LEARNING TO RESOLVE BRIDGING REFERENCES

Massimo Poesio,[♣] Rahul Mehta,[♣] Axel Maroudas,[♣] and Janet Hitzeman[♠]
[♣]Dept. of Comp. Science, University of Essex, UK `poesio at essex dot ac dot uk`
[♠]MITRE Corporation, USA `hitz at mitre dot org`

Abstract

We use machine learning techniques to find the best combination of local focus and lexical distance features for identifying the anchor of mereological bridging references. We find that using first mention, utterance distance, and lexical distance computed using either Google or WordNet results in an accuracy significantly higher than obtained in previous experiments.

1 Introduction

BRIDGING REFERENCES (BR) (Clark, 1977)–anaphoric expressions that cannot be resolved purely on the basis of string matching and thus require the reader to 'bridge' the gap using commonsense inferences–are arguably the most interesting and, at the same time, the most challenging problem in anaphora resolution. Work such as (Poesio et al., 1998; Poesio et al., 2002; Poesio, 2003) provided an experimental confirmation of the hypothesis first put forward by Sidner (1979) that BRIDGING DESCRIPTIONS (BD)[1] are more similar to pronouns than to other types of definite descriptions, in that they are sensitive to the *local* rather than the *global* focus (Grosz and Sidner, 1986). This previuous work also suggested that simply choosing the entity whose description is lexically closest to that of the bridging description among those in the current focus space gives poor results; in fact, better results are obtained by always choosing as ANCHOR of the bridging reference[2] the first-mentioned entity of the previous sentence (Poesio, 2003). But neither source of information in isolation resulted in an accuracy over 40%. In short, this earlier work suggested that a combination of salience and lexical / commonsense information is needed to choose the most likely anchor; the problem remained of how to combine this information.

In the work described in this paper, we used machine learning techniques to find the best combination of local focus features and lexical distance features, focusing on MEREOLOGICAL bridging references:[3] references referring to parts of an object already introduced (*the cabinet*), such as *the panels* or *the top* (underlined) in the following example from the GNOME corpus (Poesio et al., 2004).

(1) The combination of rare and expensive materials used on [this cabinet]$_i$ indicates that it was a particularly expensive commission. The four Japanese lacquer panels date from the mid- to late 1600s and were created with a technique known as *kijimaki-e*.

> For this type of lacquer, artisans sanded plain wood to heighten its strong grain and used it as the background of each panel. They then added the scenic elements of landscape, plants, and animals in raised lacquer. Although this technique was common in Japan, such large panels were rarely incorporated into French eighteenth-century furniture.
>
> Heavy Ionic pilasters, whose copper-filled flutes give an added rich color and contrast to the gilt-bronze mounts, flank the panels. Yellow jasper, a semiprecious stone, rather than the usual marble, forms the top.

2 Two sources of information for bridging reference resolution

2.1 Lexical information

The use of different sources of lexical knowledge for resolving bridging references has been investigated in a series of papers by Poesio *et al.* all using as dataset the Bridging Descriptions (BDs) contained in the corpus used by Vieira and Poesio

[1] We will use the term bridging descriptions to indicate bridging references realized by definite descriptions, equated here with noun phrases with determiner *the*, like *the top*.

[2] Following (Poesio and Vieira, 1998), we use the term 'anchor' as as a generalization of the term ANTECEDENT, to indicate the discourse entity which an anaphoric expression either realizes, or is related to by an associative relation; reserving 'antecedent' for the cases of identity.

[3] We make use of the classification of bridging references proposed by Vieira and Poesio (2000). 'Mereological' bridging references are one of the the 'WordNet' bridging classes, which cover cases where the information required to bridge the gap may be found in a resource such as WordNet (Fellbaum, 1998): synonymy, hyponymy, and meronymy.

