Resolving Bridging Descriptions in High-Dimensional Space

Sabine Schulte im Walde
schulte@ims.uni-stuttgart.de

Studienarbeit (Revised Version)
February 20, 1998

Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart
Germany

Supervisor: Prof. Dr. Mats Rooth

Centre for Cognitive Science (CCS)
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Scotland, UK

Supervisor: Dr. Massimo Poesio, Dr. Chris Brew

Contents

1 Introduction

The aim of the thesis[1] is to test the effectiveness of word clustering algorithms for resolving bridging descriptions. Poesio, Vieira and Teufel [11] have been working on a system which automatically resolves bridging descriptions in unrestricted written texts. They use WordNet [1] and its semantic hierarchy as a source for determining the antecedents and their relationships to the descriptions, but found that the information about lexical associations is both very incomplete and often not sufficient. Considering the hypothesis, that the usage of bridging descriptions is triggered by the semantic priming effect of the antecedent, this thesis implements an algorithm by Lund, Burgess and Atchley [7] for creating a high-dimensional space as a model of semantic closeness and semantic priming of words as base for the resolution process, alternative to WordNet.

Section 2 will present the idea in a more detailed way. In section 3 and section 4 we explain the algorithm and the data which serve as basis for the experiments in section 5. Finally, section 6 interprets the results of the method.

2 Background

2.1 Bridging Descriptions

Definite Descriptions (DDs) are the kind of noun phrases starting with the definite article *the*, as in *the lake*. An important property of these noun phrases is that they often refer back to an entity previously introduced in the text, called the *antecedent*, which is used by both the speaker/text and the listener/reader to refer to and establish the content of the description. Here is an example for illustrating the phenomenon:

> "What a wonderful *tree* opposite your house! *The leaves* are of beautiful green!"

For both the person who mentions the sentences and the perceiving person the definite description *the leaves* does not present a problem. Why that? Because mentioning the antecedent *tree* in the preceding sentence justifies the usage, and both persons (consciously or unconsciously) recognise a relation between the expressions *tree* – *the leaves*.

[1] Thanks to Massimo Poesio, Chris Brew, Scott McDonald, Renata Vieira and Richard Thacker for interesting discussions and their comments on previous versions of the thesis.

Hawkins [5] and Prince [12] developed taxonomies for the different types of usage of definite descriptions, depending on the degree of familiarity the definite description expresses in a certain context. Poesio and Vieira [13] reduced these proposals for their task of resolving definite descriptions in written natural language texts to a classification system with three classes:

- **Anaphoric same head:**
  The text explicitly mentions an antecedent for the definite description, with the same head noun as the description, as in

  > "There is still an outstanding *report*. Who is going to write it?" –
  > "As far as I know John wanted to finish *the report* for tonight."

- **Associative description:**
  The text provides an antecedent which the definite description is associated with. This may be a noun phrase, as in

  > "My son is celebrating his *wedding* next week. All his former girl-friends are *the bridesmaids*, isn't that funny?"

  as well as an event, as in

  > "I have carefully *planned* our next burglary. *The strategy* can impossibly cause a mistake."

- **Larger situation / Unfamiliar**:
  The text does not introduce an antecedent for the definite description, whose interpretation is either based on common knowledge, as in *the Iran-Iraq war ...*, or on additional information provided with the description, as in *the fact that ...*

These classes were considered to represent the different ways in which definite descriptions are processed. The class this thesis is concerned with is that of associative descriptions, henceforth called *Bridging Descriptions (BDs)*, so the first step will be to discuss this class in more detail: The resolution process for bridging descriptions, i.e. determining the reference of this kind of definite description, is considered to contain two subtasks:

1. *Finding the antecedent for the bridging description.*
   Considering our first example, the word *tree* has to be determined as the antecedent for *the leaves*.

2. *Determining the association between the antecedent and the bridging description.*
   Considering the example once more, a relationship between *tree* and *leave* has to be found, e.g. the latter is a part of the former.

It is essential to define the possible relation which may hold between antecedent and bridging description and therefore allow the kind of reasoning performed by the reader, since this is the license for the usage of a bridging description.

The following list of relationships is not complete, but represents the ones the system of Poesio, Vieira and Teufel is restricted to, based on a corpus study of bridging descriptions which observes their different processing requirements [14]:

- **Synonymy:**
  The antecedent and the bridging descriptions are synonymous, as in *new album – the record.*

- **Hypernymy/Hyponymy:**
  The antecedent and the bridging description are in a *is-a*-relation, as in *rice – the plant* (super-ordination/hypernymy) or *plant – the rice* (sub-ordination/hyponymy).

- **Meronymy:**
  The antecedent and the bridging description stand in a *part-of* relation, as in *tree – the leaves.*

- **Names:**
  The bridging description refers back to a proper name, as in *Bach – the composer.*

- **Compound Nouns:**
  The antecedent occurs as part of a compound noun, as in *stock market crash – the markets.*

- **Events:**
  The antecedent is not a noun phrase, but either a verb phrase or a sentence, e.g. *planned – the strategy.*

- **Discourse Topic:**
  The antecedent is an (often implicit) discourse topic of a text, as in *the industry* appearing in a text about oil companies.

- **Inference:**
  The bridging description is based on more complex inferential relations, as in *last week's earthquake – the suffering people*.

We integrated the first three relationships of synonymy, hypernymy/hyponymy and meronymy into one single class. They represent the part of the relations encoded in WordNet.

2.2 The System

Poesio, Vieira and Teufel are developing a system with the goal of treating the largest possible subset of definite descriptions in unrestricted written texts. For their task, they make use of linguistic information, but not of knowledge hand-coded for this purpose. The analysis is based on 20 parsed articles of the *Wall Street Journal (WSJ)*, selected at random from the Penn Treebank Corpus [8]. In the corpus, 1040 definite description were identified, 312 anaphoric uses, 492 in a larger or unfamiliar situation, and 204 bridging descriptions, the class we are concerned with. Appendix A is an example WSJ-text, with all bridging descriptions emphasised.

For each of the classes, a strategy for resolution is defined. In resolving the bridging descriptions, the system uses WordNet [1], a publicly available lexical database, as approximation of a knowledge base whose design is based on the results of psycholinguistic research in human lexical organisation and memory. Instead of the definitions in a traditional lexicon, WordNet defines sets of synonymous nouns, verbs, adjectives and adverbs, representing an underlying lexical concept and connected to other sets by lexical relations. Appendix B shows two examples of WordNet-code, hypernyms of the verb *lose* and synonyms of the noun *tree*, each with all different senses.

At the moment, all head nouns from the preceding five sentences are considered as possible antecedents for a bridging description, and the system queries WordNet for encoded lexical relationships. The emphasis is on the relationships of synonymy, hypernymy/hyponymy and meronymy (38 cases in the WSJ-texts), since those are the ones actually defined by WordNet. The following table shows how successfully WordNet finds antecedents based on those relationships:

Class	Total	Found	Not Found
Syn	12	4	8
Hyp	14	8	6
Mer	12	3	9
Total	38	15	23

Proper names are processed by first determining an entity type for each name in the text, e.g. *person*, and then searching for the semantic relation. Some proper names, usually referring to famous entities, are directly encoded in WordNet. For compound nouns, pre- and post-modifiers in noun phrases, in addition to only heads, have to be taken into account. This helps, for example, to resolve *the rules* to the pre-modifier *rule* in the compound noun *rule changes*. Resolving to events presupposes a nominalisation of the verb in order to make it accessible for resolution. A discourse topic can only be found if the antecedent is explicit in the discourse, e.g. if the word *oil* is mentioned in a text concerning oil companies which introduces the bridging description *the industry*. Since WordNet defines several different relationships, some more complex inferences may be found as well.

The total of bridging descriptions resolved with WordNet is 107 out of 204 cases, which corresponds to a recall of 52.5%; 34 of them are correctly resolved to the desired antecedent, which corresponds to a precision of 16.7%.

Since the precision of the system using WordNet is low, it might be improved considering another lexical source. At this point the high-dimensional space comes into play.

2.3 High-Dimensional Space

In recent work in lexical representation, a variety of different approaches has been introduced to model the lexical content of words, build classes of words and define their semantic relationships.

