Natural Language Engineering **8** (2/3): 235–255.
DOI: 10.1017/S1351324902002905 Printed in the United Kingdom

Robust discourse parsing via discoure markers, topicality and position

FRANK SCHILDER

Department for Informatics, University of Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany

(*Received 15 July 2001; revised 11 February 2002*)

Abstract

This paper describes a simple discourse parsing and analysis algorithm that combines a formal underspecification utilising discourse grammar with Information Retrieval (IR) techniques. First, linguistic knowledge based on discourse markers is used to constrain a totally underspecified discourse representation. Then, the remaining underspecification is further specified by the computation of a topicality score for every discourse unit. This computation is done via the vector space model. Finally, the sentences in a prominent position (e.g. the first sentence of a paragraph) are given an adjusted topicality score. The proposed algorithm was evaluated by applying it to a text summarisation task. Results from a psycholinguistic experiment, indicating the most salient sentences for a given text as the 'gold standard', show that the algorithm performs better than commonly used machine learning and statistical approaches to summarisation.

1 Introduction

The output of a discourse parser is a discourse tree that reflects the rhetorical structure of the input text. Obtaining robustness for a discourse parser is a demanding task due to the many unresolved theoretical issues regarding the derivation of the discourse structure. Although formal model-theoretic approaches such as Discourse Representation Theory (DRT) (Kamp and Reyle 1993) or its extension, Segmented DRT by Asher (1993), can provide a detailed analysis of the content of larger texts, this can only be done when world knowledge is specified *a priori* in very great detail. However, a world knowledge representation system encompassing the knowledge needed to understand, for instance, a newspaper article, does not yet exist. As a consequence, other means have to be found for a robust rhetorical parser.

The outcome of a discourse parser is a hierarchical discourse structure representing the rhetorical information of the text. The discourse structure is not only of theoretical interest, deriving the discourse tree structure can also improve performance of Information Extraction (IE) tasks or IR applications such as text summarisation (Sumita et al. 1992; Marcu 1999b). Thus, a robust rhetorical parsing algorithm that accurately determines the discourse structure of text would be useful for many practical applications.

Current robust discourse parsers, however, face the following three problems. First of all, such a parser can only ensure a very shallow processing of the text's content, since a world knowledge representation system with a wide coverage is not (yet) accessible. Consequently, any world knowledge is only indirectly captured by corpora investigations and/or inserted via genre-dependent hand-written rules. Second, the parser is tied to a discourse theory (e.g. RST) and is thus restricted by this discourse theory's constraints (e.g. predictions on discourse attachment). Third, future developments on the theoretical issues of discourse parsing are difficult to incorporate into current state-of-the-art systems. Present systems cannot simply be updated. New findings require the programming of an entirely new system.

The present paper proposes an algorithm for the derivation of discourse structure that does not need a world knowledge representation system and is not necessarily committed to one discourse theory, but is still general and applicable to all text sorts. The algorithm encompasses findings from two research strands: a formal discourse theory (Underspecified SDRT (USDRT); Schilder (2000)) and techniques developed for IR applications (vector space model; Salton (1971) and position method; Edmundson (1969)). The discourse grammar chosen employs underspecification. Hence it is flexible enough to incorporate future developments that can be encoded as more restrictive constraints. The lack of a world knowledge reasoning system is compensated by the IR techniques. For the time being, these techniques are only an approximation for the inferences that can be drawn from world and context knowledge but note that these techniques ensure the robustness of the present system.

Recent approaches to discourse processing, on the other hand, that allow robust parsing require extensive corpora investigations. Within the Rhetorical Structure (RST) framework (Mann and Thompson 1988), Marcu (1999a), for example, proposes a rhetorical parsing algorithm that applies decision-based learning techniques. A crucial prerequisite for the success of the parser, however, is a discourse corpus tagged with rich rhetorical and semantic information. Unfortunately, there is still a lack of such corpora and compiling them is quite work intensive and time-consuming. Other approaches that do not draw data from corpora employ hand-written rules (Sumita et al. 1992; Marcu 1997; Corston-Oliver 1998). The disadvantage of these is that they only work in the given text domain for which those rules were written. More importantly, these rhetorical parsers face the problem of ambiguous discourse sequences and hence may produce more than one output. Only Marcu, in his later work, offers a solution for this problem (Marcu 1998). In this more refined approach of his rhetorical parser, he proposes to combine different heuristics in order to decide on the best discourse tree. The heuristics used are similar to the ones employed by the current proposal. However, there are important differences with respect to the computation of the different metrics that will be discussed in more detail in subsequent sections. Crucially, Marcu's approach differs from the current one in the way these different metrics are combined. Marcu's refined system adds all n metrics m_i in a linear equation, such as $\alpha \times m_1(D) + \beta \times m_2(D) \ldots + \omega \times m_n(D) = s(D)$. A score $s(D)$ for a derived discourse tree D indicates how well this tree indicates the underlying discourse structure in comparison with other trees derived by the rhetori-

cal parser. The optimal weighting for the weights α to ω is found via a machine learning algorithm that can compute the optimal combinations of metrics given a training corpus. This kind of approach has the following flaws. First, an important precondition for a linear combination of metrics is the assumption that the n metrics are independent. However, this is surely not the case, because all features that are reflected by the metrics have to be taken into account as influencing each other. Second, determining the optimal weighting for the linear combination of metrics depends on the chosen training corpus. A new text only works well with this system if it has a similar writing style as the texts in the training corpus. Finally, Marcu's combined system of different heuristics still produces n different output trees D_i, each ranked with a score $s(D_i)$, whereas the current system produces only one tree structure. This is possible because ambiguities are resolved locally.

Furthermore, other approaches to discourse parsing adopt the claims of one formal framework (e.g. RST) regarding open attachment sites. The claims different theories make about which segments are still accessible for a specific discourse relation vary a lot. The rhetorical relation *Contrast*, for instance, is defined as multi-nuclear within RST. A consequence of this feature is that the content of both segments linked by the relation are combined. SDRT, on the other hand, defines for the relation *Contrast* that the second segment closes off the previous discourse. Given this theoretical framework, only the second segment would be accessible for the following discourse. To my knowledge no empirical study has been carried out to investigate this issue. The system described in this paper offers an ideal means for testing the different predictions different theories make.