(2000). In these studies, the lexical distance between a BD and its antecedent was used to choose the anchor for the BD among the antecedents in the previous five sentences. In (Poesio et al., 1997; Vieira and Poesio, 2000) WordNet 1.6 was used as a lexical resource, with poor or mediocre results. These results were due in part to missing entries and / or relations; in part to the fact that because of the monotonic organization of information in WordNet, complex searches are required even to find apparently close associations (like that between *wheel* and *car*). Similar results using WordNet 1.6 were reported at around the same time by other groups - e.g., (Humphreys et al., 1997; Harabagiu and Moldovan, 1998) and have been confirmed by more recent studies studying both hyponymy (Markert et al., 2003) and more specifically mereological BDs. Poesio (2003) found that none of the 58 mereological references in the GNOME corpus (discussed below) had a direct mereological link to their anchor: for example, *table* is not listed as a possible holonym of *drawer*, nor is *house* listed as a possible holonym for *furniture*. Garcia-Almanza (2003) found that only 16 of these 58 mereological references could be resolved by means of more complex searches in WordNet, including following the hypernymy hierarchy for both the anchor and the bridging reference, and a 'spreading activation' search.

Poesio et al. (1998) explored the usefulness of vector-space representations of lexical meaning for BDs that depended on lexical knowledge about hyponymy and synonymy. The HAL model discussed in Lund et al. (1995) was used to find the anchor of the BDs in the dataset already used by Poesio et al. (1997). However, using vectorial representations did not improve the results for the 'WordNet' BDs: for the synonymy cases the results were comparable to those obtained with WordNet (4/12, 33%), but for the hyponymy BDs (2/14, as opposed to 8/14 with WordNet) and especially for mereological references (2/12) they were clearly worse. On the other hand, the post-hoc analysis of results suggested that the poor results were in part due to the lack of mechanisms for choosing the most salient (or most recent) BDs.

The poor results for mereological BDs with both WordNet and vectorial representations indicated that a different approach was needed to acquire information about part-of relations. Grefenstette's work on semantic similarity (Grefenstette, 1993) and Hearst's work on acquiring taxonomic information (Hearst, 1998) suggested that certain syntactic constructions could be usefully viewed as reflecting underlying semantic relations. In (Ishikawa, 1998; Poesio et al., 2002) it was proposed that syntactic patterns (henceforth: CONSTRUCTIONS) such as *the wheel of the car* could indicate that *wheel* and *car* stood in a part-of relation.[4] Vector-based lexical representations whose elements encoded the strength of associations identified by means of constructions like the one discussed were constructed from the British National Corpus, using Abney's CASS chunker. These representations were then used to choose the anchor of BDs, using again the same dataset and the same methods as in the previous two attempts, and using mutual information to determine the strength of association. The results on mereological BDs–recall .67, precision=.73–were drastically better than those obtained with WordNet or with simple vectorial representations. The results with the three types of lexical resources and the different types of BDs in the Vieira / Poesio dataset are summarized in Table 1.

Finally, a number of researchers recently argued for using the Web as a way of addressing data sparseness (Keller and Lapata, 2003). The Web has proven a useful resource for work in anaphora resolution as well. Uryupina (2003) used the Web to estimate 'Definiteness probabilities' used as a feature to identify discourse-new definites. Markert et al. (2003) used the Web and the construction method to extract information about hyponymy used to resolve *other*-anaphora (achieving an f value of around 67%) as well as the BDs in the Vieira-Poesio dataset (their results for these cases were not better than those obtained by (Vieira and Poesio, 2000)). Markert *et al.* also found a sharp difference between using the Web as a a corpus and using the BNC, the results in the latter case being significantly worse than when using WordNet. Poesio (2003) used the Web to choose between the hypotheses concerning the anchors of mereological BDs in the GNOME corpus generated on the basis of Centering information (see below).

2.2 Salience

One of the motivations behind Grosz and Sidner's (1986) distinction between two aspects of the attentional state - the LOCAL FOCUS and the GLOBAL FOCUS–is the difference between the interpretive preferences of pronouns and definite descriptions. According to Grosz and Sidner, the interpretation for pronouns is preferentially found in the local focus, whereas that of definite descriptions is preferentially found in the global focus.

[4] A similar approach was pursued in parallel by Berland and Charniak (1999).

