

# ACQUIRING LEXICAL KNOWLEDGE FOR ANAPHORA RESOLUTION

Massimo Poesio,<sup>ℓ</sup> Tomonori Ishikawa,<sup>‡</sup> Sabine Schulte im Walde,<sup>\*</sup> and Renata Vieira<sup>†</sup>

<sup>ℓ</sup>University of Essex, Department of Computer Science

poesio@essex.ac.uk

<sup>‡</sup>University of Edinburgh, TAAL

<sup>\*</sup>University of Stuttgart, IMS

schulte@ims.uni-stuttgart.de

<sup>†</sup>Universidade do Vale do Rio dos Sinos

renata@exatas.unisinos.br

## Abstract

The lack of adequate bases of commonsense or even lexical knowledge is perhaps the main obstacle to the development of high-performance, robust tools for semantic interpretation. It is also generally accepted that, notwithstanding the increasing availability in recent years of substantial hand-coded lexical resources such as WordNet and EuroWordNet, addressing the commonsense knowledge bottleneck will eventually require the development of effective techniques for acquiring such information automatically, e.g., from corpora. We discuss research aimed at improving the performance of anaphora resolution systems by acquiring the commonsense knowledge require to resolve the more complex cases of anaphora, such as bridging references. We focus in particular on the problem of acquiring information about part-of relations.

## 1. Introduction

The lack of adequate bases of commonsense or even lexical knowledge is perhaps the main obstacle to the development of high-performance, robust tools for semantic interpretation (except for cases like pronoun interpretation, where a lot can be achieved on the basis of syntactic information only). It is also generally accepted that, notwithstanding the increasing availability in recent years of substantial hand-coded lexical resources such as WordNet and EuroWordNet, addressing the commonsense knowledge bottleneck will eventually require the development of effective techniques for acquiring such information automatically, e.g., from corpora. Most current work on lexical acquisition so far, however, has focused on acquiring the type of knowledge needed to improve the performance of parsers- e.g., subcategorization frames (Brent, 1993; Grefenstette, 1993; Manning, 1993; Resnik, 1993; Abney and Light, 1999) - rather than semantic interpreters. And a lot of work on lexical acquisition has not been shown to improve the performance in other semantic tasks.

The goal of our research is to improve the performance of anaphora resolution systems by acquiring the commonsense knowledge require to resolve the more complex cases of anaphora, such as bridging references. We also hope to acquire in the process insights into what kind of commonsense knowledge is actually needed for the task. In previous work, we developed a system for resolving definite descriptions that employed heuristical methods for identifying discourse new descriptions, and used a combination of access to WordNet 1.6 and heuristics to resolve bridging descriptions (Poesio and Vieira, 1998; Poesio et al., 1997; Vieira and Teufel, 1997; Vieira, 1998; Vieira and Poesio, 2000b; Vieira and Poesio, 2000a). This system was extensively evaluated, allowing us to identify where commonsense knowledge is actually needed, and what type - in particular, devising a classification for bridging descriptions depending on the sources of knowledge needed to solve

them (Poesio et al., 1997; Vieira and Teufel, 1997; Vieira, 1998). More recently, we have been using the data collected in this first project to improve the system in particular areas, and especially the resolution of bridging references. In this paper we discuss these more recent results.

## 2. A Heuristic-Based System for Robust Definite Description Resolution

The starting point for this discussion are the systems for resolving definite descriptions (expressions like *the car*) discussed in (Vieira and Poesio, 2000b; Vieira and Poesio, 2000a). A first important characteristic of this work was that the development of these systems was based on the extensive corpus analysis reported in (Poesio and Vieira, 1998); a second one was that the annotated corpus produced as a result of the analysis was then used to systematically evaluate the systems.