A common idea in this area is *clustering*, described by Charniak [2] as grouping words by defining n relevant properties and giving numerical values for each property. This creates a vector of length n with the n numerical values for each word to be classified, which can be viewed as a point in n-dimensional space. The points in space which are near one another build a class/cluster that reflects commonality of semantic features of the words. The properties used in the vector, the metric used to measure the distance of points (in order to estimate the degree of semantic similarity) and the algorithm used to cluster the points are important issues and open to variation.

The underlying concept of a semantic network goes back to Collins and Quillian [3]: The meaning of a word is represented by a node, embedded within a network of other meanings/nodes. The relations between the nodes are represented by links between the nodes, possibly with each link varying in length to reflect the strength of the relationship. A typical relationship is the *is-a*-link, which describes the sub-ordinated node as a subtype of the super-ordinated one.

Lund et al. [7] utilised high-dimensional space to represent semantic memory. Since words were found to cluster semantically, inter-word distance was interpretable as a measure of semantic similarity. This similarity was shown to be able to account for the *semantic priming* effect, Meyer and Schvaneveldt's robust and important finding in word recognition [10], that the identification of a word is made easier if a word related in meaning is presented just before it. For example, subjects respond faster to the word *doctor* if it is preceded by the related word *nurse* than by the unrelated word *butter* or presented in isolation.

We decided to adopt Lund et al.'s approach to construct a high-dimensional space; the distance between points in the space was utilised to resolve bridging descriptions as described in section 2.1. The idea was to test the hypothesis that bridging descriptions are resolved by a process comparable to semantic priming.

To start with, we describe the general concepts of creating and then interpreting a high-dimensional space in more detail:

2.3.1 *Creating high-dimensional space*

As mentioned before, a high-dimensional space consists of points representing the semantics of words. How were these words chosen in our approach? The words called *target words* are generally determined by the task, i.e. we chose those words in whose semantic content we were interested. In our case, this contained the words occurring in bridging descriptions and in the potential antecedents; these formed the points and clusters in the space.

The next step was to define n properties characterising the n dimensions the target word is described with, to determine the n-dimensional space: Since we were examining the semantic priming effect, we were interested in surrounding words, so the properties were defined by *context words*, i.e. words co-occurring with the target words. In order not to suffer from sparse data, the context words should be high-frequency words.

Now the values of the properties had to be determined. How could we find the numbers describing the co-occurrence of the context words with the target words, i.e. how often they appear close together in a text? Again, in order not to suffer from sparse data, a sufficient amount of data for this task should be considered, otherwise the co-occurrence counts would mostly be zero. The solution for this was to use a large corpus. The process of constructing the vectors can be described as running through the corpus and checking for each of the target words the co-occurrence of any of the context words. In this case, the co-occurrence count for the context word (describing a property of the target word) was increased. We ended up with an n-dimensional vector for each target word, where each dimension was the co-occurrence frequency of a context word with the target word. These vectors were stored in the form of a *co-occurrence matrix* where each row was the vector for a target word, defining a point in n-dimensional space.

Let's consider an example: We are interested in the semantic values for the words *house* and *car*. To determine their vectors in (3-dimensional) space we define the context words *door*, *wall* and *colour*. The co-occurrence matrix displays how often the target words appear in a sample text in co-occurrence with the context words:

	door	*wall*	*colour*
house	81	122	20
car	93	3	45

The table shows that the word *door* appears quite often with both *house* and *car*, *wall* very often with *house*, but almost never with *car*, and *colour* appears a few times with both words, though more often with *car*.

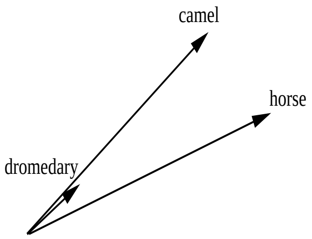
2.3.2 *Interpreting high-dimensional space*

Once we had determined the high-dimensional space we wanted to be able to estimate the semantic similarity of the words. Since semantic similarity is reflected by distance in space, this distance had to be calculated. An enormous number of distance measures had been suggested, including:

- **Manhattan Metric:**
  The Manhattan Metric measures the distance of two points in n-dimensional space by summing the absolute differences of the vectors' elements:
  $d = \sum_{i=1}^{n} |x_i - y_i|$
  An important point to mention here is that the resulting distance d has to be normalised before being compared with other distances. In order to see the need for this, consider the following example from Huckle [6]: Assume the three words *horse*, *camel* and *dromedary*, with vectors

horse	$\langle 32, 12 \rangle$
camel	$\langle 24, 20 \rangle$
dromedary	$\langle 7, 5 \rangle$

The vectors for *camel* and *dromedary* almost point into the same direction:



But since *camel* and *horse* are both high-frequency words, the distance between them will be shorter than between *dromedary* and *horse*, and there is a long distance between *camel* and *dromedary* as well. This result is due to the frequency of the words, not to the similarity in their semantics. To prevent from such mistakes, the co-occurrence counts in the vectors have to be normalised. After summing up the co-occurrences of the target words with the different context words, each count is divided by the frequency of the word:

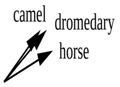
$$normalised\ co\text{-}occurrence\ count = \frac{freq(target, context)}{freq(target)}$$

(where $freq(target, context)$ means target and context word appearing together, i.e. in co-occurrence).

Considering frequencies of 4 for *horse*, 4 for *camel* and 1 for *dromedary*, the normalised vectors are

horse	$\langle 8, 3\rangle$
camel	$\langle 6, 5\rangle$
dromedary	$\langle 7, 5\rangle$

which is illustrated as follows:



Now the picture has changed: *horse* is more similar (because closer in distance) to *dromedary* than to *camel*, and *dromedary* and *camel* are represented by almost the same vector.
The normalisation applies for all measures of distances of points in space.

- **Euclidean Distance:**
  The Euclidean Distance is calculated by summing the squared differences of the vectors' elements and then determining the square root:
  $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.
  The vectors have also to be normalised before applying the measure.

- **Cosine of the Vectors' Angle:**
  This measure does not calculate the distance between points, but the angle α between the n-dimensional vectors which determine the points in n-dimensional space:
  $cos(\alpha) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$.
  The closer the $cos(\alpha)$ is to 1, the smaller the angle α is and therefore the shorter the distance is.

- **Spearman Rank Correlation Coefficient:**
  First, all vectors are normalised by their frequency, and then all elements i of a vector x are replaced by the rank R_{x_i} this element would occupy in a list comprising the elements i of all vectors. The distance is calculated by
  $d = \sum_{i=1}^{n}(R_{x_i} - R_{y_i})^2$.

- **Relative Entropy:**
  The measure of Relative Entropy determines the likelihood for two points being in the same cluster. Given a random variable Y it calculates the uncertainty of vector X, given that vector Y is known:
  $H(X|Y) = -\sum_{x,y} p(x,y) * log[p(x|y)]$.
  The higher the uncertainty H is, the larger the distance is.

- **Hellinger Distance:**
  The Hellinger Distance is defined as:
  $\sum(\sqrt{x_i} - \sqrt{y_i})^2$.

- **Kullback-Leibler Divergence:**
  The Kullback-Leibler Divergence is defined as:
  $\sum x_i * log(\frac{x_i}{y_i})$.

- **Weighted Combination:**
  In addition to those measures mentioned above it is possible to calculate the distance by a combination of two or more methods:
  $d = \alpha_1 d_1 + \alpha_2 d_2 + ... + \alpha_k d_k$.

Considering this variety of metrics we decided to start with the simplest methods before trying to apply more sophisticated suggestions.

3 Algorithm

3.1 Lund et al.'s Version

Having introduced the general idea of high-dimensional space we now come to the concrete algorithm used by Lund et al., complemented by possible changes and variations.

Lund et al. used as data a 160 million word corpus of articles extracted from all newsgroups containing English dialogue (*Usenet*), so their database was conversational and of large diversity. Both the target words and the context words were selected as the 70,000 most frequently occurring symbols within the corpus, therefore the size of the co-occurrence matrix was 70,000 times 70,000.

The co-occurrence counts were calculated as follows: Lund et al. defined a window size of 10 words to the left and to the right of the target words, and within this window, co-occurrence values were inversely proportional to the number of words separating a specific pair. So the word next to the target word got the value 10, the following word the value 9, and so on. In this way, the closeness of the co-occurring word was weighted.