	Synonymy	**Hyponymy**	**Meronymy**	**Total WN**	**Total BDs**
BDs in Vieira / Poesio corpus	12	14	12	38	204
Using WordNet	4 (33.3%)	8(57.1%)	3(33.3%)	15 (39%)	34 (16.7%)
Using HAL Lexicon	4 (33.3%)	2(14.3%)	2(16.7%)	8 (22.2%)	46(22.7%)
Using Construction Lexicon	1 (8.3%)	0	8(66.7%)	9 (23.7%)	34(16.7%)

Table 1: BD resolution results using only lexical distance with WordNet, HAL-style vectorial lexicon, and construction-based lexicon.

However, already Sidner (1979) hypothesized that BDs are different from other definite descriptions, in that the local focus is preferred for their interpretation. As already mentioned, the error analysis of Poesio et al. (1998) supported this finding: the study found that the strategy found to be optimal for anaphoric definite descriptions by Vieira and Poesio (2000), considering as equally likely all antecedents in the previous five-sentence window (as opposed to preferring closer antecedents), gave poor results for bridging references; entities introduced in the last two sentences and 'main entities' were clearly preferred. The following example illustrates how the local focus affects the interpretation of a mereological BD, *the sides*, in the third sentence.

(2) [Cartonnier (Filing Cabinet)]$_i$ with Clock

[This piece of mid-eighteenth-century furniture]$_i$ was meant to be used like a modern filing cabinet; papers were placed in [leather-fronted cardboard boxes]$_j$ (now missing) that were fitted into the open shelves.

[A large table]$_k$ decorated in the same manner would have been placed in front for working with those papers.

Access to [the cartonnier]$_i$'s lower half can only be gained by the doors at the sides, because the table would have blocked the front.

The three main candidate anchors in this example– the cabinet, the boxes, and the table–all have sides. However, the actual anchor, the cabinet, is clearly the Backward-Looking Center (CB) (Grosz et al., 1995) of the first sentence after the title;[5] and if we assume that entities can be indirectly realized– see (Poesio et al., 2004)–the cabinet is the CB of all three sentences, including the one containing the BR, and therefore a preferred candidate.

In (Poesio, 2003), the impact on associative BD resolution of both relatively simple salience features (such as distance and order or mention) and of more complex ones (such as whether the anchor was a CB or not) was studied using the GNOME corpus (discussed below) and the CB-tracking techniques developed to compare alternative ways of instantiating the parameters of Centering by Poesio et al. (2004). Poesio (2003) analyzed, first of all, the distance between the BD and the closest mention of the anchor, finding that of the 169 associative BDs, 77.5% had an anchor occurring either in the same sentence (59) or the previous one (72); and that only 4.2% of anchors were realized more than 5 sentences back. These percentages are very similar to those found with pronouns (Hobbs, 1978).

Next, Poesio analyzed the order of mention of the anchors of the 72 associative BD whose anchor was in the previous sentence, finding that 49/72, 68%, were realized in first position. This finding is consistent with the preference for first-mentioned entities (as opposed to the most recent ones) repeatedly observed in the psychological literature on anaphora (Gernsbacher and Hargreaves, 1988; Gordon et al., 1993). Finally, Poesio examined the hypothesis that finding the anchor of a BD involves knowing which entities are the CB and the CP in the sense of Centering (Grosz et al., 1995). He found that CB(U-1) is the anchor of 37/72 of the BDs whose anchor is in the previous utterance (51.3%), and only 33.6% overall. (CP(U-1) was the anchor for 38.2% associative BDs.) Clearly, simply choosing the CB (or the CP) of the previous sentence as the anchor doesn't work very well. However, Poesio also found that 89% of the anchors of associative BDs had been CBs or CPs. This suggested that while knowing the local focus isn't sufficient to determine the anchor of a BD, restricting the search for anchors to CBs and CPs only might increase the precision of the BD resolution process. This hypothesis was supported by a preliminary test with 20 associative BDs. The anchor for a BD with head noun NBD was chosen among the subset of all potential antecedents (PA) in the previous five sentences that had been CBs or CPs by calling Google (by hand) with the query "the NBD of the NPA", where NPA is the head noun of the potential antecedent, and choosing the PA with the highest hit count. 14 mereological BDs (70%) were resolved correctly this way.