Since the proposed system is embedded into a formal discourse theory that uses underspecification, the parse of an input text can never fail. The output may not be very specific when only a few constraints (e.g. only discourse markers) are considered. Future research, however, can add further constraints in a later version of the system. In the meantime, the current system disambiguates a highly underspecified discourse tree structure by employing robust techniques from IR.

More specifically, the algorithm works like this: first, some discourse structure is derived based on discourse markers. However, empirical studies of discourse structure show that only around 37% of the relations in a full analysis are signalled by discourse markers (Soria and Ferrari 1998). Hence the rest of the discourse structure remains *underspecified*. Secondly, discourse units are scored according to their topicality. A topical sentence is a sentence with a high concentration of important words and/or at a prominent position in the text (Edmundson 1969). The partly underspecified discourse structure is then fully specified with the help of the topicality scores: more topical sentences are placed higher in the discourse tree than sentences with a low topicality score.

The topicality score is obtained by computing the similarity of each discourse unit with the title and/or the lead text (i.e. the first paragraph of a news article printed in bold). Applying the vector space model to the given task of discourse parsing, the title and each sentence are represented as vectors in an n-dimensional vector space. As for a classical IR application, the task here is to find the most relevant documents (i.e. sentences) within a collection of documents (i.e. the text)

for a given query (i.e. the title). Additionally, the topicality score is adjusted for those sentences that occur in a prominent position where important sentences are expected, following the position method.

The ranking of the topical sentences adds further constraints to the partly underspecified discourse structure and as a result a fully specified tree structure is derived. The tree structure reflects the ordering of the sentences according to their importance – essentially, the most topical sentences become dominant over the less topical ones as long as the discourse structure derived from the discourse markers allow this.

The proposed algorithm is implemented in a system that takes as an input a text and produces as an output the discourse structure based on the given constraints. The system can also be run with a different set of parameters. For example, it is possible to run the system only on the topicality scores or only with respect to the position method. The system is in particular useful for experimenting with current discourse theories. The different predictions theories make regarding discourse attachment could be tested by varying the constraints certain discourse markers impose.

The algorithm described in this paper was evaluated with respect to its performance on the text summarisation task by comparing its results with an empirically derived 'gold standard', and comparing its scores with the scores of other summarisation systems. The discourse parser was used to extract the set of most important sentences from German news articles. This set of importantly marked sentences could serve as input for a text summarisation program (cf. Marcu 1999b). Results show that the algorithm performs considerably better than three different baselines that are each based solely on one of the proposed constraints (i.e. discourse markers, topicality score or position). The proposed system also outperforms two other summarisation systems (i.e. *Extractor* (Turney 2000) and Microsoft Office 98 summarizer). In addition, I compared the analysis of a short English text by Marcu (1999b) with the output for this same text produced by the current system.

The algorithm described in this paper can be seen as a deductive approach to the derivation of discourse structure. Past formal theories start with the semantic content of the sentential structure and build up a hierarchical discourse structure by consulting world knowledge. The current approach takes the opposite direction and acquires clues from discourse markers, topicality scores and formatting cues to further specify the at first totally underspecified discourse tree. Moreover, the proposed approach combines insights from formal approaches with practical methods developed for IR application in order to obtain a robust rhetorical parsing system.

2 Background

This section provides the necessary background information regarding the proposed algorithm.

2.1 Discourse grammar and discourse markers

Underspecification formalisms (e.g. Reyle 1993; Bos 1995) have become increasingly popular in the field of semantics to describe scope ambiguities of multiple quantifiers. Recently, underspecification formalisms have also been applied to discourse

grammars (Asher and Fernando 1997; Gardent and Webber 1998; Schilder 1998). In the following, I first provide the motivation for using underspecification for discourse grammars, and then give the formal details regarding Underspecified SDRT (USDRT) (Schilder 2000).

2.1.1 Underspecification

Underspecification formalisms provide a formal system that can be used for the concise representation of more than one reading for an ambiguous sentence such as (1):

(1) Every man loves a woman.

The logical form may be

a. $\forall x \exists y\ man(x) \rightarrow (woman(y) \wedge love(x, y))$, or
b. $\exists y \forall x\ man(x) \rightarrow (woman(y) \wedge love(x, y))$

Within an underspecification formalism such as that proposed by Bos (1995), a representation is derived that leaves the ordering between the two quantifiers open.

Similarly, discourse grammars have been developed that also allow underspecification. Here the scope of the to-be-derived rhetorical relations may be left open. Consider (2):

(2) (a) I try to read a novel (b) if I feel bored or (c) I am unhappy. (Gardent and Webber 1998)

The discourse in (2) is ambiguous with respect to the expressed discourse structure. Either the speaker tries to read a novel provided one of the two conditions in (b) and (c) hold, or being unhappy is the alternative to not reading a novel. As Gardent and Webber (1998) show, these two readings can be represented by leaving the structural relations between scope bearing discourse relations underspecified. A formal representation is given in Figure 1. A tree logic is used to represent several trees in one representation (i.e. forest) instead of one tree for each reading, using dominance constraints on node labels.

Such constraints on node labels are imposed, indicating the strict dominance relation or the dominance relation, which is transitive. The strict dominance relation (i.e. parent relation) is drawn with a straight line, whereas the dominance relation is indicated by the dotted line.

Have a look at the discourse in (2) again. The first argument of the second discourse relation could be filled by two discourse units: (a) the complex discourse unit in (2a) and (2b) or (b) only the one in (2b). An underspecified representation names only these two possibilities for the argument of the discourse relation via dominance constraints. Formally, the underspecification is expressed by a 'hole' K added to the first argument of the second discourse relation, which can be filled by two different 'plugs'. The two readings are covered via the two dominance relations that hold between K and the two conceivable arguments. This forest representation

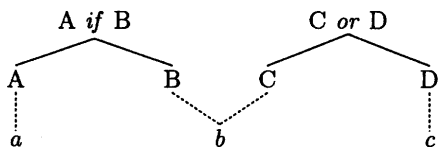

Fig. 1. The underspecified discourse structure for (2).

in Figure 1 can give rise to the following specified readings: (i) *a if (b or c)*, or (ii) *(a if b) or c*.

Apart from capturing scope ambiguities, underspecification can also be used for robust discourse processing. In addition to the underspecification of discourse relation scope, the actual discourse relation that links two segments in a text can be underspecified. A discourse grammar that has this feature is described in more detail in the following section.