Within the resulting matrix, each row represented the degree to which each context word preceded the target word, and each column represented the co-occurrence values for a context word following each target word. A full co-occurrence vector for a target word consisted of both the row and the column for that word, so the number of dimensions of the vectors was doubled, compared to the size of the matrix. Let's consider an example matrix with three dimensions (instead of 70,000):

	door	*wall*	*house*
door	1	2	11
wall	3	0	7
house	9	6	2

The row *door*, for example, represents how often the context words *door*, *wall* and *house* precede the word *door*, and the column *door* represents how often the context words follow the word *door*. The vectors are therefore determined as:

$$\begin{array}{ll} \textit{door} & \langle 1,2,11,1,3,9\rangle \\ \textit{wall} & \langle 3,0,7,2,0,6\rangle \\ \textit{house} & \langle 9,6,2,11,7,2\rangle \end{array}$$

To reduce the amount of data, the column variances of the particular vectors used in each experiment were computed, and the columns with the smallest variances were discarded. This left a 200-element vector for each target word. The analysis was performed by a multi-dimensional scaling algorithm, that projected points from a high-dimensional space into a lower-dimensional space in a non-linear fashion, preserving the distances between points as much as possible. The distance measure used was Euclidean Distance.

3.2 Our Version

This was the original idea used by the authors. Our model was as close as possible; but we used different data and also tried several variants in which parameters were changed or added, in order to find the optimal combination for the resolution process. A detailed description of our algorithm follows:

The corpus we used for training the model was the *British National Corpus (BNC)* with 100 million words of spoken (approximately 10%) and written (approximately 90%) text. We worked with a part of the BNC containing 30 million words.

Since we were interested in the semantic values and similarities of the bridging descriptions and the antecedents, the target words were determined by this task, as mentioned before. Our target words were those in the WSJ-texts that Poesio, Vieira and Teufel had used in their work.

As explained in section 2.3, the context words should be highly frequent. So, unlike Lund et al.'s algorithm, the context words were not the same words as the target words, but selected from the top of the frequency list of words from the BNC. This should provide a useful basis for co-occurrence counts. Points for consideration were the part of speech (Should we use all parts of speech equivalently?) and number (How many dimensions should we determine for the high-dimensional space?) of the context words.

Another important point concerned the word-forms: Should we adhere to all different word-forms in the texts we use, for example should we distinguish between the verb forms *plan*, *plans* and *planned* or the noun forms *tree* and *trees*? Since we were interested in the lexical semantics of the words, we ignored features like number and gender and used only one form per lexical entry, which was *(to) plan* and *tree* in the described cases. Therefore, all words were lemmatised before they entered the training process. In addition, the second experiment added part of speech tags to the lemmatised word-form in order to distinguish between the same word-form for different parts of speech, for example between *plan* as a verb or a noun.

In determining the window size for co-occurrence we asked Lund et al. about the importance of this feature. They claimed that their algorithm worked equally well for all window sizes, so we decided to vary this parameter between the values 1 and 10. Considering our results, window sizes up to 30 were tried later. As in Lund et al., co-occurrence was measured to the left as well as to the right, and the co-occurrence counts were summed. To get an idea of variation, the algorithm was changed to distinguish between the left and the right side of the target word, which doubled the number of dimensions. We thought that this could improve the algorithm, as co-occurrence in preceding and in following contexts may refine the semantic values. Another parameter varied by us concerned the weighting of closeness of the context word to the target words. For part of the second experiment, this weighting was abandoned to check if the semantic definitions changed.

The previous definitions determined a co-occurrence matrix with the dimensions *number of target words* and *number of context words*, and the dimensions of the co-occurrence vectors were also determined by the number of context words. The procedure Lund et al. applied to double the number of context words for the dimensions of the vectors as the example showed could not be used, since target words and context words were not the same.

Once equipped with the co-occurrence matrix, the distance was measured without applying multi-dimensional scaling (MDS). The measures described in section 2.3 could be used in high-dimensional space, no lower complexity was demanded. An application of MDS might be worth trying, though.

4 Data

Before discussing the experiments, this section describes the data used for training.

4.1 Target Words

As explained in the previous sections, the target words were defined by the bridging descriptions from the 20 parsed articles of the Wall Street Journal we wanted to resolve, plus the possible antecedents.

Within the articles, Renata Vieira had already determined all bridging descriptions and manually worked out the desired antecedent for each of the descriptions, which was only slightly changed, caused by subjective differences in the manual resolution. The bridging descriptions were classified according to the type of relation between the description and the antecedent, as defined in section 2.1. The resulting distribution was as follows:

Relationship	Number	Percentage
Same Head	9	4.4%
Syn/Hyp/Mer	12/13/11	17.7%
Names	44	21.7%
Events	30	14.8%
Compound Nouns	24	11.8%
Discourse Topic	14	6.9%
Inference	46	22.7%
Total	203	100.0%

It might strike the reader that the relationship of *Same Head* does not belong to the bridging descriptions according to our definitions. This was also part of the changes.

Having determined the bridging descriptions, the set of possible antecedents had to be defined. All nouns and verbs (except for auxiliaries) from the preceding five sentences and the same sentence (up to the current position) of the bridging descriptions were considered as possible antecedents. Appendix C gives an impression of which words were considered to be possible antecedents for the bridging description *the markets*.

So far, Vieira had only worked with head nouns (except for the case of resolving compound nouns), but not considering verbs and non-head nouns would have excluded the resolution for events and compound nouns, so we added these. The boundary of five sentences was found to be the best performance compromise between not getting enough possible antecedents – in the sense of not including the desired antecedent – and including too much noise.

4.2 Context Words

The University of Brighton provides an on-line frequency list of the BNC with information about the word, the frequency, the part of speech and the number of files the word occurs in. This list was assumed to give us an estimate of the frequency of words in general, since it is based on a considerably large corpus (of 100 million words). For the task of determining the context words it means that we could easily find out about high-frequent words. But there were two important points to think about:

First, how many context words should we determine? There were many opinions about the concrete number; as there was no time for arbitrarily many experiments to find the best figure, we adopted the number of 2000 dimensions from Huckle [6], determined in his thesis as the optimal value for this parameter in his semantic space model.

Secondly, should we use all kinds of words? Function words, among them prepositions and determiners (e.g. *the* and *of* were at the very top of the frequency list), would certainly co-occur with all possible words, since they appear in all kinds of contexts. But this was not what we wanted. The context words we needed should include a certain amount of semantic content. Therefore only adjectives, common nouns and proper nouns, ordinal numbers and lexical verbs were chosen to represent the context words. The top 2000 words with those parts of speech were taken from the frequency list, all other parts of speech were ignored.

5 Experiments

5.1 Baseline

As a baseline for the following experiments we prepared an algorithm which randomly chose an antecedent out of the nouns and verbs in the preceding five sentences for each of the bridging descriptions. The results are shown in the following table:

Relationship	Resolution	Total
Same Head	1 (11.1%)	9
Synonymy	-	12
Hypernymy	1 (7.7%)	13
Meronymy	-	11
Names	5 (11.4%)	44
Events	-	30
Compound Nouns	2 (8.3%)	24
Discourse Topic	1 (7.1%)	14
Inference	1 (2.2%)	46
Total	11 (5.4%)	203

The percentage of resolution was pretty low; only 5.4% of the bridging descriptions were resolved in the desired way.

5.2 Experiment 1

5.2.1 Method

Our first experiment was realized with the parameters of the data set in the following way:

- **Corpus:**
  The training corpus contained 30 million words from 1200 randomly chosen BNC-texts. All words were lemmatised by John Carroll's lemmatiser based on work at Sheffield.

- **Target Words:**
  Target words were all nouns and verbs from the 20 WSJ-texts, lemmatised in the same way as above.

- **Context Words:**
  Context Words were the 1936 most frequent adjectives, common nouns and proper nouns, ordinal numbers, and lexical verbs from the BNC, lemmatised in the same way as above.

- **Window sizes:**
  The window sizes were 1, 2, 3, 5 and 10 words to the left and to the right.

- **Measures:**
  The bridging descriptions were resolved by finding the closest antecedent in the space according to the measures *Manhattan Metric*, *Euclidean Distance* and *Cosine*.