3 Methods

The results just discussed suggest that lexical information and salience information combine to deter-

[5] The CB is Centering theory's (Grosz et al., 1995) implementation of the notion of 'topic' or 'main entity'.

mine the anchor of associative BRs. The goal of the experiments discussed in this paper was to test more thoroughly this hypothesis using machine learning techniques to combine the two types of information, using a larger dataset than used in this previous work, and using completely automatic techniques. We concentrated on mereological BDs, but our methods could be used to study other types of bridging references, using, e.g., the constructions used by Markert et al. (2003).[6]

3.1 The corpus

We used for these experiments the GNOME corpus, already used in (Poesio, 2003). An important property of this corpus for the purpose of studying BR resolution is that fewer types of BDs are annotated than in the original Vieira / Poesio dataset, but the annotation is reliable (Poesio et al., 2004).[7] The corpus also contains more mereological BDs and BRs than the original dataset used by Poesio and Vieira.

The GNOME corpus contains about 500 sentences and 3000 NPs. A variety of semantic and discourse information has been annotated (the manual is available from the GNOME project's home page at `http://www.hcrc.ed.ac.uk/ ˜ gnome`). Four types of anaphoric relations were annotated: identity (`IDENT`), set membership (`ELEMENT`), subset (`SUBSET`), and 'generalized possession' (`POSS`), which also includes part-of relations. A total of 2073 anaphoric relations were annotated; these include 1164 identity relations (including those realized with synonyms and hyponyms) and 153 `POSS` relations.

Bridging references are realized by noun phrases of different types, including indefinites (as in *I bought a book and a page fell out* (Prince, 1981)). Of the 153 mereological references, 58 mereological references are realized by definite descriptions.

3.2 Features

Our classifiers use two types of input features.

Lexical features Only one lexical feature was used: lexical distance, but extracted from two different lexical sources.

Google distance was computed as in (Poesio, 2003) (see also Markert et al. (2003)): given head nouns NBD of the BD and NPA of a potential antecedent, Google is called (via the Google API) with a query of the form "the NBD of the NPA" (e.g., *the sides of the table*) and the number of hits *NHits* is computed. Then

$$\textbf{Google distance} = \begin{cases} 1 & \text{if } \mathit{NHits} = 0 \\ \frac{1}{\mathit{NHits}} & \text{otherwise} \end{cases}$$

The query "the NBD of NPA" (e.g., *the amount of cream*) is used when NPA is used as a mass noun (information about mass vs count is annotated in the GNOME corpus). If the potential antecedent is a pronoun, the head of the closest realization of the same discourse entity is used.

We also reconsidered WordNet (1.7.1) as an alternative way of establishing lexical distance, but made a crucial change from the studies reported above. Both earlier studies such as (Poesio et al., 1997) and more recent ones (Poesio, 2003; Garcia-Almanza, 2003) had shown that mereological information in WordNet is extremely sparse. However, these studies also showed that information about hypernyms is much more extensive. This suggested trading precision for recall with an alternative way of using WordNet to compute lexical distance: instead of requiring the path between the head predicate of the associative BD and the head predicate of the potential antecedent to contain at least one mereological link (various strategies for performing a search of this type were considered in (Garcia-Almanza, 2003)), consider only hypernymy and hyponymy links.

To compute our second measure of lexical distance between NBD and NPA defined as above, **WordNet distance**, the following algorithm was used. Let **distance**(s, s') be the number of hypernim links between concepts s and s'. Then

1. Get from WordNet all the senses of both NBD and NPA;
2. Get the hypernym tree of each of these senses;
3. For each pair of senses s_{NBD_i} and s_{NPA_j}, find the Most Specific Common Subsumer s^{comm}_{ij} (this is the closest concept which is an hypernym of both senses).

[6] In (Poesio, 2003), bridging descriptions based on set relations (element, subset) were also considered, but we found that this class of BDs required completely different methods.

