word 2003 was developed for Microsoft by Bloomsbury Publ.

4.2.2 Method

We have formulated definitions for each of the features and have written Perl scripts for the automatic extraction of the feature values. Where possible, we used existing resources and tools, e.g. the syntactic analyses in the Penn Treebank. Depending on the form of the feature, its value had to be extracted for each (M-)SDU x, y and z in the triple, or for both pairs x-y and y-z. In total, 1,836 features were used, being the 1,000 best surface features, 20 syntactic features, 718 lexical features, 84 reference features and 14 discourse features. For details on the definition and extraction of the features, the reader is referred to Theijssen (2007).

We applied five machine learning algorithms: *Naive Bayes*, *k-Nearest Neighbours* (*kNN*), *Support Vector Machines* (*SVM*), *Decision Trees* and *Maximum Entropy*. The first four are present in the Orange software (Demsar et al. 2004) and we chose to employ those implementations. For Maximum Entropy we used the implementation of Zhang (2004). Since there was not enough data to establish the optimal parameters for each algorithm, we applied the algorithms with their default settings. The continuous features were made discrete by dividing their range into seven equal-frequency intervals with the 'discretization'-function in Orange, and were offered in this form to Naive Bayes and Maximum Entropy. Although potentially important knowledge about the exact distribution of the feature values is lost in this way, we believe it is a necessary step to reach (interpretable) output.

Due to the rather small number of triples, we decided to apply ten-fold cross-validation on all cases. It would not be fair to place some triples extracted from a particular Wall Street Journal text in the train data and other triples extracted from the same text in the test data. Therefore we had to manually split the data into partitions with equal numbers of triples and of texts.
The results are compared to the accuracy reached by chance: 56.0% (always choosing *right*). In order to establish an upper bound of the classification task, we randomly selected 50 triples in 50 different paragraphs. Two human analyzers who are familiar with RST reached an average accuracy of 87.0% on the classification task. Their inter-annotator agreement was substantial: Cohen's (1960) Kappa = 0.78.

4.2.3 Results

Since the machine learning task concerns choosing between only two classes (*left* and *right*), and the distribution of both classes is known, the machine learning results are represented by the *accuracy*, being the number of correctly classified cases in the test set divided by the total number of cases in the test set. The accuracies reached by the algorithms can be found in Table 4.2. Comparison with existing discourse parsers is not possible because they perform a different task.

Only Naive Bayes and Maximum Entropy reached an accuracy that is significantly better than chance. To check whether the other algorithms were affected by the large number of features and the low number of cases, we offered fewer features to them by employing Relief for feature selection, and selecting the best features for each partition. As expected, the performance of kNN, SVM and Decision Trees increased when fewer features were offered, but only SVM ever performed significantly better than chance[7]. Naive Bayes and Maximum Entropy showed the opposite effect: their classification accuracy decreased when using fewer features.

Table 4.2: Accuracies reached. * means $p<0.001$ (compared to chance)

Chance	Naive Bayes	kNN	SVM	DecTrees	MaxEnt	Human
56.0%	60.0%*	51.1%	56.9%	53.1%	60.9%*	87.0%*

4.2.4 Discussion

Despite our efforts to include good representations of all potentially relevant information, the accuracies reached by the machine learning algorithms were only slightly better than chance (56.0%). An explanation for the results could be that the default settings in Orange were not optimal for the given task and data. The default k in kNN, for example, is the square root of the number of cases in the training set. It is possible that a lower k could increase the accuracy reached and thereby the suitability of the system and its model. Adjusting the parameter setting is thus highly recommended for future research.
Since it is not our goal to reach high accuracy on the classification task, but to establish what information (which features) are useful in the detection of rhetorical relations, the problem is less severe than it seems. Still, an important consequence of the low accuracies reached by these classification algorithms is that analyzing the models is speculative and should thus be performed with care.