2.1.2 Underspecified SDRT

Within Underspecified SDRT (USDRT), similar to other approaches to underspecification, the underspecification between the sub-formulae (i.e. Segmented Discourse Representation Structures (SDRSs)[1]) is expressed by the immediate dominance relation (i.e. $\lhd$) and the dominance relation ($\lhd^*$) that hold between node labels (Schilder 2000). Formally, an Underspecified SDRS (USDRS) is defined as follows:

Definition 1 (USDRS)
Let S be a set of DRSs, L a set of labels, $\mathscr{R}$ a set of discourse relations. Then U is a USDRS confined to the tuple $\langle S, L, \mathscr{R} \rangle$ where U is a finite set consisting of conditions of the following form:

- *structural information*
  - — immediate dominance relation: $K_1 \lhd K_2$, where $K_1, K_2 \in L$
  - — dominance relation: $K_1 \lhd^* K_2$, where $K_1, K_2 \in L$
  - — precedence relation: $K_1 \prec K_2$, where $K_1, K_2 \in L$
  - — equivalence relation: $K_1 \approx K_2$, where $K_1, K_2 \in L$
- *content information*
  - — sentential (i.e. universe): $s_1 : \alpha$, where $s_1 \in L, \alpha \in S$
  - — segmental (i.e. conditions):
    - – discourse relation(s) connecting two segments:
      $K_{R1} : relation(\mathscr{P}, K'_{R1}, K''_{R2})$, where $\mathscr{P} \subseteq \mathscr{R}$, and K_{R1}, K''_{R1}, and $K''_{R1} \in L$
    - – topic information: $K^{\mathscr{T}}_{R1} : \mathscr{T} \subseteq \{\alpha, \beta\}$

The definition distinguishes between structural and content information. The structural constraints determine how the labels are related to each other, determining the scope information for the relations. However, how the constraints are realised

[1] The semantic content of a sentence would be represented by a DRS, larger sequences by an SDRS.

for a given discourse structure depends on the derived discourse relation(s). An open discourse relation, such as *Elaboration*, keeps former segments open for attachment and comes with a different set of constraints than, for example, *Narration*.

The content information described by the definition is derived from the basic discourse units (i.e. sentences or clauses) and the discourse relation(s). It is important to note that in contrast to standard SDRT, a discourse relation *set* $\mathscr{P}$ provides the link between (S)DRSs. The set $\mathscr{P}$ is introduced by Schilder (2000), because former approaches to underspecification of discourse (Asher and Fernando 1997; Schilder 1998) do not provide an appropriate formalisation for the underspecification of the discourse relations. These approaches deal with an underspecified discourse relation in the same way as scope ambiguity by introducing a condition such as $R =?$. However, there is a crucial difference between these two forms of ambiguity: scope ambiguity can easily be resolved by computing all combinations of scope-bearing operators. The discourse relation, on the other hand, cannot be resolved by determining the scope of all relations. Instead, the relation has to be inferred from world knowledge and the information provided by context. Within the standard SDRT account, for instance, this relation has to be obtained by considering world knowledge as well as additional discourse knowledge by applying a non-monotonic reasoning system called DICE (Lascarides and Asher 1993).

However, since formalising world knowledge and context information is not feasible for a robust system, another route is taken by USDRT. Within USDRT the discourse relation set $\mathscr{R}$ contains at first all conceivable discourse relations. In contrast to other approaches to discourse processing, the discourse relation does not have to be determined while constructing the discourse tree. In the absence of any further cues, the discourse relation between two segments may be totally underspecified (i.e. $\mathscr{P} = \mathscr{R}$). After taking into account further restrictions, the discourse relation set $\mathscr{P}$ can be restricted to only a subset of $\mathscr{R}$.

Constraining clues for $\mathscr{R}$ can come from discourse markers such as *but, however* or *because*. In order to ensure robustness for the further specification of a more restricted set of discourse relations, a discourse relation lattice is employed, as in Schilder (2000) (see Figure 2).[2]

Evidence for the idea that a set of discourse relations can connect two segments has recently been provided by research on multiple discourse markers by Oates (2000). She collected corpus data that provide evidence that a 'strong' discourse marker has to follow a 'weak' discourse marker restricting the set of possible discourse relations even further:

(3) The pores in the skin are a classic example: they cannot become perceptible to us by themselves, but yet (but/yet/*yet but) their presence in the skin can be deduced from sweat. (BNC)

The marker *but* signals a set of discourse relations (i.e. {*Contrast, Concession, Antithesis, Exception*}) and the marker *yet* refers to a more restrictive sub-set (i.e. {*Contrast,*

[2] For the sake of readability, I present only a small lattice containing only four relations (i.e. *Cont(rast), Conc(ession), Anti(thesis)* and *Ex(ception)*).

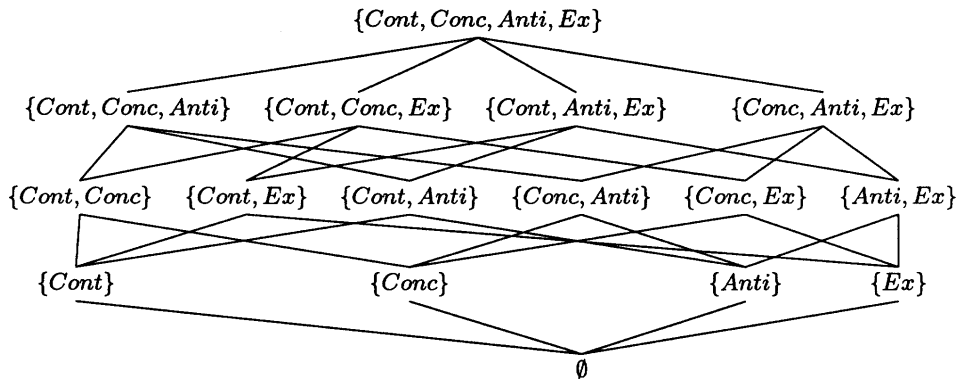

Fig. 2. The discourse relation lattice for four discourse relations.