[7] A serious problem when working with bridging references is the fact that subjects, when asked for judgments about bridging references in general, have a great deal of difficulty in agreeing on which expressions in the corpus are bridging references, and what their anchors are (Poesio and Vieira, 1998). This finding raises a number of interesting theoretical questions concerning the extent of agreement on semantic judgments, but also the practical question of whether it is possible to evaluate the performance of a system on this task. Subsequent work found, however, that restricting the type of bridging inferences required does make it possible for annotators to agree among themselves (Poesio et al., 2004). In the GNOME corpus only a few types of associative relations are marked, but these can be marked reliably, and do include part-of relations like that between *the top* and *the cabinet* that we are concerned with.

4. The *ShortestWNDistance* between NBD and NPA is then computed as the shortest distance between any of the senses of NBD and any of the senses of NPA:

   $$\textit{ShtstWNDist}(NBD, NPA) =$$
   $$\min_{i,j}(\textbf{distance}(s_{NBD_i}, s^{com}_{ij}) + \textbf{distance}(s^{com}_{ij}, s_{NPA_j}))$$

5. Finally, a normalized **WordNet distance** in the range 0..1 is then obtained by dividing *ShtstWNDist* by a *MaxWNDist* factor (30 in our experiments). **WordNet distance** = 1 if no path between the concepts was found.

$$\textbf{WN distance} = \begin{cases} 1 & \text{if no path} \\ \frac{\textit{ShtstWNDist}}{\textit{MaxWNDist}} & \text{otherwise} \end{cases}$$

Salience features In choosing the salience features we took into account the results in (Poesio, 2003), but we only used features that were easy to compute, hoping that they would approximate the more complex features used in (Poesio, 2003). The first of these features was **utterance distance**, the distance between the utterance in which the BR occurs and the utterance containing the potential antecedent. (Sentences are used as utterances, as suggested by the results of (Poesio et al., 2004).) As discussed above, studies such as (Poesio, 2003) suggested that bridging references were sensitive to distance, in the same way as pronouns (Hobbs, 1978; Clark and Sengul, 1979). This finding was confirmed in our study; all anchors of the 58 mereological BDs occurred within the previous five sentences, and 47/58 (81%) in the previous two. (It is interesting to note that no anchor occurred in the same sentence as the BD.)

The second salience feature was boolean: whether the potential antecedent had been realized in **first mention** position in a sentence (Poesio, 2003; Gernsbacher and Hargreaves, 1988; Gordon et al., 1993). Two forms of this feature were tried: **local first mention** (whether the entity had been realized in first position within the previous five sentences) and **global first mention** (whether it had been realized in first position anywhere). 269 entities are realized in first position in the five sentences preceding one of the 58 BDs; 298 entities are realized in first position anywhere in the preceding text. For 31/58 of the anchors of mereological BDs, 53.5%, **local first mention** = 1; **global first mention** = 1 for 33/58 of anchors, 56.9%.

3.3 Training Methods

Constructing the data set The data set used to train and test BR resolution consisted of a set of positive instances (the actual anchors of the mereological BRs) and a set of negative instances (other entities mentioned in the previous five sentences of the text). However, preliminary tests showed that simply including all potential antecedents as negative instances would make the data set too unbalanced, particularly when only bridging descriptions were considered: in this case we would have had 58 positive instances vs. 1672 negative ones. We therefore developed a parametric script that could create datasets with different positive / negative ratios - 1:1, 1:2, 1:3 - by including, with each positive instance, a varying number of negative instances (1, 2, 3, ...) randomly chosen among the other potential antecedents, the number of negative instances to be included for each positive one being a parameter chosen by the experimenter. We report the results obtained with 1:1 and 1:3 ratios.

The dataset thus constructed was used for both training and testing, by means of a 10-fold cross-validation.

Types of Classifiers Used Multi-layer perceptrons (MLPs) have been claimed to work well with small datasets; we tested both our own implementation of an MLP with back-propagation in MatLab 6.5, experimenting with different configurations, and an off-the-shelf MLP included in the Weka Machine Learning Library[8], Weka-NN. The best configuration for our own MLP proved to be one with a sigle hidden layer and 10 hidden nodes. We also used the implementation of a Naive Bayes classifier included in the Weka MLL, as Modjeska et al. (2003) reported good results.