5.2.2 Results and Tendencies

The following table shows the resolution of the bridging descriptions; the percentage is based on the total number of 203. The best result is printed in bold:

	Window Size				
Metric	1	2	3	5	10
Man	37 (18.2%)	36 (17.7%)	39 (19.2%)	41 (20.2%)	37 (18.2%)
Euc	37 (18.2%)	36 (17.7%)	39 (19.2%)	39 (19.2%)	40 (19.7%)
Cos	39 (19.2%)	36 (17.7%)	39 (19.2%)	42 (20.7%)	**45 (22.2%)**

The best resolution for *Manhattan Metric* was achieved at window sizes of three and five, for *Euclidean Distance* the results seemed to get (slightly) better with larger windows, and for *Cosine*, the results were also the better the larger the window was, and the results for a window size of one were only slightly worse. *Cosine* appeared as the most successful measure, with 45 correct resolutions out of 203 bridging descriptions.

5.3 Experiment 2

5.3.1 Method

In the first experiment, the lemmatiser had worked on non-tagged words and therefore introduced ambiguities and mistakes. These cases were now avoided by first tagging the data with the respective parts of speech, and then lemmatising on that basis.

In addition, the differences between American (in the Wall Street Journal) and British English (in the British National Corpus) had not been considered. This was improved by transferring all words of the Wall Street Journal into British English.

With these refinements, our second experiment was realized with the parameters of the data set in the following way:

- **Corpus:**
  The training corpus contained 30 million words, this time plus their parts of speech, from 1200 randomly chosen BNC-texts. All word-tag pairs were lemmatised by CELEX, a morphological database.

- **Target Words:**
  Target words were all nouns and verbs from the 20 WSJ-texts plus their part of speech tags. Words in American English were transformed into British English, the tags converted into the BNC-taxonomy. Then the word-tag pairs were lemmatised in the same way as above.

- **Context Words:**
  Context Words were the 2061 most frequent adjectives, common nouns and proper nouns, ordinal numbers, and lexical verbs from the BNC, lemmatised in the same way as above.

- **Window sizes:**
  Since there was a tendency of some metrics in the first experiment to improve the resolution with an increasing window, it was possible that the result could improve again, so a bigger window was added. The window sizes were 1, 2, 3, 5, 10 and 15, and later on 20 and 30, words to the left and to the right.

- **Measures:**
The bridging descriptions were resolved by finding the closest antecedent in the space according to the measures *Manhattan Metric*, *Euclidean Distance* and *Cosine*.

To check whether the parameters of (a) weighting the closeness between two co-occurring words and (b) not distinguishing between the left and the right of target words – as explained in section 3 –, this experiment consisted of three sub-experiments:

- **Part 1:** Standard algorithm
The algorithm was performed in the same way as in the first experiment, i.e. in the way it was described in section 3.

- **Part 2:** Closeness not weighted
This time the algorithm was changed: The closeness of the context word to the target word was no longer weighted, so each occurrence of a context word in the window counted equally.

- **Part 3:** Distinction between left and right
This time the algorithm did distinguish between words on the left and on the right in the window, so there were two counts for each target word - context word - pair, and therefore the number of columns in the matrix was doubled.

5.3.2 Results and Tendencies

- **Part 1:** Standard algorithm

  Since preliminary results for *Manhattan Metric* showed a trend towards more successful resolution with increasing window size and might improve even more with a bigger window, the experiment was extended to a window size of 20. The best score is printed in bold:

	Window size				
Metric	1	2	3	5	10
Man	34 (16.8%)	35 (17.2%)	41 (20.2%)	41 (20.2%)	42 (20.7%)
Euc	35 (17.2%)	37 (18.2%)	37 (18.2%)	36 (17.7%)	37 (18.2%)
Cos	41 (20.2%)	45 (22.1%)	**46 (22.7%)**	41 (20.2%)	41 (20.2%)

	Window size	
Metric	15	20
Man	44 (21.7%)	44 (21.7%)
Euc	38 (18.7%)	39 (19.2%)
Cos	38 (18.7%)	38 (18.7%)

The tables show that *Euclidean Distance* resolved the bridging descriptions worst. There was a trend towards better resolution with increasing window size, but the number of resolved bridging descriptions did not even reach 40.

Manhattan Metric had a strong trend towards resolving better with an increasing window size and got the third highest successful number of 44.

Cosine was the most successful measure when a window size of 3 was used. Unlike the first experiment, the success did not increase with an increasing window size, but improved until a window of 3 and got worse afterwards.
To have an idea about what the results look like, the resolution can be found as appendix D.

The window sizes of 15 and 20 emphasised this impression. So there was no common tendency for all methods considering the different window sizes and the achieved success. Actually, each method had to be interpreted by itself. Again, *Cosine* appeared as the most successful measure, but this time not with the largest window, but a window of 3 words.

- **Part 2:** Closeness not weighted

	Window size	
Metric	5	10
Man	41 (20.2%)	44 (21.7%)
Euc	38 (18.7%)	39 (19.2%)
Cos	39 (19.2%)	39 (19.2%)

Again, there were no common tendencies: The results for *Cosine* got worse for both window sizes, for *Manhattan Metric* they stayed the same for a window size of 5, but got better for a window size of 10. For *Euclidean Distance*, the successful number of resolution improved for both window sizes, but was still below 40.

The experiment was extended to window sizes of 15 and 20 to see whether *Manhattan Metric* or *Euclidean Distance* would get better results:

	Window size	
Metric	15	20
Man	42 (20.7%)	45 (22.1%)
Euc	39 (19.2%)	38 (18.7%)

The results for *Manhattan Metric* improved for a window size of 20, but all other combinations did not improve. Since *Manhattan Metric* still seemed to improve with an increasing window, another larger window size of 30 was tested as well. The result was again a successful score of 45 (22.1%), so it did not improve once more.

- **Part 3:** Distinction between left and right

	Window size	
Metric	5	10
Man	39 (19.2%)	41 (20.2%)
Euc	37 (18.2%)	39 (19.2%)
Cos	44 (21.7%)	41 (20.2%)

The results showed, as before, different behaviour for the different measures: Compared to the standard method, the scores for *Manhattan Metric* got worse for all window sizes, but improved for *Cosine* for a window size of 5. For *Euclidean Distance*, again all results were improved. Indicating possible general improvement, the experiment was extended to examine the window size of 3 for *Cosine*, which was the most successful size for the standard algorithm, and the window size of 15 for *Euclidean Distance*, the third best result, to see whether the trend towards better resolution continued:

	Window size	
Metric	3	15
Euc		38 (18.7%)
Cos	43 (21.1%)	

Both results turned out not to be similarly successful, the resolution was worse.

It can be repeated that there was no common tendency for all methods considering their success in resolution for different window sizes. Summarising the results and tendencies, the resolution for *Manhattan Metric* improved for the standard algorithm with an increasing window size. The best results were for window sizes of 15 and 20. Without weighting the closeness between the target word and the context word there was no tendency observable. The results were as good as in the standard version for the window sizes 10 and 20. Distinguishing between co-occurrence on the left and co-occurrence on the right made the results slightly worse.

The resolution with *Euclidean Distance* was the worst of all three measures. Though the results showed tendency to improve with increasing window size in all three parts, the successful count got worse at a certain point before it ever reached 40.

The resolution for measuring *Cosine* improved with the standard method until a window size of 3, then it got worse. Without weighting closeness the successful counts were worse, distinguishing between left and right improved for window sizes of 5 and 10, but not for the best count with a window size of 3.

The best overall result was 46 correct resolutions out of 203 possible ones, 22.7%. So the results of this second experiment hardly improved those from the first one. Taking different parts of speech into account did not support the resolution process in a positive way.

6 Interpretation

For the goal of interpreting the results we examined the most successful resolutions in both experiments more closely. The following table represents the respective numbers for the different relationships:

Relationship	Resolution				Total
	Exp. 1		Exp. 2		
Same Head	9	(100.0%)	9	(100.0%)	9
Synonymy	3	(25.0%)	4	(33.3%)	12
Hypernymy	2	(15.4%)	2	(15.4%)	13
Meronymy	4	(36.4%)	2	(18.2%)	11
Names	1	(2.3%)	1	(2.3%)	44
Events	3	(10.0%)	5	(16.7%)	30
Compound Nouns	16	(66.7%)	16	(66.7%)	24
Discourse Topic	2	(14.3%)	1	(7.1%)	14
Inference	5	(10.9%)	6	(13.0%)	46
Total	45	(22.2%)	46	(22.7%)	203

Compared to the baseline experiment with 5.4% and also the WordNet resolution of 16.6% we achieved a good number of right resolutions, but still there were only 22.7% resolved in the best case. This is especially a problem for the precision: Since our algorithm always suggested an antecedent for a bridging description (i.e. the recall was 100%) the precision was identical to the result of the resolution.