4.3 Feature ranking

In order to discover which of the features in our feature set are most useful, we ranked them on the basis of three different metrics: (1) a metric based on the model of Naive Bayes, (2) a metric based on the model of Maximum Entropy, and (3) a feature ranking metric developed by one of the authors (Van Halteren, see below). Earlier in our research we employed the feature selection algorithm that comes standard with Orange (Demsar et al. 2004), being Relief (Kononenko 1994). Since Relief shows only a low to medium correlation with the other three metrics (the Spearman's rank correlation coefficient is 0.33 to 0.51 with $p<0.001$), we excluded it from the ranking described in this section. Exploring the differences between Relief and the other three metrics is beyond the scope of this paper.

[7] When provided with the best 100 features, the accuracy is 58.7% with chi-square 6.17, $p<0.05$

4.3.1 Method

Given the significant improvement over chance, we believe Naive Bayes and Maximum Entropy were able to sift the information from the sets of features with some success. Assuming that this sifting is expressed in the model parameters, we attempted to extract an indication of feature importance. As for the systems that were not able to perform better than chance, they were obviously unable to discover the information and any ranking is not likely to provide a useful measurement of feature importance.

To find a relevance score for the features following the model of Naive Bayes, we established the probability of each feature given the class. We approached this by considering both classes *left* and *right* and counting the number of times a certain feature value occurred with that class, and divided it by the total number of cases with the class in the training set. We then looped through all cases in the test set and divided the probability of the feature value given the correct class by the probability of the feature value given the incorrect class, and took the log. The result was the contribution of the feature value for that particular case. We then averaged the attributions over all cases in each fold to achieve a single relevance score for the feature.

Maximum Entropy considers each feature with each value separately and therefore established a weight (relevance score) for each feature-value combination. Since we need a relevance score per feature rather than per feature-value combination, we calculated a weighted average relevance score for each feature, using the frequencies of the feature values as weights. The result was averaged over the 10 training sets. The model also shows which class is best selected for which feature value, enabling us to establish the preferred class when a feature is *present* (binary features) or relatively high (continuous features). Sometimes, the preferred class of a continuous feature varied per frequency range and no general trend could be detected.

The third metric is the feature ranking algorithm *Cluster Separation Score* (*CSS*), developed by Van Halteren. CSS is determined for each feature by dividing the difference between the means of the values with class *left* and class *right* by the sum of the standard deviations of the values with class *left* and class *right*. The resulting relevance score is an indication of the extent to which the feature is able to distinguish the cases with class *left* from those with class *right*. As with the model of Maximum Entropy, the formula shows which class is best selected for which feature value. *CSS* requires that the feature values are continuous, which was problematic for our data since the great majority consists of discrete (nominal) features. We converted these features (such as the presence or absence of a POS tag) to numerical features with values *0* and *1*. We assumed that despite the fact that the features are not truly continuous, the metric will still be able to estimate the relevance of the features.

Since it is undesirable to draw conclusions on features that occur in only one partition of the data, we removed those from the three rankings found. They are features that only have the values *absent* and *not applicable* (for binary features) or *0* (for continuous features) in nine or ten partitions. From the total of 8,664 features[8], 806 features[9] remained after this removal.

In order to reach a final ranking of the 806 features, we averaged the rankings in the three methods. In these rankings, we have given features with equal ranking scores the same rank by averaging over their positions (e.g. rank 4.5 for two equally useful features at positions 4 and 5).

4.3.2 Results

Since an overview of all 806 features would be too extensive to suit this paper and would include a discussion of irrelevant features at the bottom of the list, we limit ourselves to the 50 best features following our ranking. Note that the features have either been determined for all three (M-)SDUs *x*, *y* or *z*, or for both (M-)SDU pairs *x-y* and *y-z* in the triple. Therefore, the features have forms such as *the (x)*, being the presence of the word *the* in *x*, or *anaphora (y-z)*, being the presence of an anaphoric relation between *y* and *z*. This section presents the findings for each feature type. The top 10 can be found in Table 4.3. The last column shows the range of the ranks found with the three different metrics.