4 Experimental Results

In the first series of experiments only mereological Bridging Descriptions were considered (i.e., only bridging references realized by *the*-NPs). In a second series of experiments we considered all 153 mereological BRs, including ones realized with indefinites. Finally, we tested a classifier trained on balanced data (1:1 and 1:3) to find the anchors of BDs among all possible anchors.

4.1 Experiment 1: Mereological descriptions

The GNOME corpus contains 58 mereological BDs. The five sentences preceding these 58 BDs contain a total of 1511 distinct entities for which a head could be recovered, possibly by examining their antecedents. This means an average of 26 distinct potential antecedents per BD, and 5.2 entities per sentence. The simplest baselines for the task of finding

[8] The library is available from `http://www.cs.waikato.ac.nz/ml/weka/`.

the anchor are therefore 4% (by randomly choosing one antecedent among those in the previous five sentences) and 19.2% (by randomly choosing one antecedent among those in the previous sentence only). As 4.6 entities on average were realized in first mention position in the five sentences preceding a BD (269/58), choosing randomly among the first-mentioned entities gives a slighly higher accuracy of 21.3%.

A few further baselines can be established by examining each feature separately. Google didn't return any hits for 1089 out of 1511 distinct PAs, and no hit for 24/58 anchors; in 8/58 of cases (13.8%) the entity with the minimum Google distance is the correct anchor. We saw before that the method for computing WordNet distance used in (Poesio, 2003) didn't find a path for any of the mereological BDs; however, not trying to follow mereological links worked much better, achieving the same accuracy as Google distance (8/58, 13.8%) and finding connections for much higher percentages of concepts: no path could be found for only 10/58 of actual anchors, and for 503/1511 potential antecedents.

Pairwise combinations of these features were also considered. The best such combination, choosing the first mentioned entity in the previous sentence, achieves an accuracy of 18/58, 31%. These baseline results are summarized in the following table. Notice how even the best baselines achieve pretty low accuracy, and how even simple 'salience' measures work better than lexical distance measures.

Baseline	**Accuracy**
Random choice between entities in previous 5	4%
Random choice between entities in previous 1	19%
Random choice between First Ment. entities in previous 5	21.3%
Entity with min Google distance	13.8%
Entity with min WordNet distance	13.8%
FM entity in previous sentence	31%
Min Google distance in previous sentence	17.2%
Min WN distance in previous sentence	25.9%
FM and Min Google distance	12%
FM and Min WN distance	24.1%

Table 2: Baselines for the BD task

The features utterance distance, local first mention, and global f.m. were used in all machine learning experiments. But since one of our goals was to compare different lexical resources, only one lexical distance feature was used in the first two experiment.

The three classifiers were trained to classify a potential antecedent as either 'anchor' or 'not anchor'. The classification results with Google distance and WN distance for all three classifiers and the 1:1 data set (116 instances in total, 58 real anchor, 58 negative instances), for *all* elements of the data set, and averaging across the 10 cross-validations, are shown in Table 3.

	WN Distance (Correct)	**Google Distance** (Correct)
Our own MLP	92(79.3%)	89(76.7%)
Weka NN	91(78.4%)	86(74.1%)
Weka Naive Bayes	88(75.9%)	85(73.3%)

Table 3: Classification results for BDs

These results are clearly better than those obtained with any of the baseline methods discussed above. The differences between WN distance and Google distance, and that between our own MLP and the Weka implementation of Naive Bayes, are also significant (by a sign test, $p \leq .05$), whereas the pairwise differences between our own MLP and Weka's NN, and between this and the Naive Bayes classifier, aren't. In other words, although we find little difference between using WordNet and Google to compute lexical distance, using WordNet leads to slightly better results for BDs. The next table shows precision, recall and f-values for the positive data points, for the feature sets using WN distance and Google distance, respectively:

	Precision	Recall	F-value
WN features	75.4%	84.5%	79.6%
Google features	70.6%	86.2%	77.6%

Table 4: Precision and recall for positive instances

Using a 1:3 dataset (3 negative data points for each anchor), overall accuracy increases (to 82% using Google distance) and accuracy with Google distance is better than with Wordnet distance (80.6%); however, the precision and recall figures for the positive data points get much worse: 56.7% with Google, 55.7% with Wordnet.