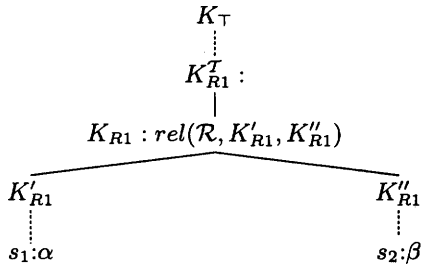

Fig. 3. Underspecified discourse structure.

Concession, Antithesis}). The discourse relation set lattice in Figure 2 reflects the different levels of more specific discourse relation sets.

The starting point of any derivation is a structure such as that shown in Figure 3. Two situations are connected by the set $\mathcal{R}$ of all discourse relations.

Note that the topic node dominates the node with the discourse relation(s). This representation is chosen, because the topic of this segment is the argument of another discourse relation, but it is not the discourse relation from that segment.

The topic node $K^{\mathcal{T}}_{R1}$ is further determined during the derivation process of a more specific discourse relation. For example, $\mathcal{R}$ can be restricted to {*Explanation*} due to the discourse marker *because*, and in this case the topic node contains only α.

2.1.3 Topics

Discourse grammars with underspecification reflect the fact that texts are not necessarily fully specified with respect to their discourse structure. However, it is not clear how such underspecified approaches describe the semantic content of an entire text (i.e. the topic of the text). Normally, the derivation of one possible discourse structure results in a detailed description of the whole text in the root node. No abstraction takes place as in the analysis of (2) by Gardent and Webber (1998). Such an abstraction may not be necessary for a short discourse such as in (2), but becomes more urgent when longer sequences are to be analysed.

Within SDRT a so-called topic feature is introduced. A common topic for two segments is demanded, for instance, when they are connected via *Narration*. However, how this common topic is derived is not explained by the formalism.

The same issue is addressed from a more applied angle by research projects on keyword extraction or topic detection. Systems for keyword extraction such as *Extractor* use a genetic learning algorithm and some linguistic knowledge to derive a set of keywords for a given text (Turney 2000).

2.2 Vector space model and topicality

This section gives a brief introduction to the Vector Space Model (VSM) before it is applied to the derivation of the topicality score.

2.2.1 Vector space model

The vector space model is a very robust approach to IR applications (Salton 1971). Given a collection of documents, many IR tasks require the retrieval of a set of documents that match a user query. The VSM sees documents as bags of words and generates for each document an n-dimensional vector representing the content of the document. These vectors consist of all the terms that occur in the collection. In a document vector a value is given for every single term depending on whether this term is present in the document or not.

The values are normally weighted according to the $tf \times idf$ weighting. A weight for a term i in a document j is determined by the term frequency (i.e. $tf_{i,j}$) in a given document times the inverse document frequency (i.e. idf_i) (Jones 1972).

To compute the set of relevant documents for a given query, all documents are represented as n-dimensional vectors. The query is also represented as an n-dimensional vector. A similarity-function determines how close the vectors are within the n-dimensional vector space by computing the angle between the two vectors. More precisely, the cosine for a document vector $\vec{d}$ and a query vector $\vec{q}$ is computed as follows:

$$sim(\vec{q_k}, \vec{d_j}) = \frac{\Sigma_{i=1}^{N} w_{i,k} \times w_{i,j}}{\sqrt{\Sigma_{i=1}^{N} w_{i,k}^2} \times \sqrt{\Sigma_{i=1}^{N} w_{i,j}^2}} \quad (1)$$

Since the formula in equation (1) computes the cosine for the angle between the two vectors, the following interpretation of this formula is possible: if the two documents are identical, the cosine will be 1; if the two documents do not share any common terms, the cosine will be 0 (i.e. the two vectors are orthogonal).

2.2.2 Topicality scores

For the proposed algorithm topicality scores are derived within the VSM. This is done by transferring the IR task described in the previous section from a document collection to a text. Instead of a collection of documents we have a collection of sentences and clauses and the query for the retrieval task is the title (and possibly a lead text).

The justification for doing this comes from the following observation: the title and the lead text are already a concise summary of the text, covering the most important topics. Presumably the author of a well-written text has organised that text so that the most topical sentences are salient in the discourse structure. Since it is still an open research question as to how the topicality of a sentence or segment can be determined compositionally within a model-theoretical framework, the topic score for each sentence and clause is computed within the VSM.

For a given title vector $\vec{t_k}$ and a discourse unit vector $\vec{du_j}$ the topicality score $top(\vec{t_k}, \vec{du_j})$ has to be computed. However, the formula (1) cannot be so easily adopted for this task. For example, Marcu (1998) in his more elaborated system that also allows to take into account certain heuristics, uses a cosine metric. This metric is not appropriate for the task of determining a topical discourse unit for two reasons.[3] First, note that the vectors in (1) are normalised (i.e. $\sqrt{\Sigma_{i=1}^{N} w_i^2}$). Within a standard IR task this makes sense since the normalisation eliminates the exact length of the vector in the n-dimensional vector space. Instead, only the direction of the vectors are considered. Although normalising the vectors seems to be right for a document retrieval task, this is not the case for the comparison between a sentence and the title in order to derive a topicality score. This score has to be high when there is a high concentration of keywords that are crucial for the text. A sentence with the same number of keywords but with a higher word count should not obtain the same topicality score. Instead, the score should be lower for the longer sentence. The idea behind this way of defining topicality is that a topic sentence is a sentence that expresses central concepts to the article in a concise way. A longer sentence that shares the same number of keywords also expresses additional concepts that elaborate further. Such sentences are not expected to be as high in the discourse tree. They are probably found at a more intermediate level of the tree.

Hence the dot product between vectors is used to compute an adequate topicality score. The topicality score between the title t_k and a discourse unit du_j is derived as follows:

$$top(\vec{t_k}, \vec{du_j}) = \Sigma_{i=1}^{N} w_{i,k} \times w_{i,j} \tag{2}$$

The second reason that a cosine metric is inappropriate for determining a topic score for a discourse unit is that the weighting of the terms has to be reconsidered for the task at hand. It seems questionable to use the inverse document frequency idf. This frequency is normally used to penalise terms that occur often in the overall document collection. In a text, however, frequent terms are often an indicator for topical keywords. Thus the weights $w_{i,j}$ should consist only of the term frequency regarding the discourse unit du_j.

After determining scores for all sentences and clauses in the text, a partially ordered list of topical sentences and clauses can be obtained. These scores can then be used to further specify a partially underspecified discourse parse, as already described.