The corpus study discussed in (Poesio and Vieira, 1998) revealed that on average, our annotators classified more than half of all definite descriptions as first-mention- descriptions without an explicit antecedent in the previous discourse. Of these, about 22% were ‘larger-situation,’ such as *the Querecho Plains of New Mexico* in (1a); 22% ‘unfamiliar,’ like *the economic know-how to steer the city through a possible fiscal crisis* in (1b); and 8% ‘associative’ (or ‘bridging references’ (Hawkins, 1978; Clark, 1977)), like *the kitchen* in (1c).

- (1)
  - a. Out here on *the Querecho Plains of New Mexico*, however, the mood is more upbeat trucks rumble along the dusty roads and burly men in hard hats sweat and swear through the afternoon sun.
  - b. They wonder whether he has *the economic know-how to steer the city through a possible fiscal crisis*, and they wonder who will be advising him.

- c. Once inside, she spends nearly four hours measuring and diagramming each room in *the 80-year-old house*, gathering enough information to estimate what it would cost to rebuild it. While she works inside, a tenant returns with several friends to collect furniture and clothing. One of the friends sweeps broken dishes and shattered glass from a countertop and starts to pack what can be salvaged from *the kitchen*.

The result concerning the prevalence of discourse-new descriptions led us to develop heuristic methods for identifying them, based in part on the discussion in (Hawkins, 1978). Particularly effective heuristics were checking whether the head of the clause was a predicate like *fact*, *result*; and if the definite was post-modified by a relative clause. Overall, these heuristics achieved R=69% and P=72%. For anaphora, we developed heuristics for dealing with premodification, to avoid, e.g., a match between *the red car* and *the blue car*, while allowing *the house* to match *the 80-year-old Victorian house*. We also develop segmentation heuristics, to restrict the number of potential antecedents. Overall, the system achieved 62% recall and 83% precision on anaphoric same head descriptions.

Our work on bridging descriptions—the main topic of this paper—was based on a classification of bridging descriptions into classes depending on the sort of information needed to resolve them (Poesio et al., 1997; Vieira and Teufel, 1997; Vieira, 1998) which heavily relied on previous classifications proposed by (Clark, 1977; Sidner, 1979; Strand, 1996). We identified the following classes:

- cases based on well-defined lexical relations, such as synonymy, hypernymy and meronymy, that can be found in a lexical database such as WordNet (Fellbaum, 1998)—as in *the flat ... the living room*;
- bridging descriptions in which the antecedent is a proper name and the description a common noun, whose resolution requires some way of recognizing the type of object denoted by the proper name (as in *Bach ... the composer*);
- cases in which the anchor is not the head noun but a noun modifying an antecedent, as in *the company has been selling discount packages ... the discounts*
- cases in which the antecedent (anchor) is not introduced by an NP but by a VP, as in *Kadane oil is currently drilling two oil wells. The activity ...*
- descriptions whose the antecedent is not explicitly mentioned in the text, but is implicitly available because it is a discourse topic—e.g., *the industry* in a text referring to oil companies;
- cases in which the relation with the anchor is based on more general commonsense knowledge, e.g., about cause-consequence relations.

Our corpus contained 204 bridging descriptions, distributed among the classes above as follows:

Class	Total	Percentage
Syn / Hyp / Mer	12/14/12	19%
Names	49	24%
Compound Nouns	25	12%
Events	40	20%
Discourse Topic	15	7%
Inference	37	18%
Total	204	100%

The algorithms we proposed were based on a combination of access to WordNet 1.6 and heuristics (Poesio et al., 1997; Vieira and Teufel, 1997; Vieira and Poesio, 2000a). The system achieved R=29% and P=38% on bridging descriptions. If we only consider the resolutions due to information present in WordNet, the results are as follows:

Class	Total	Percentage
Syn / Hyp / Mer	4/8/3	39%
Names + Compound Names	19	25.7%
Total	34	16.7%

We found three types of problems with WordNet. First of all, there was a problem of missing data - certain words, like *crocidolite*, were not in the lexicon. Secondly, we found that certain lexical relations were context-dependent: e.g., *slump*, *crash* and *bust* are all virtually synonyms in the Wall Street Journal corpus, but not in WordNet. And finally, we found that in order to ensure monotonicity, information otherwise present in WordNet was sometimes located in positions that made it difficult to find. Thus, for example, WordNet contains the information that **floor** is part of **room**, but it does not contain the information that **rooms** are part of **houses** or **homes**, but only that they are part of **buildings**; so in order to resolve a bridging description *wall on house* involves a fairly complex search mechanisms. Our experiments indicated that while these mechanisms improve recall, they affect precision very badly.