Before we start with a more elaborated interpretation we consider some more general influences on the specific classes:

- The precision in the class *Same Head* had to be 100%, since the bridging descriptions were resolved to the same word as an antecedent, and the same word was always the most similar one, since the vectors were identical. Only if the desired antecedent was not among the considered ones, the resolution could possibly have been wrong.

- The three relationships *Synonymy*, *Hypernymy/Hyponymy* and *Meronymy* might be considered as typical WordNet relationships, since these are the relationships explicitly coded in the semantic hierarchy. The resolution with our algorithm varied between 15% and 36%, which is around the average of our resolution. Comparing with WordNet, where the average number of resolution is lower than with our algorithm, the resolution for these three classes is better: We resolved an average of 25.0%/22.2% percent, whereas WordNet was able to resolve 39.5%. One reason for that is certainly due to the fact that the structure in WordNet relations is based on the definition of lexical relationships, and synonymy, hypernymy/hyponymy and meronymy are among those relationships. Therefore, these relationships could directly be found as links between words, and given two words – in our task the bridging description and a possible antecedent – conceivable relationships between those words could easily be looked up in WordNet. Mistakes in the resolution considering these relationships were according to Poesio, Vieira and Teufel mostly due to (a) missing links between words, and (b) the unexpected way in which knowledge is organised in WordNet. In our approach, the clustering algorithm did not create any specific relationship between the words in one cluster.

- The *Names* were resolved with a bad result. But these bridging descriptions were also the most difficult ones to resolve, since they were so specific in the Wall Street Journal – like names of persons who had been interviewed, conductors of concerts – that they could hardly (with some exceptions: not at all) be found in the part of the BNC which was examined. An exception to this were well known names like *Bach*.

- The resolution for *Events* was improved in the second experiment, considering the different parts of speech. But still, there were only 16.7% right. The algorithm should have been able to find more associations, since the verbs which appeared in the texts were neither low-frequent words nor semantically far away.

- The recall of *Compound Nouns* should be total, since the bridging descriptions resolved to the same word. The reason why only two thirds of them were resolved correctly is, that the desired antecedents of the missing third were not among the considered ones.

- The class containing the relationship *Discourse Topic* is difficult to resolve, since (i) some topics were not explicitly mentioned in the text, but only implicit in the background, and (ii) those topics explicitly mentioned were often very general – e.g. *oil* – and difficult to determine as antecedent for a description like *market*, for which there are words more closely related.

- The relationship *Inference* is a mixture of sometimes very closely related and therefore easy to resolve word pairs, like *buyer* and the antecedent *sale*, and sometimes word pairs which are only related to each other in specific contexts, like *carrier* and the antecedent *pollination*. For that reason, the result for this class was below the average.

We once more had a look at the restrictions we had posed on the algorithm and their consequences in the result. First, we examined the parameters:

- *Corpora:*
  The different corpora for the training and the resolution process, the British National Corpus for one, the Wall Street Journal for the other task, certainly influenced the result. As mentioned before, several expressions in the WSJ, as proper names for example, do not appear in the BNC, so that their semantic values were based on sparse data. Using the same corpus for both tasks should improve the resolution.
  Another parameter concerning corpora is the size of the training corpus: Again, to overcome the bottleneck of sparse data, the training corpus should be as large as possible. Therefore, increasing the size from 30 million words up to possibly the whole BNC (100 million words) should refine the co-occurrence counts and therefore the resolution.

- *Target Words:*
  The choice of the target words was also an important point of consideration. Those words which definitely had to be considered were the bridging descriptions themselves, since their semantic values had to be determined. But the parameter determining which kinds of words were considered as possible antecedents was variable. So far, all nouns and verbs from the preceding five sentences had been chosen. A possibility to reduce the number might be to include syntactic information, for

example based on a partial parser. First, typical (syntactic) relationships between bridging descriptions and antecedents would have to be determined; applying these relationships to the bridging descriptions in the parse should make the set of possible antecedents smaller.

- *Context Words:*
  In section 4.2 we have already argued that we adopted the number of context words from Huckle's PhD-thesis [6], since there was no time to vary this parameter. But since the number was very much task-dependent, it might be worth trying other dimensions, especially considering that Lund et al. started with 70,000, compared to 2,000 we used in our version.
  In addition, the context words depended on the subjective judgement on which parts of speech contained most content information. An alternative would be to choose the context words by a method considering the reliability of words as indicators of context, as developed in McDonald's PhD-work [9].

- *Word-Forms:*
  The issue of which word-forms were useful for target words, context words and the corpus was varied in the two experiments. Surprisingly, the change from the purely lemmatised words to lemmatised and tagged words did not change (or rather improve) the result. How can we interpret this? Probably, the semantic content in the lemmatised form did not need any extra distinction between the different word-forms to achieve the necessary degree of semantic similarity between bridging description and antecedent.

- *Window Size:*
  As mentioned before, the window size is an important parameter. We only tried sizes between 1 and 30 (concentrating on 1 to 10) in this thesis, not discovering any strong tendencies towards improvement with a certain size. So this parameter is still open to further variation.

- *Metric:*
  The best performing metric for measuring distances in semantic space seems to be the *Cosine*. In both experiment it achieved the best results. Of course, not all possible measures were tried, so there is still room for extensions. Another possibility is to apply combinations of measures instead of one measure on its own, as mentioned before.

In addition, we found three main classes of mistakes we had caused in preparing the data and the algorithm:

1. In some cases, the desired antecedent could not be found since it was not in the part of the text we had considered for resolution, i.e. it was not on the list of possible antecedents for the bridging description we had created. This happened if the right word in the text was outside the area we considered for possible antecedents; they were either before the preceding five sentences (20 cases – 9.9%) or after the description (2 cases – 1.0%). Another antecedent had to be suggested which had to be the wrong result.

2. In other cases, the lemmatisation of either the bridging description or the desired antecedent was wrong, so that it was not possible to resolve the description in the way we had determined. For example, the noun *evening* was lemmatised to *even*.

3. Some wrong resolutions were caused by mistakes in the automatic extraction of words: In the first experiment, for instance, we did not consider adjectives as possible bridging descriptions, but in fact the words *two* and *half* appeared as heads of bridging descriptions.
   In addition, in some cases our algorithm extracted the wrong word as head of the bridging description, for example *office* instead of *company* in the noun phrase *the combined companies' offices*.

Resolutions which fell under these cases were excluded from the possibility of being resolved in the desired way right from the beginning.
The following table shows the distribution of the described phenomena over the different relationships:

Relationship	Case 1	Case 2		Case 3		Total
	Both Exp.	Exp. 1	Exp. 2	Exp. 1	Exp. 2	
Same Head	-	-	-	-	-	9
Synonymy	4 (33.3%)	-	-	-	-	12
Hypernymy	2 (15.4%)	-	-	-	-	13
Meronymy	-	-	-	-	-	11
Names	4 (9.1%)	-	-	3 (6.8%)	-	44
Events	1 (3.3%)	-	-	1 (3.3%)	-	30
Compound Nouns	3 (12.5%)	1 (4.2%)	1 (4.2%)	-	-	24
Discourse Topic	3 (21.4%)	1 (7.1%)	-	-	-	14
Inference	5 (10.9%)	1 (2.2%)	-	1 (2.2%)	-	46
Total	22 (10.8%)	3 (1.5%)	1 (0.5%)	5 (2.5%)	-	203

There is no strong tendency which could evoke a certain relation between these cases of wrong resolution and a certain class of bridging descriptions.

But the total numbers are impressive: While cases 2 and 3 only present about 2% of all bridging descriptions, case 1 describes the fact that 10% of them were completely lost because of the strong restriction of considering only the five preceding sentences as range for antecedents. This raises the question whether there would have been a better division. But as mentioned before, considering more sentences introduces more noise than it supports the algorithm. We should look for alternative divisions, like paragraphs, for example: For further hints, one of the (longer) texts of the Wall Street Journal, containing 26 bridging descriptions, was examined:

- Considering sentences:
  10 of the bridging descriptions were in the same sentence, 7 in the sentence before, 5 in another sentence before and further 4 up to the threshold of five sentences.

- Considering paragraphs:
  19 of the bridging descriptions were in the same paragraph, and the other 7 in the preceding paragraph.