Table 4.3: Top 10 of 806 features

	Feature	*Pos.*	*Feature type*	*Av. rank*	*Range*
1	pers. pronoun in first clause	*z*	reference	2	1–3
2	cont. quotation marks	*y-z*	discourse	3	1–6
3	word similarity (Lin)	*y-z*	lexical	4	4–4
4	word similarity (Lin)	*x-y*	lexical	7	5–11
5	token overlap	*y-z*	lexical	10	7–17
6	missing modifier	*y-z*	reference	12	2–23
7	def. article in first clause	*z*	reference	17	11–20
8	proper noun sg.	*z*	surface	17	9–22
9	proper noun sg.	*y*	surface	20	11–36
10	past tense	*x*	syntactic	21	12–36

Surface features

Of the 579 surface features, 10 features are in our top 50 (being words and POS tags). The trigrams, which were also included as surface features, are not present in the top 50. This is probably because of the small size of our data set.

The following word features are included in the top 50: *as (y)*, *farmer (z)*, *little (y)*, *the (z)* and *to (y)*. The presence of the word *farmer* in this top 50 is probably

[8] 7,828 different surface, 20 syntactic, 718 lexical, 84 reference and 14 discourse features.

[9] 579 surface, 20 syntactic, 136 lexical, 61 reference and 10 discourse features.

caused by the specific text type and data set[10]. The word *little* in *y* seems to refer back to *x*, because the relation for this feature is expected between *x* and *y* by the metrics of Maximum Entropy and CSS:

> $_x$*[If the pound falls closer to 2.80 marks, the Bank of England may raise Britain's base lending rate by one percentage point to 16%, says Mr. Rendell.]*$_x$ $-$ $_y$*[But such an increase, he says, could be viewed by the market as "too little too late."]*$_y$ $_z$*[The Bank of England indicated its desire to leave its monetary policy unchanged Friday by declining to raise the official 15% discount-borrowing rate that it charges discount houses, analysts say.]*$_z$
> — wsj_0693

The relevance of the definite article *the* in *z* (ranked 15th) seems to confirm our intuition that *the* can be used as a reference word, and that references are important in discourse. However, in cases where *the* is present in *z*, CSS expects a relation between *x* and *y*, not between *y* and *z*, as in the example below:

> $_x$*[He made numerous trips to the U.S. in the early 1980s, but wasn't arrested until 1987 when he showed up as a guest of then-Vice President George Bush at a government function.]*$_x$ $-$ $_y$*[A federal judge in Manhattan threw out the indictment, finding that the seven-year delay violated the defendant's constitutional right to a speedy trial.]*$_y$ $_z$*[The appeals court, however, said the judge didn't adequately consider whether the delay would actually hurt the chances of a fair trial.]*$_z$
> — wsj_0617

Apparently, *the* is more often used to introduce a new topic or argument in the text than to refer back to the previous (M)SDU. Journalists of the Wall Street Journal probably assume that readers are familiar with certain notions and topics (in this case *the appeals court*), thus mentioning them with the definite article.

POS tags that are in the top 50 are: *personal pronoun (z)*, *proper noun singular (y,z)*[11] and *third person singular verb (x)*. Personal pronouns are cues that the (M-)SDU in which it appears is rhetorically related to the previous (M-)SDU. Proper nouns are common in financial newspaper texts. The metrics of Maximum Entropy and CSS show that when a proper noun is present in the *z*, the relation is most likely between *x* and *y*, and when in *y*, between *y* and *z*. Apparently, a person or company often introduces a new topic. According to CSS, a *third person singular verb* in *x* is an indicator of a relation between *y* and *z*, while no consistent pattern is displayed by the metric of Maximum Entropy. Its presence in the top 50 may hint at the relevance of syntactic structure, as we will also find below.