[3] Two of Marcu's metrics are very similar to the constraints used for the proposed system (i.e. the title- and position-based metric).

2.3 Position method and term adjustment

Finally, further formatting cues are taken into account. According to the position method, sentences at the beginning of a paragraph are considered to be important. I will show how this feature can be incorporated into the vector representation. Moreover, I will provide a general guideline for attachment when there is no discriminating topicality score to be derived between the discourse units.

2.3.1 Position method

Many standard IR approaches already assume that certain positions in a text bear important information, following Edmundson (1969). For example, the first sentence of a paragraph is usually a good candidate as a sentence that describes the topic of this segment. Lin and Hovy (1997), for example, show that within certain genres important sentences are often located at the beginning or the end of a paragraph.

Consequently, it seems useful to consider this constraint for the derivation of the discourse structure. Certain sentences should benefit from their position and obtain a better topicality score than sentences in the middle of a paragraph. Thus, the vector representing a sentence or clause is extended by another term indicating whether the sentence occurs at a prominent position or not.

2.3.2 Low attachment

Even after adding the position as another term to the vector representation, there may still be discourse units that possess the same topicality score. This is in particular the case when the sentences do not share any terms with the title and the lead nor occur in a prominent position in the text. In such a case, the assumed default is to attach low to the previous discourse. This disambiguation strategy can be explained by the observation that in texts the most important ideas are usually mentioned first and subsequently elaborated on.

2.4 Combining the constraints

In this section I will show how the different constraints can be combined and how they interact with respect to an example text in Figure 4.[4] Bear in mind that I do not assume a linear combination of the constraints, as proposed by Marcu (1998). Instead, I assume the following ordering of constraints: *discourse marker* > *topicality score (+ position)* > *low attachment*.

2.4.1 Discourse structure and discourse marker

Following Schilder (2000), a totally underspecified discourse structure is assumed. The text in Figure 4 contains 10 sentences that are connected by nine underspecified

[4] These two paragraphs, taken from a longer article in *Scientific American*, have been analysed by Marcu (1999b) and used for the evaluation of his summarisation system. I added the title and lead text from the original text.

(t) Global Climatic Change on Mars

(*l*) Today a frozen world, Mars at one time may have had more temperate conditions, with flowing rivers, thawing seas, melting glaciers and, perhaps, abundant life
$(1)_{(0.0059)}$ *With* its distant orbit – 50 percent farther from the sun than Earth – and slim atmospheric blanket, $(2)_{(0.0222)}$ Mars experiences frigid weather conditions. $(3)_{(0.0028)}$ Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles. $(4)_{(0.0044)}$ Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, $(5)_{(0)}$ *but* any liquid water formed in this way would evaporate almost instantly $(6)_{(0)}$ *because* of the low atmospheric pressure.
$(7)_{(0.0067)}$ *Although* the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, $(8)_{(0.0033)}$ most Martian weather involves blowing dust or carbon dioxide. $(9)_{(0.0049)}$ Each winter, *for example*, a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap. $(10)_{(0.0028)}$ *Yet* even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water. (From Scientific American)

Fig. 4. Example text plus title and lead.

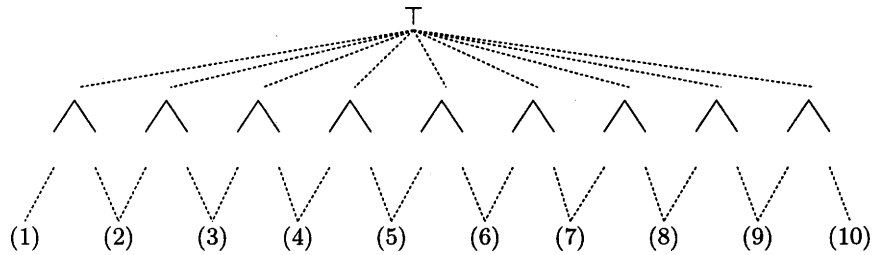

Fig. 5. The discourse structure totally underspecified.

discourse relations. The resulting underspecified discourse structure can be found in Figure 5.

Without taking into account any further constraints the system would produce this totally underspecified discourse structure as output. The following constraints can now be added until a fully specified discourse tree derives.

Most importantly, discourse markers that signal discourse relations are extracted during the parse. This information is used to restrict the totally underspecified discourse structure. The marker *because* linking (5) and (6), for instance, signals an *Explanation* relation. Since *Explanation* is defined as an open discourse relation by SDRT, both segments have to be available on the right frontier. As a consequence, the discourse unit in (5) fills the topic node. See Figure 6 for the discourse structure that evolves after taking into account the available discourse markers.[5] Although I follow the SDRT definition for openness for this example, I would like to emphasise that further tests are necessary. In SDRT the *Contrast* relation, for instance, only

[5] All discourse markers for the text in Figure 4 are marked by bold and italic fonts following Marcu (1999b).

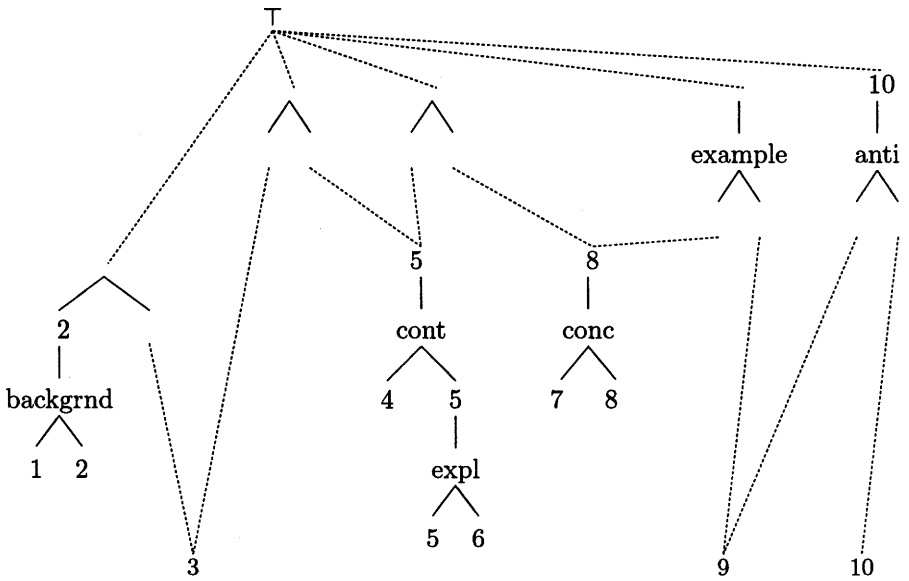

Fig. 6. The discourse structure underspecified.

allows access to the last segment, whereas in RST the relation is multi-nuclear (i.e. both segments are open for attachment). An underspecified discourse theory such as USDRT is an ideal test bed for experimenting with different definitions for openness.