Two versions of the system were evaluated. One version only resolved directly anaphoric definites and identified discourse-new descriptions; this system had R=53%, P=76%. Version II also attempted to find the anchors of bridging descriptions; this version had a higher recall (57%) but a lower precision (70%).

### 3. Addressing the commonsense knowledge bottleneck using lexical acquisition

The results just mentioned, while comparable to those obtained by other systems on the same task, are clearly not sufficient for any real application. The evaluation of the system pointed out a number of areas where improvement was needed to improve this performance.

The worse results by far were obtained on bridging descriptions. Part of the problem was that our system didn't include methods for resolving bridging descriptions based on 'causal' information or on thematic roles, but even on the 19% of bridging descriptions depending for their resolution on information contained in WordNet (synonymy, hyponymy, and meronymy) we only had 39% recall. Better lexical resources were needed for this purpose, but improving on these results was clearly going to be difficult.

One type of bridging descriptions in which we could expect to improve the performance of the system was definite descriptions referring to antecedents introduced by proper names. Our methods for resolving these bridging descriptions worked better than those for other classes (R=66%, P=95%) but state-of-the-art named entity recognition systems can achieve about 90% precision and recall (Mikheev et al., 1999).

Some researchers's solution to the shortcomings of WordNet has been to augment it by adding further hand-coded information (Harabagiu, 1997). In part because we had observed that many lexical relations were context-dependent, we decided instead to test vector-based methods for unsupervised lexical acquisition (Lund et al., 1995; Schütze, 1997). These methods are based on the assumption that the meaning of each (sense of a) word  $w$  is simply a vector of 'features'—which, in the simplest cases, are simply other words that occur in the vicinity of  $w$ . The reason for our interest in these methods were the results by Lund et al., (1995), who found a high correlation between the lexical associations acquired in this way and the lexical associations discovered by means of semantic priming experiments (Moss et al., 1995). Lund et al's results encouraged us to test whether the anchor for a bridging description could be found simply by finding the antecedent most strongly associated with that description—henceforth, the PRIMING HYPOTHESIS.<sup>1</sup> We also hoped that these methods could eventually be used to train domain-specific lexica.

In a series of experiments discussed in (Schulte im Walde, 1997; Poesio et al., 1998), we used the BNC corpus to acquire this type of lexical meanings for the bridging descriptions and their antecedents in the cases of bridging references tested in (Vieira and Poesio, 2000a). Part of our goal was to find the best values for the corpus acquisition method. We tried various window sizes, vicinity measures, and various types of corpus preprocessing, including lemmatization and tagging; the best results were obtained with the configuration using windows of size 10, lemmatized but untagged, and using the cosine metric as a vicinity measure. With this configuration, we found that the priming hypothesis didn't hold in the simple form sketched above—for only 22.7% of bridging descriptions the strongest lexical associate in the previous five sentences was also the correct anchor. The results per class of bridging descriptions obtained using the lexical knowledge bases trained this way are summarized in the following table:

Class	Total	Percentage
Syn / Hyp / Mer	4/2/2	22.2%
Names + Compound Nouns	17	23%
Events	5	16.7%
Discourse Topic	1	7%
Inference	6	13%
Total	46	22.7%

The following table compares instead these results with those obtained with WordNet (and no other heuristics) overall, and on the WordNet categories only:

Class	Total	WordNet	Vector-Based
Syn / Hyp / Mer	38	4/8/3 (39%)	4/2/2 (22.2%)
Overall	204	34 (16.7%)	46 (22.7%)

As the table shows, at least for synonymy the results with the automatically acquired lexica were comparable to those we had obtained with WordNet: the accuracy for synonymy was 36%, identical to that obtained with WordNet. The worse results were obtained with meronymy (accuracy =16.7% vs 25% with WordNet) and hyponymy (accuracy = 14.3%, vs. 57.1%) (Schulte im Walde, 1997; Poesio et al., 1998). And in all of these cases, we found that we were far from achieving the best possible performance, at least for synonymy: for example, doubling the size of the training corpus from 50M to 100M words increased the accuracy by almost 50%.