4.2 All mereological references

Clearly, 58 positive instances is a fairly small dataset. In order to have a larger dataset, we included every bridging reference in the corpus, including those realized with indefinite NPs, thus bringing the total to 153 positive instances. We then ran a second series of experiments using the same methods as before. The results were slightly lower than those for BDs only, but in this case there was no difference between using Google and using WN. F-measure on positive instances was 76.3% with WN, 75.8% with Google.

4.3 A harder test

In a last experiment, we used classifiers trained on balanced and moderately unbalanced data to determine the anchor of 6 randomly chosen BDs among

	WN Distance (Correct)	**Google Distance** (Correct)
Weka NN	227(74.2%)	230(75.2%)

Table 5: Classification results for all BDs

all of their 346 possible antecedents in context. For these experiments, we also tried to use both Google and WordNet simultaneously. The results for BDs are shown in Table 6. The first column of the table specifies the lexical resource used; the second the degree of balance; the next two columns percentage correct and F value on a testing set with the same balance as the training set; the final two columns perc. correct and F value on the harder test set.

The best results,F=.5, are obtained using both Google and WN distance, and using a larger (if unbalanced) training corpus. These results are not as good as those obtained (by hand) by Poesio (which, however, used a complete focus tracking mechanism), but the F measure is still 66% higher than that obtained with the highest baseline (FM only), and not far off from the results obtained with direct anaphoric definite descriptions (e.g., by (Poesio and Alexandrov-Kabadjov, 2004)). It's also conforting to note that results with the harder test improve the more data are used, which suggests that better results could be obtained with a larger corpus.

5 Related work

In recent years there has been a lot of work to develop anaphora resolution algorithms using both symbolic and statistical methods that could be quantitatively evaluated (Humphreys et al., 1997; Ng and Cardie, 2002) but this work focused on identity relations; bridging references were explicitly excluded from the MUC coreference task because of the problems with reliability discussed earlier. Thus, most work on bridging has been theoretical, like the work by Asher and Lascarides (1998).

Apart from the work by Poesio *et al.*, the main other studies attempting quantitative evaluations of bridging reference resolution are (Markert et al., 1996; Markert et al., 2003). Markert et al. (1996) also argue for the need to use both Centering information and conceptual knowledge, and attempt to characterize the 'best' paths on the basis of an analysis of part-of relations, but use a hand-coded, domain-dependent knowledge base. Markert et al. (2003) focus on *other* anaphora, using Hearst' patterns to mine information about hyponymy from the Web, but do not use focusing knowledge.

6 Discussion and Conclusions

The two main results of this study are, first of all, that combining 'salience' features with 'lexical' features leads to much better results than using either method in isolation; and that these results are an improvement over those previously reported in the literature. A secondary, but still interesting, result is that using WordNet in a different way –taking advantage of its extensive information about hypernyms to obviate its lack of information about meronymy–obviates the problems previously reported in the literature on using WordNet for resolving mereological bridging references, leading to results comparable to those obtained using Google. (Of course, from a practical perspective Google may still be preferrable, particularly for languages for which no WordNet exists.)

The main limitation of the present work is that the number of BDs and BRs considered, while larger than in our previous studies, is still fairly small. Unfortunately, creating a reasonably accurate gold standard for this type of semantic interpretation process is slow work. Our first priority will be therefore to extend the data set, including also the original cases studied by Poesio and Vieira.

Current and future work will also include incorporating the methods tested here in an actual anaphora resolution system, the GUITAR system (Poesio and Alexandrov-Kabadjov, 2004). We are also working on methods for automatically recognizing bridging descriptions, and dealing with other types of (non-associative) bridging references based on synonymy and hyponymy.

Acknowledgments

The creation of the GNOME corpus was supported by the EPSRC project GNOME, GR/L51126/01.

