Word Sense Disambiguation: Why Statistics When We Have These Numbers?

Kavi Mahesh, Sergei Nirenburg, Stephen Beale,
Evelyne Viegas, and Victor Raskin
Computing Research Laboratory (3CRL)
New Mexico State University
Las Cruces, NM 88003, USA
{mahesh,sergei,sb,viegas,raskin}@crl.nmsu.edu

Boyan Onyshkevych

US Dept. of Defense
ATTN. R525
Fort Meade, Maryland, 20755 USA
baonysh@afterlife.ncsc.mil

http://crl.nmsu.edu/Research/Projects/mikro/

Abstract. Word sense disambiguation continues to be a difficult problem in machine translation (MT). Current methods either demand large amounts of corpus data and training or rely on knowledge of hard selectional constraints. In either case, the methods have been demonstrated only on a small scale and mostly in isolation, where disambiguation is a task by itself. It is not clear that the methods can be scaled up and integrated with other components of analysis and generation that constitute an end-to-end MT system. In this paper, we illustrate how the Mikrokosmos Knowledge-Based MT system disambiguates word senses in real-world texts with a very high degree of correctness. Disambiguation in Mikrokosmos is achieved by a combination of (i) a broad-coverage ontology with many selectional constraints per concept, (ii) a large computational-semantic lexicon grounded in the ontology, (iii) an optimized search algorithm for checking selectional constraints in the ontology, and (iv) an efficient control mechanism with near-linear processing complexity. Moreover, Mikrokosmos constructs complete meaning representations of an input text using the chosen word senses.

1 Word Sense Ambiguity

Word sense disambiguation continues to be a difficult problem for machine translation (MT) systems.The most common current methods for resolving word sense ambiguities are based on statistical collocations or static selectional preferences between pairs of word senses. The real power of word sense selection seems to lie in the ability to constrain the possible senses of a word based on selections made for other words in the local context. Although methods using selectional constraints and semantic networks have been delineated at least since Katz and Fodor (1963), computational models have not demonstrated the effectiveness of knowledge-based methods in resolving word senses in real-world texts on a large scale. This has resulted in a predominant shift of attention from knowledge-based to corpus-based, statistical methods for word sense resolution, despite the far greater potential of knowledge-based methods for advancing the development of large, practical, domain independent NLP/MT systems.[1]

In this article, we illustrate how the semantic analyzer of the Mikrokosmos machine translation system resolves word sense ambiguities in real-world Spanish texts (news articles on company mergers and acquisitions from the EFE newswire) with a high degree of correctness. We begin by presenting the results from Mikrokosmos and then illustrate how they were obtained.

[1] See Guthrie et al (1996) and Wilks et al (1995) for recent surveys of related work.

Text	#1	#2	#3	#4	Average
# words	347	385	370	353	364
# words/sentence	16.5	24.0	26.4	20.8	21.4
# open-class words	183	167	177	177	176
# ambiguous open-class words	57	42	57	35	48
# resolved by syntax	21	19	20	12	18
total # correctly resolved	51	41	45	34	43
% correct	97%	99%	93%	99%	97%

Table 1. Mikrokosmos Results in Disambiguating Open Class Words in Spanish Texts.

# words	390
# words/sentence	26
# open-class words	104
# ambiguous open-class words	26
# resolved by syntax	9
total # correctly resolved	23
% correct	97.1%

Table 2. Mikrokosmos Results on an Unseen Text.

2 Results

Experiment 1: Mikrokosmos semantic analyzer applied on 4 out of 400 Spanish texts used in knowledge acquisition.

Table 1 shows sample disambiguation results from Mikrokosmos. These are results from analyzing four real-world texts. The average text was 17 sentence long, with over 21 words per sentence. For evaluation purposes, correct senses for all the open class words in the texts were determined by a native speaker. Mikrokosmos selects the right sense of open-class words about 97% of the time.

The performance on the first and third texts was worse than the performance on the other two texts. The first and third texts had longer sentences, many more ambiguous words, and constructs that make disambiguation hard (e.g., ambiguous words embedded in appositions). Moreover, just a handful of difficult words led to significantly worse performance in these texts. For example, the Spanish word "operacion" occurred several times in these texts and was hard to disambiguate between its WORK-ACTIVITY, MILITARY-OPERATION, SURGERY, and FINANCIAL-TRANSACTION senses (although the SURGERY sense was easily eliminated).