The last surface feature present in the top 50 is the length (in words) of *y*. CSS and Maximum Entropy disagree on the direction of the rhetorical relation for this feature.

[10] The word *farmer* is quite frequent in the RST Treebank (and consequently in our data): in triple-final position (*z*), it is present in 25 triples extracted from 9 different paragraphs in 3 different texts in the Treebank.

[11] The notation *proper noun singular (y, z)* represents two features: *proper noun singular (y)* and *proper noun singular (z)*.

Syntactic features

The top 50 includes 5 of the 20 syntactic features. Both *present (x)* and *past (x)* tense are present in the top 50. Also included are *gerunds (y)* and *infinitives (y)*. The relatively high ranking of these syntactic features seems to indicate that syntactic structure is related to discourse structure.

Syntactic similarity is also in the top 50, but only for the left pair (*x*-*y*). CSS expects a rhetorical relation on the left when the syntactic similarity between *x* and *y* is high. This is in accordance with the literature and our intuitions based on inspections of the data. For Maximum Entropy, the direction depends on the similarity range: the expected class varies per interval (in the discretized version of syntactic similarity), and no general trend can be found.

Lexical features

Of the 718 lexical features, 12 belong to the 50 best features. Despite the fact that cue phrases are used in all systems discussed in the beginning of this paper, none of LeThanh's (2004) cue phrases come forward in our approach. NP cues (also taken from LeThanh) present in the top 50 are *speculation (x)* and *goal (y)*, and VP cues *affect (y)*, *assume (y)*, *have to (y)* and *bring (z)*. An (M-)SDU with *goal*, *affect*, *assume* or *have to* is most likely rhetorically related to the following (M-)SDU. The presence of *speculation* in *x* indicates a relation between *y* and *z*. Apparently, it is used to close a topic. The verb *bring* appears to introduce a new topic, since a relation is expected between *x* and *y* when *bring* is present in *z*.

Word overlap is ranked in the top 50 only for the right pair in the triple (*y*-*z*). A relatively high word overlap implies there is a rhetorical relation between the two (M-)SDUs concerned, which is what we expected.

The same expected pattern is found for *word similarity Lin (x-y, y-z)*. The higher the similarity, the higher the chance that a rhetorical relation exists. Word similarity on basis of WordNet is not present in the top 50. It is commonly known that the wide coverage of WordNet may lead to problems when applied to specific domains such as financial newspaper texts. Because Lin's Thesaurus was trained on Wall Street Journal texts, it is not surprising that the similarity based on Lin's Thesaurus is more useful for our task than the similarity based on WordNet.

Time references are only useful enough to be in the top 50 when they occur in *z*. According to both CSS and the model of Maximum Entropy, the presence of a time reference in *z* indicates that the relation is probably between *x* and *y*. This would mean that time references introduce new topics that are not rhetorically related to the previous (M-)SDUs, for example as in:

> $_x$*[Witnesses have said the grand jury has asked numerous questions about Jacob F. "Jake" Horton, the senior vice president of Gulf Power who died in the plane crash in April.]*$_x$ – $_y$*[Mr. Horton oversaw Gulf Power's governmental-affairs efforts.]*$_y$ $_z$*[On the morning of the crash, he had been put on notice that an audit committee was recommending his dismissal because of invoicing irregularities in a company audit.]*$_z$

— wsj_0619

Note, however, that this example differs from the example in Section 4.2.1 where both *(M-)*SDUs in the relation contain a time reference. In order to also capture such instances, the presence of time references in two adjacent (M-)SDUs is best included as a separate feature in future research.