2.4.2 Topicality, position and low attachment

The next constraint imposed on the partially underspecified discourse structure comes from the topicality score. Vector representations are derived for each discourse unit according to the formula in (2). The topicality score is computed by comparing the vector for a discourse unit with the vector representing the title, and if present, the lead of the text.

Given the topicality scores for each discourse unit, the topical clauses can be spotted and ranked. A high score indicates high topicality. Within the discourse structure, a high topicality score is usually reflected by a high position in the discourse tree.

However, some texts may not have their structural ordering signalled by formatting cues such as titles or lead texts. In such a case, and in order to include formatting cues, the first sentence of a paragraph is marked as a sentence containing some topical information. Remember that according to the Position Method a good location for a topical sentence is the beginning of a paragraph. In order to incorporate this into the topicality score for each discourse unit, the vectors representing the clauses (1)–(2) and (7)–(8) get an additional term that stands for the feature of being first sentence in a paragraph. The title and lead vector also contain this term.

The final discourse structure of the text in Figure 4 is generated as follows: based on the ranking of the discourse units, the discourse structure, where it is still

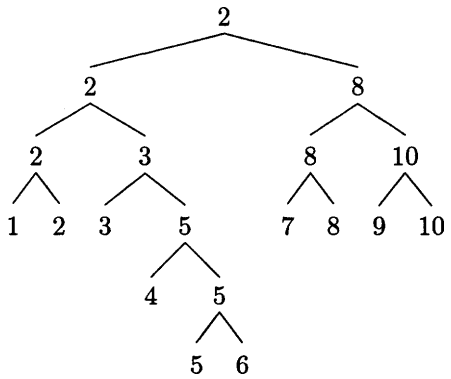

Fig. 7. The fully specified discourse tree for the example text.

underspecified, is further determined.[6] The first two units that need to be specified are (3) and (2). Since (3) scores only 0.0028 and (2) scores 0.0222, the resulting discourse structure has (2) dominating (3). When the units earn the same score the next unit is attached low in the tree.[7] Continuing this procedure with the remaining units, the discourse structure in figure 7 is derived, which coincides with the discourse tree derived by Marcu.

It is important to note that the discourse marker *Although* in (7) prevents the unit (7) from ending up higher in the tree than unit (8). Although the unit in (7) has a higher topicality score than the one in (8), the discourse structure already determined by the discourse marker wins.

3 Implementation

The approach to discourse parsing is implemented as a system that takes an HTML or Text file[8] as input and generates as output the preferred rhetorical structure of the text. The discourse structure disambiguation is done locally within one parse according to the constraints described above.

In the following I will show how the discourse can be parsed via context-free rules. Based on these rules a syntax/semantics interface is defined that precisely determines the effects discourse syntax can have on the discourse semantics.

3.1 Discourse syntax

Recently, Webber *et al.* (1999) presented a discourse grammar based on Lexicalised Tree Adjoining Grammar (LTAG). Within this approach discourse parsing is seen

[6] The topicality scores for each unit in Figure 4 are given as an index.

[7] The low attachment default is not required for the example text in Figure 4 because the order in the tree is either determined by the topicality score or by discourse markers (e.g. unit (5) and (6)).

[8] A decision-tree-based POS tagger developed by Schmid (1994) was integrated into the system. The tagger is used to produce an XML-file that reflects the POS-information (see Figure 8). The title and lead information is still tagged by hand. However, it should not be so difficult to develop an automatic tagger that extracts this information and to then incorporate it into the present system.

Fig. 8. A snapshot of the discourse parsing system.

as an extension of sentence parsing going beyond the sentence level. The semantic effects that are imposed by discourse markers are modelled as anaphoric links often invoking defeasible inferences. In contrast to other approaches (e.g. RST, SDRT etc.) the approach by Webber *et al.* (1999) abandons the idea that the semantic discourse structure is built by rhetorical relations. Instead the focus is on the *syntactic* parsing of discourse.

For the implementation I adopted the view that discourse can be parsed much like a sentence, by giving a grammar of context-free rules. However, based on the derived syntactic structure I still assume a discourse semantic structure. This structure is modelled as a discourse tree that satisfies the constraints imposed by a USDRS defined earlier.

Thus, I first programmed several context-free rules within DCG notation that is supported by the PROLOG system.[9] See Figure 9 for some example rules.

3.2 Derivation of a discourse tree

The derivation of the rhetorical structure is done in parallel to the syntactic parse via the coupling of the discourse syntactic with the discourse semantic structure. First,

[9] The system is programmed in SWI-PROLOG and XPCE (see `http://www.swi-prolog.org`). The discourse syntactic and semantic output trees are visualised by the daVinci visualisation software v2.1.

```
d(d/[S], Sem) -->
      s(S, Sem).

d(d/[S,D], Delta) -->
      s(S, Alpha),
      d(D, Beta),
      {spec_discourse(Alpha, Beta, Delta)}.

s(s/[CL], Sem) -->
        cl(CL,final,Sem).

s(s/[CL, S], Delta) -->
         cl(CL, Alpha),
         s(S, Beta),
         {spec_discourse(Alpha, Beta, Delta)}.
```

Fig. 9. Some simplified discourse syntax rules.

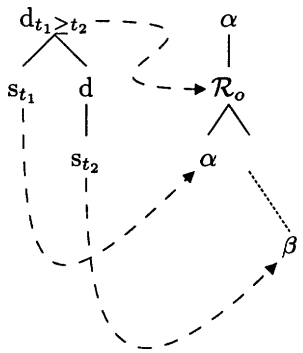

Fig. 10. The discourse syntax/semantics interface for discourse units without discourse markers.

a discourse syntax tree for two sentences is derived using the rules specified earlier. Then the discourse semantic tree is generated by a precise interface definition.[10] For example, the syntactic structure for a two sentence discourse is reflected by the left tree in figure 10. In case the topicality score t_1 for the first sentence is higher than the score t_2 for the second sentence, an open discourse structure is derived. The right tree in figure 10 indicates a discourse semantic tree that is open: $\mathscr{R}_o$ is the set of open discourse relations, α and β are the semantic representation of the first and the second sentence, respectively.