This data evoked the idea of basing the search for antecedents on paragraphs instead of sentences, since the distribution seems to be more uniform.

The preceding discussion shows that at least 10% of the bridging descriptions were resolved in the wrong way because of the restrictions we had posed on the resolution algorithm. But what about the remaining cases which were resolved wrongly? Is it possible to identify a semantic relation between bridging description and antecedent, or were the descriptions resolved in a somehow arbitrary way?

First, there is a certain number of bridging descriptions which were resolved to the same word-form, though the resolution should have been to a related, but not identical, word. Of course, as soon as there was a word-form among the antecedents which was identical to the bridging description, this word-form succeeded in the process of resolution, since the vector was identical as well. There were two main reasons for that: (a) the desired antecedent was more specific than the chosen one, e.g. the text was about companies and mentioned the word *company* quite often, and then it mentioned the specific company called *Pinkerton*, so the following bridging description *the company* should refer to that name, but resolved to the word *company*, which had appeared in the preceding five sentences, (b) in the first experiment, lemmatising had sometimes created two identical word-forms out of two different lexemes, usually noun and verb, e.g. *to plan* and *the plan*, and since we did not distinguish between different parts of speech, there was no difference in the word-form.

As the following table shows, there were again about 10% of such cases:

Relationship	Identical Word-Form		Total
	Exp. 1	Exp. 2	
Same Head	-	-	9
Synonymy	2 (16.7%)	1 (8.3%)	12
Hypernymy	2 (15.4%)	3 (23.1%)	13
Meronymy	-	-	11
Names	9 (20.5%)	10 (22.7%)	44
Events	1 (3.3%)	1 (3.3%)	30
Compound Nouns	-	-	24
Discourse Topic	2 (14.3%)	2 (14.3%)	14
Inference	6 (13.0%)	4 (8.7%)	46
Total	22 (10.8%)	21 (10.3%)	203

All these cases of bridging description bear a semantic relationship with the chosen antecedent, since it is the identical word. But this was not the semantic association we had asked for. The high percentages for some of the relationships and the total of more than 10% of all bridging descriptions clearly show that this kind of wrong resolution should not be ignored. We will come back to this issue later on.

Secondly, we had a look at the remaining cases which have not be explained yet. Having resolved 45/46 bridging descriptions in the right way and explained that 30/23 descriptions were wrongly resolved due to mistakes in the algorithm and 22/21 due to an identical word-form, there are 106/113 cases still to consider, in the first and second experiment, respectively. It appeared that several of these bridging descriptions were semantically very close to the antecedent found by the algorithm, sometimes even closer than the desired antecedent, for example *market* resolved to *customer* instead of *phone service*. Other bridging descriptions were resolved to a completely unrelated word. The following table shows the numbers of how semantically associated antecedent and bridging description were (subjectively judged, of course) in the two experiments, compared to the 106/113 cases we still want to explain.

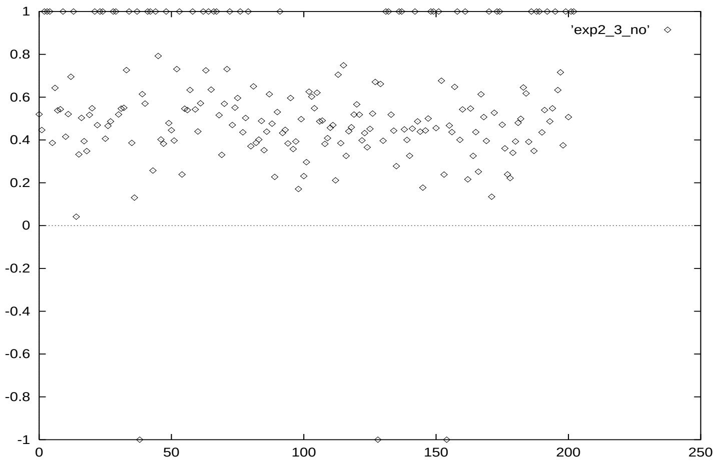
Semantic Association	Exp. 1	Exp. 2
Total Number	106	113
Semantically Associated	67 (63.2%)	69 (61.1%)
Not Semantically Associated	39 (36.8%)	44 (38.9%)

The most striking point in this table is that almost two thirds of the cases we could have resolved by our algorithm were resolved in a semantically associated way, but still not right. So the lack in the resolution process was not caused by creating an insufficient semantic space, as one could have assumed. The table above clearly shows that we can generally observe semantic association between bridging description and antecedent, so we successfully resolved most bridging descriptions to semantically close words. We can therefore follow that semantic similarity is not the only kind of information we need for the resolution process.

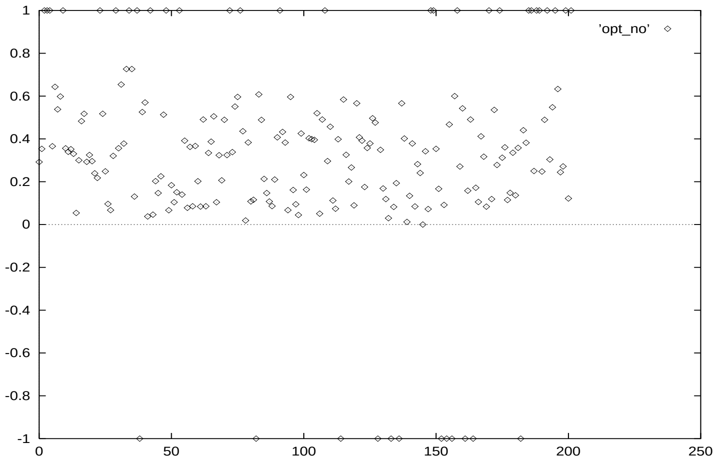
We considered another possibility of what might had been arranged in an insufficient way: What about assessing the time period between mentioning the antecedent and the bridging description? Was the desired antecedent generally mentioned more recently than the wrongly chosen one, so that we should have relied on this kind of closeness? We had a look on the best resolution of the second experiment, and there 73 (61.9%) of the desired antecedents would have been mentioned more recently to the description than the actually chosen antecedent, 45 (38.1%) would not. So there is no strong tendency towards a better resolution with taking recency of the antecedent into account; this was no lack in the algorithm.

We are left with the surprising observation that the main reason for the insufficient resolution was not mainly caused by the implementation of the algorithm, but by the connection between the idea of the model and our motivation of resolving bridging descriptions. It seems as if the model created by Lund et al. which mirrors the degree of semantic priming between words is not the model we need for the resolution process. If it was the right model there should not have been so many word pairs satisfying Lund et al.'s semantic hypothesis but still not sufficient for the resolution process.

This hypothesis is emphasised by a further illustration. Consider the following pictures, based on data of the best resolution in the second experiment. The first figure shows the cosine of the distances of the 203 antecedents to the bridging descriptions, as chosen by our algorithm. The higher the cosine is (i.e. the closer to +1), the shorter the distance between antecedent and description was. The numbers vary from -1 to +1, but are concentrated in the area between 0.3 and 0.6:



The next figure shows the cosine of the distances of the 203 desired antecedents to the bridging descriptions. Surprisingly, also these distances show variation, but in the area between 0 and 0.6:



This indicates that also the antecedents which should have been chosen show a certain distance to the bridging descriptions, in average larger than in our resolution, which is expressed in the wider distribution in the second figure. This fact undermines our hypothesis that not the word which is semantically most strongly priming a bridging description is the one needed for its resolution. Even if we had resolved all bridging descriptions to an antecedent in a very short distance, we would not have succeeded, since – as the second figure clearly shows – the resolution should be to an antecedent in a certain distance.

This raises the final question if it is possible at all to utilise Lund et al.'s model for our purposes, or if the underlying hypotheses of the model are too strong. How could we possibly determine the optimal distance? One potential solution comprises two steps:

1. The cases on the +1 line are those which were resolved to the same word-form, i.e. the bridging descriptions with the relationship *Same Head* or *Compound Noun*, and in addition (in the first figure) those cases where the resolution was wrongly to an antecedent with the same word-form – as mentioned before. If we left those cases outside our considerations we could concentrate on the relationships which are resolved in a certain distance, as illustrated by the second figure.
   A reasonable argument for this is the fact that we could concentrate on the bridging descriptions which require an active inference between antecedent and bridging description, excluding identity.

2. Then, instead of looking for the closest word in space, we could determine the antecedent by setting a certain range. For example, the distance between bridging description and antecedent had to be between 0 and 0.6, and the one closest to 0.3 is chosen. This would at the same time improve our precision, because not all bridging descriptions would be resolved.