Reference features

Reference features are the most frequent (18) in the top 50[12]. The best feature according to our ranking method is a reference feature: a *personal pronoun in the first clause (y-z)*. The same feature but in *x-y* can be found at place 36 in the ranking. Also in the top 10 is the feature *definite article in the first clause (y-z)*. As we already saw in the discussion of the surface feature *the* above, the presence of a definite article in the first clause of z indicates that the relation is expected between *x* and *y*. Apparently, definite articles are most often used to refer to what is assumed to be known, not to what has previously been mentioned in the article.

Still, most reference features in the top 50 confirm our intuitions that anaphoric references in different forms are cues for rhetorical relations. When there is an *anaphoric relation (x-y, y-z)* between two (M-)SDUs, it is an indication that there is a rhetorical relation. A high *relative number of demonstrative (y-z)* and *personal pronouns (x-y, y-z)*, and the *presence of personal pronouns (x-y, y-z)* also predict a rhetorical relation with the preceding (M-)SDU. The relevance of personal pronouns has already been found in the surface features, as discussed above. Again, the definite article is especially used to introduce a new topic: a relatively high *number of definite articles* in z is a predictor of a relation between *x* and *y*.

Of the reference words in the top 50 (*added (x-y, y-z)*, *further (x-y)*, *less (x-y, y-z)* and *other (x-y, y-z)*), only *further* is not used to refer to the previous (M-)SDU. In our data, *further* seems to ask for an elaboration, for example:

> $_x$*[Operating profit grew 57% to* $269 million from 171$ *million, while operating margins rose to 16.1% from 15.9% the previous quarter and 12.6% a year ago.]*$_x$ $_y$*[Daniel Akerson, MCI chief financial officer, said the company sees further improvements in operating margins.]*$_y$ $-$ $_z$*["We think we can take it to the 18% range over next 18 to 24 months," he said.]*$_z$
> — wsj_1999

The last reference feature we defined, NP simplification, is present in the top 50 in the form of *missing modifiers (y-z)* and *missing head (x-y)*. When there are missing modifiers or the head is missing, a relation is indeed expected in that (M-)SDU pair.

[12]In our feature set, reference features are always determined for (M-)SDU *pairs*. Since we expect reference items to refer back, features such as the presence of reference words or of a personal pronoun in the first clause always concern the second (M-)SDU of a pair.

Discourse features

The top 50 includes 4 of the 14 discourse features available. The feature describing the position of the (M-)SDU in the paragraph (*sentence number in the paragraph (y, z)*) is included. We believe the relevance of this feature can be explained by the general structure of paragraphs in newspaper articles. This newspaper structure also makes itself felt in the feature *internal discourse structure (z)*. Testing these features on different text genres is necessary to establish whether our intuitions about newspaper structure are valid.

As expected, the presence of *continuous quotation marks (y-z)* is a cue for the presence of a rhetorical relation in both CSS and the model of Maximum Entropy.

4.3.3 Discussion

In the ranking, we have seen that some features are highly ranked by two of the three algorithms, but lower by the third. This explains the sometimes wide ranges in the last column of Table 4.3. Note that the average rankings are still relatively good given the large number of features (806). The best feature, *personal pronoun in first clause*, appears to be important for all three algorithms.

The list of best 50 features following our ranking strategy contains 'positive' features (expecting a rhetorical relation) as well as 'negative' features (introducing a new item in the text). Positive features are syntactic similarity, word overlap, word similarity (following Lin's (1998) Dependency Thesaurus), continuous punctuation and almost all reference features. Negative features include time references, proper nouns, definite articles, the word *further* and the verb *bring*. Obviously, both positive and negative features are useful for discourse analysis.

Some features unexpectedly come forward from our approach as relevant. The word *farmer*, for example, is likely to be dependent on the data set and therefore a bad predictor. Similarly, the high ranking of discourse features such as the position of (M-)SDUs in the paragraph and their internal discourse structure is probably caused by the general (financial) newspaper structure of the data. Results such as these can be prevented in future by extending the data with more texts of different genres. This may be difficult since no such data are available with discourse (i.e. RST) annotations yet.