If a discourse marker triggers a certain discourse relation, two types of connections to the discourse have to be considered.[11] Depending on the discourse marker, the

[10] The proposed interface is inspired by Kallmeyer (1997), who defines a syntax/semantics interface for Tree Description Grammar at the sentence level.

[11] The set of German discourse markers the system uses stems mainly from Rehm (1998). For the analysis of English text a very limited set of discourse markers was sufficient for the example texts analysed.

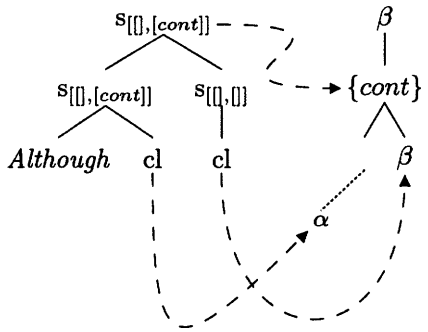

Fig. 11. The discourse syntax/semantics interface when discourse markers are present.

discourse relation is a link to the previous or the following discourse segment. The sequences in (4) exemplify these two possibilities.

(4) a. **Although** the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.
b. (...) any liquid water formed in this way would evaporate almost instantly **because** of the low atmospheric pressure.

In (4a) the discourse marker *although* indicates a discourse relation *Contrast* that provides a link to a following discourse unit. In (4b), on the other hand, a discourse relation *Explanation* is derived that is a link to a previous discourse unit. During the parse this information is stored in a two-element-long list that is passed up the syntactic tree structure. When the discourse relation is bound, the relation is removed from this list. In Figure 11, the rhetorical relation *Contrast* is put into the forward looking slot of the list (i.e. the second argument slot) until the relation is removed at the first sentence node that has a right daughter.

Finally, the system opens two new windows with the discourse syntactic and the semantic tree. Figure 12 shows such a window indicating a small part of a discourse semantic output tree. Three discourse relations are visible. One indicates a closed discourse relation because the topicality score of the second discourse unit (i.e. *s*9) is higher than the first one (i.e. *s*8). In addition, an *Explanation* relation is derived between *s*10 and *s*14 due to a discourse marker.

4 Evaluation

The method of determining the discourse structure of a text was evaluated via a summarisation task that involved extracting the most salient sentences of a text. An experiment was carried out that required participants to read a newspaper article and to underline the main ideas of the text.

4.1 Experimental design

Instructions 71 computer science students participated in the experiment. They were asked to carefully read a newspaper article at their normal reading speed. They then had to underline the sentences or parts of the sentences that represented the core

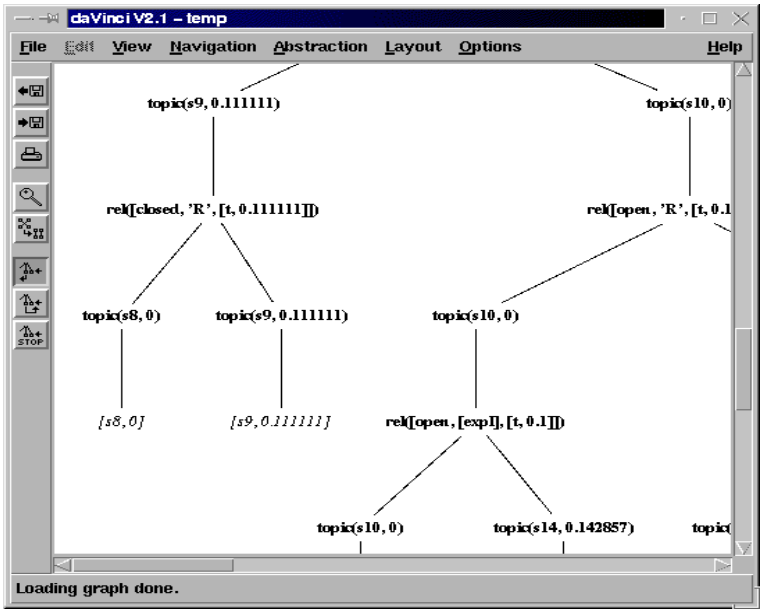

Fig. 12. A part of a discourse semantic output tree.

meaning of the article. The participants were instructed to only underline sentences in the main part of the text and not the title or the lead. The task was limited to 4–7 minutes depending on the length of the article.

Materials Five different articles from German newspapers and an online-information service were read. Articles ranged from 8–26 sentences and contained between 187 and 475 words. Each article was read by at least 24 Participants. All articles were economic news articles.

4.2 Results

The underlined sentences were counted for each article. Sentences that were underlined by more than half of the participants were considered to reflect the core ideas of the text. Those sentences are also most likely to show up in a summarisation.

The results of the experiment were compared with the outputs of four baseline algorithms, the keyword extraction program *Extractor* (Turney 2000) (employs a genetic learning algorithm),[12] a commercial program (Microsoft Office 98 summarizer), and the proposed algorithm. Table 1 shows the results as F-measures. The F-measure (F) is a combined value of precision (P) and recall (R) values and is computed by[13]

$$F = \frac{2PR}{P+R}.$$

[12] The demo web page of *Extractor* also extracts a set of 'highlights' of the processed texts. This output was used for the evaluation.

[13] Since the two programs, *Extractor* and Microsoft Office 98, do not recognise the title and the lead text, these sentences were added to the set of sentences to be highlighted by the program. Recall and precision rates were calculated accordingly.

Table 1. *F-measures for systems and baselines*

	Random	*Office 98*	Topic	Disc.	Position	*Extractor*	Discourse & Topic & Position
F-Meas.	34.00	35.19	36.82	46.50	47.85	49.99	**68.46**

The baseline algorithms are based on the following ideas:

Random A program was run for 1000 times that produced a random selection of sentences (Marcu 1999b).

Discourse Only the discourse markers were considered. New sentences that did not contain any discourse markers were attached low in the discourse tree. Marcu's original rhetorical parser mainly uses this strategy for disambiguating discourse trees (Marcu 1997).