Syntactic analysis contributed to about 38% of word sense disambiguation

Experiment 2: Mikrokosmos semantic analyzer applied on a Spanish text not used in knowledge acquisition.

The above four texts were among about 400 Spanish texts used in the general lexicon and ontology acquisition process in Mikrokosmos. Table 2 shows the results on a previously unseen text. The results were essentially similar to those for the training texts in Table 1.

The unseen text used in the experiment contained 19 words missing from the Mikrokosmos lexicon. In such cases, the Mikrokosmos analyzer produces dummy entries, marked as nouns and semantically mapped to ALL, the root concept in the ontology (this has the effect of essentially not including any semantic constraints in the definition). No changes were made in the lexicon, ontology, or the programs. There were many syntactic binding problems with this text. We did not fix any of them. We could get even better results if we assumed perfect syntactic output and fixed all the binding problems. Unknown words (which were mapped to ALL) were treated as unambiguous. 12 of the 19 unknown words appeared to be proper names and only 3-4 of them were in fact ambiguous.

Experiment 3: Adding New Senses.

This experiment was designed to test the effect of acquiring additional word senses on disambiguation results. We added 40 new word senses to about 30 words in the lexicon. As a result, correctness of disambiguation dropped by only 3.6%.

In the rest of the paper, we describe the resources and algorithms used by Mikrokosmos for semantic analysis and, in particular, word sense disambiguation.

3 Static Resources: Ontology and Lexicon

Mikrokosmos uses two primary static resources: a language-specific lexicon and a language-independent ontology. The Spanish lexicon has 7,000 manually acquired entries that have been expanded to about 37,000 virtual entries using lexical rules (Onyshkevych and Nirenburg, 1995; Viegas et al, 1996; Viegas and Raskin, in preparation). Entries have many types of information including (a) syntactic patterns in which words occur, (b) semantic patterns that represent the meanings of the words, and (c) mappings between syntactic and semantic patterns that establish semantic relationships between the constituents of the syntactic patterns. The semantic patterns are built using the concepts and inter-concept relations in the ontology. The Mikrokosmos ontology is a broad-coverage classification of about 5000 concepts in the world, including nearly 3000 OBJECTS, 1200 EVENTS, and over 600 ATTRIBUTES and RELATIONS among OBJECTS and EVENTS (Mahesh, 1996; Mahesh and Nirenburg, 1995).[2] The ontology is a richly connected network of concepts with an average of. 16 attributes and relations per concept. Each relation links a concept to other concepts in the ontology and serves as a selectional constraint. Typical thematic roles such as AGENT and INSTRUMENT are included in the nearly 400 relations present in the ontology. Further information about the lexicon and the ontology as well as on-line access for browsing is available on the World Wide Web at the URL http://crl.nmsu.edu/Research/Projects/mikro/

4 Semantic Analysis and Disambiguation

Mikrokosmos uses the Panglyzer Spanish syntactic analyzer from Pangloss, an earlier MT project (Nirenburg, ed., 1994). Given the syntactic structures produced by Panglyzer, lexical entries from the Mikrokosmos lexicon are instantiated and syntactic variables denoting heads and arguments are bound to one another. Ontological concepts present in the semantic mappings in the lexical entries are instantiated from the ontology.

Selectional constraints are encoded in the ontology. Any language-specific relaxations of such constraints are noted in the semantic patterns in the lexicon. The Mikrokosmos analyzer gathers both ontological and lexical constraints for every pair of instantiated entries that have a

[2] Concepts in the ontology are shown in SMALL CAPITALS.

syntactic dependency between them. It checks each such constraint by searching in the ontology for a path that establishes how well the candidate filler meets the constraint.