A case-by-case analysis suggested that while the results were not very good, part of the problem was the resolution method, and in particular the fact that we didn't keep track of the current local focus. The 'window' heuristic for tracking global salience was simultaneously too restrictive and not restrictive enough. On the one hand, fully 19.6% of the actual anchors were outside the 5-sentence window we were using. On the other hand, in a number of cases the anchor suggested by the algorithm is can be argued to be semantically closer than the actual antecedent, which is however the local focus or at least closer. In one case, the algorithm suggested the lexical associate *customer* as the antecedent of *market*, whereas the actual anchor is the (genre-specific) hyponym *phone service*. In an extreme case, the algorithm picks up *investigative companies* as antecedent for *the company*, whereas the actual antecedent is a specific company, *Pinkerton*. We classified each suggested resolution as either Acceptable, F (the resolution is arguably the closest semantic associate of the bridging description, but the 'correct' anchor is more in focus), Lexically plausible (although the desired one should have perhaps been classified as closer) and Wrong. 29% of resolutions were Acceptable, 21.2% F, 9.8% in class L, and 39.9% W. F

These results suggested, first, the need to integrate bridging resolution with a focus tracking mechanism; second, that at least for synonymy, these automatically acquired resources were comparable to WordNet, and increasing the accuracy might just be a matter of increasing the training corpus. On the other hand, we felt that these methods didn't work too well for hyponymy and meronymy. These considerations led us, on the one hand, to undertake extensive empirical investigations of local and global salience, discussed in (Poesio et al., 2000; Poesio et al., 2002). In order to improve our lexical resources, we considered using a mixture of lexical sources for the different types of bridging descriptions, acquired in different ways.

#### 4. Syntactic patterns and the acquisition of meronymic information

We were particularly interested in testing whether using syntactic information would help, given the results by (Grefenstette, 1993) and especially Hearst's work on automatically acquiring hyponymy information (Hearst, 1998).

<sup>1</sup>This possibility has been raised, among others, by Carter, (1987).

We concentrated on meronymy, since the acquisition of information about hyponymy has been intensively studied in the last years—see, e.g., (Caraballo, 1999). Again, we used the British National Corpus as our training corpus; and we used Abney’s CASS parser (Abney, 1993) to parse it.

We hypothesized that we would get better information about meronymy by taking as ‘mereological neighbors’ of a word *W* not all words occurring in its vicinity, but only those occurring in certain syntactic constructions, such as *the Z of W* or *Z’s W*. E.g., *window* would be considered a ‘mereological neighbor’ or *car* if it frequently occurred in constructions such as *the car’s window*. We identified the constructions of interest by following a method suggested by Hearst, (1998) –record the pairs of words occurring in the corpus in the *the Z of W* construction, and then look for other constructions in which these pairs of words occurred. By this method we identified the following constructions as potentially relevant:

**THE-NPs with OF:** *the NP of NP*

**BARE NPs with OF:** *NP of NP*

**Possessives:** *NP’s NP*

**Nominal Compounds:** *NP N*

We used the same texts as (Vieira and Poesio, 2000a; Poesio et al., 1998) to evaluate our models. The algorithm constructs for each head noun *N* of a bridging description in the test corpus a vector recording the frequency with which other words occur together with *N* in one of the constructions listed above. During evaluation, we use mutual information (Brown et al., 1992) to identify the closest ‘mereological antecedent’ of each bridging description among the antecedents contained in the five previous sentences:

$$I(x; y) = \log \frac{P(x, y)}{P(x)P(y)}$$

(for antecedents realized by proper names, we use the named entity software discussed in (Mikheev et al., 1999) to assign to each proper name a type like ORGANIZATION, PERSON or LOCATION, and then compute the mutual information between the bridging description and this type).