Topic The first *n* most topical sentences, according to the topicality score, were selected.

Position The first sentence of each paragraph was picked for the set of most important sentences.

The three last baselines correspond to the three constraints considered by the proposed algorithm. However, in the current system the tree structure indicated by the discourse markers cannot be overruled and the two other constraints are used to further specify the partly underspecified discourse tree that is obtained only by discourse markers.

4.3 Discussion

The results of the experiment show that the proposed algorithm performs better than the baseline algorithms and two other programs for keyword extraction/text summarisation. In particular, the results for the baselines show that the combination of all constraints is crucial for a high recall and precision value.

5 Conclusions

This article proposed a robust algorithm to discourse processing. A formal discourse grammar employing underspecification was combined with IR techniques such as the Vector Space Model and the Position Method. Discourse markers determine at first a partly underspecified discourse structure that gets further specified using a topicality score. Discourse units are scored according to their topic centrality, as defined by the similarity between the discourse unit and the title (and if present sub-titles and/or lead text). The obtained scores give a crucial clue for the full determination of the discourse structure. However, a high topicality score does not necessarily mean that the discourse unit ends up high in the discourse structure. The

discourse structure derived from discourse markers (e.g. *because*) make sure that this does not happen.

A system was implemented that possesses a precise discourse syntax/semantic interface. The output of the system was evaluated with respect to an IR task, namely summarisation. The system outperforms the baseline algorithms and two other keyword extraction/summarisation programs. In addition, the results suggest that the algorithm is an improvement over other approaches to text summarisation that also consider discourse structure.

References

Asher, N. (1993) *Reference to Abstract Objects in Discourse. Studies in Linguistics and Philosophy 50*. Kluwer Academic, Dordrecht.

Asher, N. and Fernando, T. (1997) Labeling representations for effective disambiguation. *Proceedings 2nd International Workshop on Computational Semantics (IWCS-II)*, pp. 1–14. Tilburg, The Netherlands.

Bos, J. (1995) Predicate logic unplugged. In: Dekker, P. and Stokhof, M., editors, *Proceedings 9th Amsterdam Colloquium*. Amsterdam, The Netherlands.

Corston-Oliver, S. H. (1998) Beyond string matching and cue phrases: Improving efficiency and coverage in discourse analysis. *AAAI Spring Symposium on Intelligent Text Summarization*, pp. 9–15. AAAI.

Edmundson, H. P. (1969) New methods in automatic extracting. *J. ACM*, **16**(2).

Gardent, C. and Webber, B. (1998). Describing discourse semantics. *Proceedings 4th TAG+ workshop*, Philadelphia, USA.

Jones, K. S. (1972) A statistical interpretation of term specificity and its application in retrieval. *J. Documentation* **28**: 11–21.

Kallmeyer, L. (1997) A syntax-semantics interface with synchronous tree description grammars. *Proceedings Formal Grammar conference*, pp. 112–124. Aix-en-Provence, France.

Kamp, H. and Reyle, U. (1993) *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language. Studies in Linguistics and Philosophy 42*. Kluwer Academic, Dordrecht.

Lascarides, A. and Asher, N. (1993) Temporal interpretation, discourse structure and commonsense entailment. *Linguistics and Philosophy* **16**: 437–493.

Lin, C. and Hovy, E. (1997) Identifying topics by position. *Proceedings 5th Conference on Applied Natural Language Processing (ANLP–97)*, pp. 283–290.

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

Marcu, D. (1997) From discourse structures to text summaries. *Proceedings ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp. 82–88. Madrid, Spain.

Marcu, D. (1998) Improving summarization through rhetorical parsing tuning. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pp. 206–215. Montreal, Canada.

Marcu, D. (1999a) A decision-based approach to rhetorical parsing. *Proceedings 37th Annual Meeting of the ACL*, pp. 365–372. Maryland, USA.

Marcu, D. (1999b) Discourse trees are good indicators of importance in text. In: Mani, I. and Maybury, M., editors, *Advances in Automatic Text Summarization*, pp. 123–136. MIT Press, Cambridge, MA.

Oates, S. L. (2000). Multiple discourse marker occurrence: Creating hierarchies for natural language generation. Technical Report ITRI-00-03, Information Technology Research Institute (ITRI), University of Brighton. (Also published in *Proceedings ANLP-NAACL Student Research Workshop, ANLP-NAACL*, 2000.)

Rehm, G. (1998) Vorüberlegungen zur automatischen Zusammenfassung deutschsprachiger Texte mittels einer SGML- und DSSSL-basierten Repräsentation von RST-Relationen. Magisterarbeit, Sprach- und Literaturwissenschaft, Universität Osnabrück.

Reyle, U. (1993) Dealing with ambiguities by underspecification: construction, representation, and deduction. *J. Semantics* **10**: 123–179.

Salton, G., editor (1971) *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall, New York.

Schilder, F. (1998) An underspecified segmented discourse representation theory (USDRT). *Proceedings 17th International Conference on Computational Linguistics (COLING '98) and of the 36th Annual Meeting of the Association for Computational Linguistics (ACL '98)*, pp. 1188–1192. Montréal, Québec, Canada.

Schilder, F. (2000) Robust text analysis via underspecification. *Proceedings Workshop on Robust Methods in Analysis of Natural Language Data (ROMAND 2000)*, pp. 105–120. Lausanne, Switzerland.

Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees. *Proceedings International Conference on New Methods in Language Processing*.

Soria, C. and Ferrari, G. (1998) Lexical marking of discourse relations – some experimental findings. In *Proceedings of the Conference workshop Discourse Relations and Discourse Markers at COLING-ACL'98*, pp. 36–42. Montréal, Québec, Canada.

Sumita, K., Ono, K., Chino, T., Ukita, T. and Amano, S. (1992) A discourse structure analyzer for Japanese text. *Proceedings International Conference on Fifth Generation Computer Systems*, **2**: 1133–1140.

Turney, P. D. (2000) Learning algorithms for keyphrase extraction. *Information Retrieval* **2**(4): 330–336.

Webber, B., Knott, A., Stone, M., and Joshi, A. (1999) Discourse relations: A structural and presuppositional account using lexicalised tag. *Proceedings 37th Annual Meeting of the ACL*, pp. 41–48. Maryland, USA.