Consider the example sentence “Fuentes financieras consultadas no preciso el monto” (“Financial sources consulted did not specify the amount”). The word “fuente” has three senses in the Mikrokosmos lexicon: MEDIA-SOURCE, FOUNTAIN, and PLATE. An ontological constraint on CONSULT (the meaning of “consultadas”) restricts its SOURCE to be HUMAN. Similar constraints exist between the meanings of “fuente” and “financieras”, between “fuente” and “preciso”, and between “preciso” and “monto”. In the following, we first illustrate how the constraints are checked and then how the results of checking individual constraints are combined in an efficient control structure to select the best combination of word senses for an entire sentence.

4.1 Checking Selectional Constraints: Onto-Search

A constraint is checked by finding a path in the ontology between the candidate (e.g., FOUNTAIN in the above example) and the constraint (HUMAN above) that estimates how well the candidate meets the constraint.

Relations in the ontology have two levels of selectional constraints: an overall constraint as well as an expected (default) filler that meets the constraint. The advantage that we have over previous constraint-based approaches to word-sense disambiguation is that we have a much richer set of constraints derived from the broad-coverage ontology, which allows fairly fine-grained constraints on some relations, along with a knowledge-intensive constraint checking method, as described briefly below.

In the easiest case, the selectional constraints on the correct set of senses are all satisfied, and are violated for incorrect combinations of senses. Satisfied selectional constraints appear in the method as a simple taxonomic path over the IS-A hierarchy between the candidate concept and the constraint. But because natural language use is not literal or precise (because of vagueness, metonymy, etc.), we often need to relax constraints; however, relaxing or discarding semantic constraints unrestrictedly would result in egregious proliferation of readings in semantic analysis.

In our method, controlled constraint satisfaction is managed by considering all relations, not just IS-A arcs, and by levying a cost for traversing any of those non-taxonomic relations. We treat the ontology as a directed (possibly cyclic) graph, with concepts as nodes and relations as arcs. Thus constraint satisfaction is treated as a cheapest path problem, between the candidate concept node and the constraint nodes; the best path thus reflects the most likely underlying semantic relation, whether it be metonymic or literal.

The cost assessed for traversing a metonymic (or other) arc may be dependent on the previous arcs traversed in a candidate path, because some arc types should not be repeatedly traversed, while other arcs should not be traversed if certain other arcs have already been seen. We use a state transition table to assess the appropriate cost for traversing an arc (based on the current path state) and to assign the next state for each candidate path being considered. Our weight assignment transition table has about 40 states, and has individual treatment for 40 types of arcs; the other arcs (of the nearly 400 total relation types) are treated by a default arc-cost mechanism.

The weights that are in the transition table are critical to the success of the method. We learn them by an automatic training method. After building a training set of inputs (candidate fillers and constraints) and desired outputs (the "correct" paths over the ontology, i.e., the preferred relation), we used a simulated annealing numerical optimization method (Kirkpatrick et al, 1983; Metropolis et al, 1953) for identifying the set of arc costs that results in the optimal

Sense of "fuente"	Is it a HUMAN?	Path in ontology
MEDIA-SOURCE	score=1.0	through IS-A arcs
FOUNTAIN	score=0.80	through PRODUCED-BY relation
PLATE	score=0.74	through PRODUCED-BY relation

Table 3. Constraint Checking Results for "is fuente a HUMAN?".

set of solutions for the training data. A similar approach is used to optimize the arc costs so that the cheapest cost reflects the preferred word sense from a set of candidates.

Although any of a variety of shortest-path graph search algorithms could be used, we use an A*-style modification of the Dijkstra algorithm with heap-based priority queues (Dijkstra, 1959; Gibbons, 1985). This algorithm gives us the desired expected-case complexity, with worst-case complexity of only $O(Elog_2N)$, where E is the number of edges and N is the number of nodes, and $E << N^2$.

Table 3 shows the results from checking constraints on the three senses of "fuente" in the above example. MEDIA-SOURCE IS-A HUMAN (through the intermediate concepts COMMUNICATION-ROLE and SOCIAL-ROLE in the hierarchy) whereas the other two are only related to HUMANS through a PRODUCED-BY relation.[3] If this was the only constraint in the sentence, the analyzer would immediately pick the MEDIA-SOURCE sense of “fuente” as the most appropriate one for “being consulted.” However, an average of 240 such constraints are checked per sentence in the texts reported in Table 1 and the results of all 240 constraints must be combined before an optimal selection can be made.