4.4 Conclusion

In this paper, we have aimed at answering the question "Can we identify features that can be used to predict the presence of rhetorical (RST) relations between (Multi-)Sentential Discourse Units within paragraphs in English?" By reducing RST parsing to a classification problem, using an inventory of potentially relevant features (Section 4.2) and ranking them on the basis of three different metrics (Section 4.3), we have succeeded in this.[13] Some of the relevant features we have

[13]The feature values, the Perl scripts and the feature relevance scores found can be downloaded from `http://lands.let.ru.nl/~daphne`.

found predict the presence of a rhetorical relation (e.g. word similarity), while others are more often used to introduce new topics or arguments (the definite article for example).

In our research, we have reduced discourse analysis to a rather artificial classification task that is only a first step towards automatic discourse analysis. Also, we have limited ourselves to existing implementations of algorithms and metrics, without adjusting the parameters and without closely examining their capability to deal with our data. As this may have resulted in low accuracy and therefore only speculative rankings, we advise other researchers who plan to use our ranking to test the features on the real task with systems and settings that are more tuned to this kind of data.

References

Baayen, R.H., R. Piepenbrock, and L. Gulikers (1995), The CELEX Lexical Database (CD-ROM), *Technical report*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA.

Bernstein, Y. and J. Zobel (2005), Redundant documents and search effectiveness, *Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM Press, New York, Bremen, Germany, pp. 736–743.

Black, E., S. Abney, S. Flickenger, C. Gdaniec, C. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini, and T. Strzalkowski (1991), Procedure for quantitatively comparing the syntactic coverage of English grammars, *Proceedings of the workshop on Speech and Natural Language*, Leiden, the Netherlands, pp. 306–311.

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

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20** (1), pp. 37–46.

Corston-Oliver, S.H. (1998), *Computing Representation of Discourse Structure*, PhD thesis, Dept. of Linguistics, University of California, Santa Barbara, CA, USA.

Demsar, J., B. Zupan, and G. Leban (2004), Orange: From Experimental Machine Learning to Interactive Data Mining, *Technical report*, Faculty of Computer and Information Science, University of Ljubljana. Software available at `http://www.ailab.si/orange`.

Kononenko, I. (1994), Estimating Attributes: Analysis and Extensions of RELIEF, *Proceedings of the European Conference on Machine Learning*, pp. 171–182.

LeThanh, H. (2004), *Investigation into an Approach to Automatic Text Summarisation*, PhD thesis, Middlesex University, UK.

Lin, D. (1998), Automatic retrieval and clustering of similar words, *Proceedings of the 17th international conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 768–774.

Mann, W. and S. Thompson (1988), Rhetorical structure theory: Toward a functional theory of text organization, *Text* **8** (3), pp. 243–281.

Marcu, D. (1999), A decision-based approach to rhetorical parsing, *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, Maryland, USA, pp. 365–372.

Marcu, D. (2000), The rhetorical parsing of unrestricted texts: a surface-based approach, *Computational Linguistics* **26** (3), pp. 395–448.

Marcus, P.M., M.A. Marcinkiewicz, and B. Santorini (1993), Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics* **19** (2), pp. 313–330.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004), WordNet::Similarity – Measuring the Relatedness of Concepts, *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, Boston, MA, USA, pp. 38–41.

Poesio, M. and M. Alexandrov-Kabadjov (2004), A general-purpose, off the shelf anaphoric resolver, *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Soricut, R. and D. Marcu (2003), Sentence Level Discourse Parsing using Syntactic and Lexical Information, *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada.

Theijssen, D. (2007), *Features for automatic discourse analysis of paragraphs*, Master's thesis, Radboud University Nijmegen, The Netherlands. Available at `http://lands.let.ru.nl/~daphne/MA_thesis.html`.

Zhang, L. (2004), *Maximum Entropy Modeling Toolkit for Python and C++*.