With this procedure the success of the resolution would necessarily improve. But still, a variety of guesses are included, because we cannot see a certain relationship between the distances of the bridging description and the antecedent. So we leave this open as a further possibility and come back to the point of asking why utilising Lund et al.'s model was not optimal for our task.

Lund et al. claim that their system is able to distinguish between semantic and associative relations. They present an example by determining a difference between the two word-pairs *bed* – *table* with a purely semantic relation and *cradle* – *baby* with a purely associative relation, for example. They argue that semantic similarity is required in order to show a priming effect in the simulation. Therefore, semantic priming comes only with the former, but not with the latter word-pair. And exactly those (purely) semantic relationships between words are determined by a similar context, so that the words cluster together in high-dimensional space. (Purely) Associative word-pairs, on the other hand, tend to appear in the same sentence, but are not interchangeable in context. So the clusters we got by creating co-occurrence matrices were semantically determined, purely associated relations were not considered.

Compare this with the demands for resolving bridging descriptions. We were looking for **both** kinds of relationships, for semantically similar word-pairs which appear in similar contexts, as *home – the house*, as well as for associated word-pairs, often appearing in the same sentence, as *tree – the leaves*. Considering that Lund et al.'s simulation only mirrors semantic relations, as they explain, we missed the associative relations for our resolution process.

The different classes of relationships should be examined once more: Antecedents which bear the relationships of *Same Head*, *Synonymy* or *Names* to their bridging descriptions certainly appear in the same contexts; thus the semantic relation as described above is high. But the relationships of *Hypernymy/Hyponymy* and *Meronymy* fall into the class of associative relation, i.e. the related words tend to appear in the same sentence, but not necessarily in the same context. For *Events*, already the syntactic pattern excludes that noun and verb might be in the same context, at least not within a small window. *Compound Nouns* strongly depend on the head of the compound, since that does not necessarily have to be associated with the bridging description and therefore appear in a similar context. Concerning *Discourse Topics* and *Inferences*, the relation cannot be determined concretely, since the possibilities are more variable.

What chance is there to improve the usefulness of Lund et al.'s model for our task? One point worth to consider in the cases of associative relation might be to work with different, larger window sizes to determine the co-occurrence matrix, since a large enough window also grasps words-pairs which tend to appear in the same sentence instead of only in the same context. Another important issue is keeping track of the focus. Since our interpretation strongly shows that the semantic priming effect is not sufficient on its own to resolve bridging descriptions, some more information is missing. And since the focus is always the information on top of the common ground, this might be the relevant point. The focus should, for example, support the resolution in cases I described before, when the focus changes to a more specific subject (remember the example of focus change from the general word *company* to a specific company named *Pinkerton*).

Concluding the interpretation, we can say that the psychological process of resolving bridging descriptions is not sufficiently modelled by Lund et al.'s semantic priming space. Some relationships are possible to be resolved; for others, more elaborated algorithms are needed.

A Text from Wall Street Journal

Investors are appealing to the Securities and Exchange Commission not to limit their access to information about stock purchases and sales by corporate insiders.

A SEC proposal to ease reporting requirements for some company executives would undermine the usefulness of information on insider trades as a stock-picking tool, individual investors and professional money managers contend.

They make *the argument* in letters to *the agency* about rule changes proposed this past summer that, among other things, would exempt many middle-management executives from reporting trades in their own companies' shares. The proposed changes also would allow executives to report exercises of options later and less often.

Many of the letters maintain that investor confidence has been so shaken by the 1987 stock market crash – and *the markets* already so stacked against the little guy – that any decrease in information on insider-trading patterns might prompt individuals to get out of stocks altogether.

"*The SEC* has historically paid obeisance to the ideal of a level playing field," wrote Clyde S. McGregor of Winnetka, Ill., in one of the 92 letters the agency has received since the changes were proposed Aug. 17. "Apparently the commission did not really believe in this ideal."

Currently, *the rules* force executives, directors and other corporate insiders to report purchases and sales of their companies' shares within about a month after *the transaction*. But about 25% of the insiders, according to SEC figures, file their reports late.

The changes were proposed in an effort to streamline federal bureaucracy and boost compliance by the executives "who are really calling the shots," said Brian Lane, special counsel at the SEC's office of disclosure policy, which proposed the changes.

Investors, money managers and corporate officials had until today to comment on *the proposals*, and *the issue* has produced more mail than almost any other issue in memory, Mr. Lane said. The SEC will probably vote on the proposal early next year, he said.

Not all those who wrote oppose the changes. The Committee on Federal Regulation of Securities for the American Bar Association argues, for example, in its lengthy letter to the SEC, that the proposed changes "would substantially improve *the law* by conforming it more closely to contemporary business realities."

What the investors who oppose the proposed changes object to most is the effect they say the proposal would have on their ability to spot telltale "clusters" of trading activity – buying or selling by more than one officer or director with in a short period of time. According to some estimates, the rule changes would cut insider filings by more than a third.

The SEC's Mr. Lane vehemently disputed those estimates. *The rules* will eliminate filings by "vice presidents of maintenance and personnel," but will still require reports from vice presidents of sensitive or policy-making divisions, such as sales, marketing, finance and research and development, Mr. Lane said.

The proposed rules also would be tougher on the insiders still required to file reports, he said. Companies would be compelled to publish in annual proxy statements the names of insiders who fail to file reports on time.

Considered as a whole, Mr. Lane said, the filings required under the proposed rules "will be at least as effective, if not more so, for investors following transactions."

But Robert Gabele, president of Invest/Net, a North Miami, Fla., company that packages and sells the insider-trading data, said the proposal is worded so vaguely that key officials may fail to file the reports.

Many investors wrote asking the SEC to require insiders to report their purchases and sales immediately, not a month later. But Mr. Lane said that while the SEC regulates who files, the law tells them when to do so. Investors who want to change *the required timing* should write their representatives in Congress, he added. The SEC would likely be amenable to legislation that required insiders to file transactions on a more timely basis, he said.

B Encoding in WordNet

B.1 Hypernyms of the verb *lose*

```
Sense 1
lose, fail to keep, fail to maintain

Sense 2
lose, fail to win

Sense 3
lose
       => suffer, suffer emotionally, endure distress
           => feel, experience

Sense 4
misplace, mislay, lose
       => put, set, place, pose, position, lay
           => move, displace, make move

Sense 5
lose, miss from one’s possessions, lose sight of

Sense 6
lose, lose money, make a loss, fail to profit

Sense 7
lose, fail to get

Sense 8
lose

Sense 9
fall back, lose, drop off, fall behind, recede
       => regress, retrograde, undergo regress, retrogress
           => worsen, decline, grow worse, get worse
               => change state, turn
                   => change
```

```
Sense 10
sweat off, lose
       => reduce, melt off, lose weight, slim, slenderize, thin, slim down
           => change state, turn
               => change
```

B.2 Synonyms of the noun *tree*

```
Sense 1
tree
       => woody plant, ligneous plant

Sense 2
tree, tree diagram
       => plane figure, two-dimensional figure
```

C Possible antecedents of *the markets*

All possible antecedents are printed in italic fonts, the desired antecedents and the bridging description itself in bold:

Investors are *appealing* to the *Securities* and *Exchange Commission* not to *limit* their *access* to *information* about *stock purchases* and *sales* by corporate *insiders*.

A *SEC proposal* to *ease reporting requirements* for some *company executives* would *undermine* the *usefulness* of *information* on *insider trades* as a stock-picking *tool*, individual *investors* and professional *money managers contend*.

They *make* the *argument* in *letters* to the *agency* about *rule changes proposed* this past *summer* that, among other *things*, would *exempt* many middle-management *executives* from *reporting trades* in their own *companies*' *shares*. The *proposed changes* also would *allow executives* to *report exercises* of *options* later and less often.

Many of the *letters maintain* that *investor confidence* has been so *shaken* by the 1987 *stock* **market** *crash* – and the **markets** already so stacked against the little guy – that any decrease in information on insider-trading patterns might prompt individuals to get out of stocks altogether.