References

N. Asher and A. Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–13.

M. Berland and E. Charniak. 1999. Finding parts in very large corpora. In *Proc. of the 37th ACL.*

H. H. Clark and C. J. Sengul. 1979. In search of referents for nouns and pronouns. *Memory and Cognition*, 7(1):35–41.

H. H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason, editors, *Thinking: Readings in Cognitive Science*. Cambridge.

C. Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. The MIT Press.

A. Garcia-Almanza. 2003. Using WordNet for mereological anaphora resolution. Master's thesis, University of Essex.

Lex Res	**Balance**	**Perc on bal**	**F on bal**	**Perc on Hard**	**F on Hard**
WN	1:1	70.2%	.7	80.2%	.2
	1:3	75.9%	.4	91.7%	0
Google	1:1	64.4%	.7	63.6%	.1
	1.3	79.8%	.5	88.4%	.3
WN + Google	1:1	66.3%	.6	65.3%	.2
	1.3	77.9%	.4	92.5%	**.5**

Table 6: Results using a classifier trained on balanced data on unbalanced ones.

M. A. Gernsbacher and D. Hargreaves. 1988. Accessing sentence participants. *Journal of Memory and Language*, 27:699–717.

P. C. Gordon, B. J. Grosz, and L. A. Gillion. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–348.

G. Grefenstette. 1993. SEXTANT: extracting semantics from raw text. *Heuristics*.

B. J. Grosz and C. L. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

B. J. Grosz, A. K. Joshi, and S. Weinstein. 1995. Centering. *Computational Linguistics*, 21(2):202–225.

S. Harabagiu and D. Moldovan. 1998. Knowledge processing on extended WordNet. In (Fellbaum, 1998), pages 379–405.

M. A. Hearst. 1998. Automated discovery of Wordnet relations. In (Fellbaum, 1998).

J. R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, and H. Cunningham Y. Wilks. 1997. Description of the LaSIE-II System as used for MUC-7. In *Proc. of the 7th Message Understanding Conference (MUC-7)*.

T. Ishikawa. 1998. Acquisition of associative information and resolution of bridging descriptions. Master's thesis, University of Edinburgh.

F. Keller and M. Lapata. 2003. Using the Web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3).

K. Lund, C. Burgess, and R. A. Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Conf. of the Cogn. Science Soc.*, pages 660–665.

K. Markert, M. Strube, and U. Hahn. 1996. Inferential realization constraints on functional anaphora in the centering model. In *Proc. of 18th Conf. of the Cog. Science Soc.*, pages 609–614.

K. Markert, M. Nissim, and N.. Modjeska. 2003. Using the Web for nominal anaphora resolution. In *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46.

N. Modjeska, K. Markert, and M. Nissim. 2003. Using the Web in ML for anaphora resolution. In *Proc. of EMNLP-03*, pages 176–183.

V. Ng and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Meeting of the ACL*.

M. Poesio and R. Vieira. 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216, June.

M. Poesio, R. Vieira, and S. Teufel. 1997. Resolving bridging references in unrestricted text. In R. Mitkov, editor, *Proc. of the ACL Workshop on Robust Anaphora Resolution*, pages 1–6, Madrid.

M. Poesio, S. Schulte im Walde, and C. Brew. 1998. Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*, pages 82–89.

M. Poesio, T. Ishikawa, S. Schulte im Walde, and R. Vieira. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proc. of the 3rd LREC*.

M. Poesio and M. Alexandrov-Kabadjov. 2004. A general-purpose, off the shelf anaphoric resolver. In *Proc. of the 4th LREC*, Lisbon.

M. Poesio, R. Stevenson, B. Di Eugenio, and J. M. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Comp. Linguistics*. 30(3).

M. Poesio. 2003. Associative descriptions and salience. In *Proc. of the EACL Workshop on Computational Treatments of Anaphora*.

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

C. L. Sidner. 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.

O. Uryupina. 2003. High-precision identification of discourse-new and unique noun phrases. In *Proc. of ACL 2003 Stud. Workshop*, pages 80–86.

R. Vieira and M. Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4), December.