We ran a series of experiments to identify the most useful constructions—first using single constructions to train the lexical knowledge base, then combinations of them. The recall and precision figures for the four constructions for the case of meronymy are as follows:

Construction	Recall	Precision
the Z of W	7 (58.3%)	70%
Z of W	3 (25%)	50%
Possessives	2 (16.7%)	66.7%
Compound Nouns	0 (0/	

The best results were obtained dropping the ‘Compound Nouns’ construction and combining the lexical knowledge bases acquired with the other three constructions: in this way we obtained 66.7% recall and 72.7% precision on the mereological descriptions, as opposed to 25% using

WordNet and 16.7% using pure vector association obtained before. The overall comparisons with the other methods are as follows:

Class	Total	WordNet	Vector-Based	Constructions
Syn / Hyp / Mer	38	4/8/3 (39%)	4/2/2 (22.2%)	1/0/8 (23.7%)
Overall	204	34 (16.7%)	46 (22.7%)	34 (16.7%)

Further details in (Ishikawa, 1998).

## 5. Related Work

As mentioned above, the idea of using syntactic constructions to acquire relation-specific information was inspired by work by Hearst, (1998) on acquiring hyponyms. In parallel with us<sup>2</sup> Berland and Charniak, (1999) also attempted to extract part-of information by identifying constructions in corpora. The main advantage of that work with respect to ours is that log-likelihood was used instead of mutual information, which has well-known problems with rare events; on the other hand, the knowledge bases acquired by Berland and Charniak were not used to resolve anaphoric expressions, so we do not have direct ways of comparing the results.

## 6. Topics for Future Research and Conclusions

Addressing the problem of resolving bridging descriptions is a promising way to attack the problem of the use of lexical and commonsense knowledge in interpretation. This problem however is fairly hard: in part because ‘knowledge poor’ methods don’t work very well in this case, in part because it’s not clear how humans themselves deal with them.

Our experiences so far suggest that although our current results are far from satisfactory, lexical acquisition methods can be made to work for certain types of bridging descriptions, especially when a combination of methods is used to acquire the type of knowledge needed to resolve different types of bridging descriptions (traditional clustering methods work well for synonymy, but not other ‘WordNet relations’; using constructions works well for meronymy).

Longer term, our methods need to be supplemented with robust methods for focus tracking, and with ways for acquiring encyclopedic and causal knowledge.

## Acknowledgments

This work in part supported by an EPSRC Advanced Research Fellowship (Massimo Poesio). Massimo Poesio wishes to thank Chris Brew, Will Lowe, Scott MacDonald, Maria Teresa Pazienza, and Peter Wiemer-Hastings for comments. Thanks also to audiences at the University of Edinburgh, University of Essex, and Università di Roma Tor Vergata.

## 7. References

Abney, S., and M. Light, 1999. Hiding a Semantic Class Hierarchy in a Markov Model . In *Proc. of ACL Workshop on Unsupervised Learning in NLP*. University of Maryland.

<sup>2</sup>Our original experiments were part of (Ishikawa, 1998).