D Resolution of bridging descriptions in the second experiment with a window size of 3 and *Cosine* as measure

The pattern for the resolution is as follows:

```
bridging description / tag number ---> antecedent / tag number
```

where 1 is the tag number for a (lemmatised) noun, and 4 the tag number for a (lemmatised) verb.

```
argument/1 ---> proposal/1
agency/1 ---> company/1
market/1 ---> market/1
sec/1 ---> sec/1
rule/1 ---> rule/1
transaction/1 ---> company/1
proposal/1 ---> propose/4
issue/1 ---> change/1
law/1 ---> regulation/1
rule/1 ---> rule/1
timing/1 ---> consider/4
problem/1 ---> work/4
work/1 ---> work/4
audience/1 ---> audience/1
clarinetist/1 ---> museum/1
row/1 ---> set/1
record/1 ---> case/1
half/1 ---> row/1
composer/1 ---> sonata/1
image/1 ---> choose/4
two/3 ---> choose/4
composer/1 ---> composer/1
crowd/1 ---> keep/4
audience/1 ---> audience/1
half/1 ---> half/1
evening/1 ---> way/1
technique/1 ---> use/4
structure/1 ---> work/1
piece/1 ---> piece/1
music/1 ---> music/1
```

```
generation/1 ---> create/4
development/1 ---> system/1
part/1 ---> effect/1
female/1 ---> male/1
plant/1 ---> plant/1
female/1 ---> company/1
pollen/1 ---> plant/1
plant/1 ---> plant/1
rape-seeds/1 ---> say/4
company/1 ---> add/4
approach/1 ---> technique/1
company/1 ---> company/1
organ/1 ---> organ/1
carrier/1 ---> call/4
currency/1 ---> currency/1
others/1 ---> expect/4
opening/1 ---> others/1
economist/1 ---> others/1
feed/1 ---> feed/1
drop/1 ---> fall/1
unit/1 ---> remain/4
currency/1 ---> market/1
company/1 ---> business/1
watch/1 ---> watch/1
veteran/1 ---> bring/4
owner/1 ---> buy/4
unit/1 ---> company/1
two/3 ---> work/4
firm/1 ---> firm/1
acquisition/1 ---> acquire/4
agency/1 ---> business/1
employee/1 ---> business/1
company/1 ---> company/1
firm/1 ---> company/1
company/1 ---> company/1
sale/1 ---> company/1
contract/1 ---> contract/1
lawsuit/1 ---> lawsuit/1
two/3 ---> say/4
equilibrium/1 ---> demand/1
change/1 ---> activity/1
```

```
business/1 ---> company/1
price/1 ---> price/1
market/1 ---> price/1
time/1 ---> year/1
issue/1 ---> problem/1
share/1 ---> share/1
prospect/1 ---> look/4
oil/1 ---> exploration/1
company/1 ---> company/1
volatility/1 ---> price/1
activity/1 ---> work/1
slump/1 ---> price/1
industry/1 ---> official/1
staff/1 ---> company/1
crash/1 ---> leave/4
well/1 ---> others/1
area/1 ---> site/1
bust/1 ---> get/4
concern/1 ---> say/4
company/1 ---> say/4
segment/1 ---> segment/1
price/1 ---> buy/4
housewife/1 ---> get/4
government/1 ---> home/1
problem/1 ---> issue/1
blame/1 ---> think/4
population/1 ---> land/1
nation/1 ---> land/1
legislation/1 ---> administration/1
ceiling/1 ---> bill/1
penalty/1 ---> bill/1
measure/1 ---> effect/1
change/1 ---> effect/1
critic/1 ---> others/1
argument/1 ---> others/1
proportion/1 ---> amount/1
policy/1 ---> measure/1
city/1 ---> home/1
campaign/1 ---> politician/1
president/1 ---> david/1
mayoralty/1 ---> borough/1
```

```
candidacy/1 ---> president/1
payment/1 ---> effort/1
debt/1 ---> company/1
guy/1 ---> say/4
post/1 ---> president/1
candidate/1 ---> position/1
remark/1 ---> mind/1
problem/1 ---> come/4
house/1 ---> home/1
chimney/1 ---> foot/1
lawn/1 ---> porch/1
rubble/1 ---> wall/1
fund/1 ---> help/4
floor/1 ---> room/1
wall/1 ---> room/1
kitchen/1 ---> room/1
hammacks/1 ---> coverage/1
people/1 ---> say/4
field/1 ---> bring/4
insurer/1 ---> insurer/1
company/1 ---> company/1
foot/1 ---> chimney/1
yard/1 ---> come/4
roll/1 ---> foot/1
foot/1 ---> foot/1
house/1 ---> house/1
neighbourhood/1 ---> home/1
division/1 ---> position/1
debris/1 ---> ground/1
floor/1 ---> ground/1
cost/1 ---> cost/1
calculation/1 ---> check/1
floor/1 ---> hit/4
roadway/1 ---> wish/4
market/1 ---> continue/4
move/1 ---> continue/4
fcc/1 ---> fcc/1
discount/1 ---> discount/1
agency/1 ---> plan/1
action/1 ---> action/1
firm/1 ---> business/1
```

```
mold/1 ---> london/1
57-year-old/1 ---> nothing/1
purchase/1 ---> share/1
one/1 ---> people/1
client/1 ---> individual/1
loan/1 ---> loan/1
stake/1 ---> company/1
acquisition/1 ---> acquire/4
company/1 ---> company/1
tycoon/1 ---> company/1
problem/1 ---> business/1
demise/1 ---> people/1
one/1 ---> people/1
financier/1 ---> company/1
neighbour/1 ---> say/4
measure/1 ---> grant/4
figure/1 ---> pattern/1
increase/1 ---> increase/1
government/1 ---> turtle/1
process/1 ---> program/1
program/1 ---> program/1
country/1 ---> country/1
issue/1 ---> country/1
production/1 ---> program/1
war/1 ---> show/1
proceeds/1 ---> market/1
issue/1 ---> president/1
requirement/1 ---> procedure/1
measure/1 ---> continue/4
side/1 ---> end/1
plan/1 ---> strategy/1
proposal/1 ---> plan/1
tax/1 ---> revenue/1
rate/1 ---> rate/1
governor/1 ---> add/4
policy/1 ---> policy/1
charge/1 ---> charge/1
expense/1 ---> continue/4
site/1 ---> plan/4
acre/1 ---> acre/1
loan/1 ---> mortgage/1
```

```
buyer/1 ---> sale/1
report/1 ---> report/1
plan/1 ---> plan/4
rule/1 ---> principle/1
jump/1 ---> say/4
earnings/1 ---> earnings/1
firm/1 ---> operation/1
line/1 ---> line/1
city/1 ---> city/1
```

References

[1] Richard Beckwith, Christiane Fellbaum, Derek Gross, and George A. Miller. Wordnet: A lexical database organized on psycholinguistic principles. In Uri Zernik, editor, *Lexical Acquisition – Exploiting On-Line Resources to Build a Lexicon*, chapter 9, pages 211–232. Lawrence Erlbaum Associates, Hillsdale - New Jersey, 1991.

[2] Eugene Charniak. *Statistical Language Learning*. MIT Press, 1993.

[3] A.M. Collins and M.R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–247, 1969.

[4] Trevor A. Harley. *The Psychology of Language – From Data to Theory*. Psychology Press, 1st edition, 1995.

[5] John A. Hawkins. *Definiteness and Indefiniteness*. Croom Helm, London, 1978.

[6] Christopher Huckle. *Unsupervized Categorization of Word Meanings using Statistics and Neural Network Models*. PhD thesis, University of Edinburgh, 1996.

[7] Kevin Lund, Curt Burgess, and Ruth Ann Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society of America*, pages 660–665, 1995.

[8] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. In Susan Armstrong, editor, *Using Large Corpora*, pages 273–290. MIT Press, Cambridge - London, 1994.

[9] Scott McDonald. Exploring the validity of corpus-derived measures of semantic similarity. 1997.

[10] D.E. Meyer and R.W. Schvaneveldt. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90:227–235, 1971.

[11] Massimo Poesio, Renata Vieira, and Simone Teufel. Bridging references in definite description resolution. In *Proceedings of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, 1997.

[12] E.F. Prince. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–256. Academic Press, 1981.

[13] Renata Vieira and Massimo Poesio. Corpus-based processing of definite descriptions. In Botley and McEnery, editors, *Corpus-based and Computational Approaches to Anaphora*. Forthcoming.

[14] Renata Vieira and Simone Teufel. Towards resolution of bridging descriptions. In *Proceedings of the ACL-EACL'97 Joint Conference: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997.