- Berland, M. and Charniak, E. 1999. Finding Parts in Very Large Corpora. In *Proc. of the 37th ACL*. University of Maryland.
- Brent, 1993. . From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(3):243–262.
- Brown, P., V. J. D. Della Pietra, P. V. DeSouza, J. C. Lai, and R. L. Mercer, 1992. Class-based n-grams models of natural language. *Computational Linguistics*, 18(4):467–479.
- Carballo, S. A., 1999. Automatic acquisition of a hypernym-labeled noun hierarchy from text. In *Proc. of the ACL*.
- Carter, D. M., 1987. *Interpreting Anaphors in Natural Language Texts*. Chichester, UK: Ellis Horwood.
- Clark, H. H., 1977. Bridging. In P. N. Johnson-Laird and P.C. Wason (eds.), *Thinking: Readings in Cognitive Science*. London and New York: Cambridge University Press.
- Fellbaum, Christiane, editor. 1998. *WordNet: An electronic lexical database*. The MIT Press.
- Grefenstette, G., 1993. SEXTANT: extracting semantics from raw text. *Heuristics*.
- Harabagiu, S., 1997. *WordNet-Based Inference of Textual Context, Cohesion and Coherence*. PhD Dissertation, University of Southern California.
- Hawkins, J. A., 1978. *Definiteness and Indefiniteness*. London: Croom Helm.
- Hearst, M. A., 1998. Automated discovery of wordnet relations. In C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*. MIT Press.
- Ishikawa, T., 1998. *Acquisition of Associative Information and Resolution of Bridging Descriptions*. MSc thesis, University of Edinburgh, Department of Linguistics.
- Lappin, S. and H. J. Leass, 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.
- Lund, K., C. Burgess, and R. A. Atchley, 1995. Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*.
- Manning, C. D., 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proc. of the 31st ACL*. p. 235–242.
- Mikheev, A., M. Moens, and C. Grover, 1999. Named Entity recognition without gazetteers. In *Proc. of EACL*. Bergen, Norway: EACL.
- Mitkov, R., 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 18th COLING*. Montreal.
- Morton, T., 2000. Coreference for NLP applications. In *Proc. of the 38th ACL*. Hong Kong.
- Moss, H. E., R. K. Ostrin, L. K. Tyler, and W. D. Marslen-Wilson, 1995. Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*:863–883.
- Poesio, M., H. Cheng, B. Di Eugenio, J. M. Hitzeman, and R. Stevenson, 2002. A corpus-based evaluation of centering theory. Submitted.
- Poesio, M., H. Cheng, R. Henschel, J. M. Hitzeman, R. Kibble, and R. Stevenson, 2000. Specifying the parameters of Centering Theory: a corpus-based evaluation using text from application-oriented domains. In *Proc. of the 38th ACL*. Hong Kong.
- Poesio, M., 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*. Stanford, CA: AAAI.
- Poesio, M. and R. Vieira, 1998. A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216. Also available as Research Paper CCS-RP-71, Centre for Cognitive Science, University of Edinburgh.
- Poesio, M., R. Vieira, and S. Teufel, 1997. Resolving bridging references in unrestricted text. In R. Mitkov (ed.), *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*. Madrid. Also available as HCRC Research Paper HCRC/RP-87, University of Edinburgh.
- Resnik, P., 1993. *Selection and information: a class-based approach to lexical relationships*. PhD Dissertation, University of Pennsylvania.
- Schulte im Walde, S., 1997. *Resolving bridging descriptions in high-dimensional space*. Studienarbeit, Universities of Stuttgart and Edinburgh.
- Schütze, H., 1997. *Ambiguity Resolution in Language Learning*. Stanford: CSLI.
- Sidner, C. L., 1979. *Towards a computational theory of definite anaphora comprehension in English discourse*. Ph.D. thesis, MIT.
- Strand, K., 1996. A taxonomy of linking relations. Manuscript. A preliminary version presented at the Workshop on Indirect Anaphora, Lancaster University, 1996.
- Tetreault, J. R., 2001. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520.
- Vieira, R., 1998. *Definite Description Resolution in Unrestricted Texts*. Ph.D. thesis, University of Edinburgh, Centre for Cognitive Science.
- Vieira, R. and M. Poesio, 2000a. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4).
- Vieira, R. and M. Poesio, 2000b. Processing definite descriptions in corpora. In S. Botley and T. McEnery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, chapter 10. Amsterdam / New York: John Benjamins, pages 189–212.
- Vieira, R. and S. Teufel, 1997. Towards resolution of bridging descriptions. In *Proc. of the 35th Joint Meeting of the Association for Computational Linguistics*. Madrid.