4.2 Efficient Control and Synthesis

The Mikrokosmos analyzer utilizes a new constraint-based control architecture called Hunter-Gatherer (Beale et al, 1996) to combine the results from constraint checking and pick the best combination of word senses for a sentence. Hunter-Gatherer (HG) optimizes search not by seeking the optimal configurations for constraint pruning, as previous systems have done, but by maximizing the pruning using its novel branch-and-bound methods. These branch-and-bound methods determine which variables of a problem cannot be improved by further processing and “freeze” their optimal values, reducing the overall complexity significantly. For example, in a “simple” problem with a million exhaustive combinations, a comparable constraint-based architecture (Tsang and Foster, 1990) required almost 19,000 combinations to be checked while HG required only 848 to guarantee an optimal solution. Furthermore, depending on the topology of the input problem, these results can be shown to be even more impressive. Problems in semantic analysis with exhaustive complexities over 10^{60} have been optimally solved by HG with less than a thousand combinations checked.

Table 4 shows actual results of analyses of various size problems. We have tested the Hunter-Gatherer algorithm extensively on a wide variety of sentences in several real-world texts and the claims of near-linear time processing and guaranteed optimal solutions have been verified.

It is interesting to note that a 20% increase in the total number of word senses for all the words in the sentence (79 to 95) results in a 626% increase (7.8M to 56M) in the number

[3] PLATE got a lower score than FOUNTAIN because the path for PLATE had to traverse many more IS-A links before finding the inherited relation PRODUCED-BY. There is a small penalty (score=0.96) for traversing an IS-A arc in conjunction with a metonymic relation.

Problem	#1	#2	#3
Number of word senses	79	95	119
Exhaustive combinations	7,864,320	56,687,040	235,092,492,288
Hunter-Gatherer combinations	179	254	327

Table 4. Near-Linear Processing for Natural Language Semantics.

of exhaustive combinations possible, but only a 42% increase (179 to 254) in the number of combinations considered by Hunter-Gatherer. As one moves on to even more complex problems, a 25% increase (95 to 119) in the number of word senses catapults the exhaustive complexity by 414,600% (56M to 235B) and yet only increases the Hunter-Gatherer complexity by 29% (254 to 327). As the problem size increases, the minor effects of "local multiplicative" influences diminish with respect to the size of the problem. We expect, therefore, the behavior of this algorithm to move even closer to linear with larger problems (e.g., discourse analysis). And, again, it is important to note that Hunter-Gatherer is guaranteed to produce the same results as an exhaustive search.

Although time measurements are often misleading, it is important to state the practical outcome of this type of control advancement. Prior to implementing Hunter-Gatherer, our analyzer failed to complete processing large sentences. The largest sentence above was analyzed for more than a day with no results. Using Hunter-Gatherer, on the other hand, the same sentence was finished in 17 **seconds.** It must be pointed out as well that this is not an artificially selected example. It is a real sentence occurring in one of the texts reported in Table 1, and not an overly large sentence at that.

4.3 Text Meaning Representation: The Output

Apart from disambiguating word senses, the Mikrokosmos analyzer builds a complete text meaning representation (TMR) for a text. Figure 1 shows the TMR for our example sentence.

```
CONSULT-001                               SPECIFY-001
  SOURCE: MEDIA-SOURCE-001                  AGENT: MEDIA-SOURCE-001
                                            THEME: AMOUNT-001
MEDIA-SOURCE-001                            polarity: negative
  SOURCE-OF: CONSULT-001
  AGENT-OF:   SPECIFY-001                 AMOUNT-001
  AREA-OF-ACTIVITY: FINANCE                 coreference: *unknown*
```

Figure 1. Text Meaning Representation for the Example Sentence "Fuentes financieras consultadas no preciso el monto."

5 Previous Work

World knowledge to be applied to disambiguate word senses is usually represented as selectional constraints either in a lexicon or some form of semantic network with binary links between nodes that denote word meanings. Previous methods for applying selectional constraints can be classified into the following three categories:

- Semantic networks: Constraints are represented by the links in the network with the assumption that closeness of nodes in the network signifies "conceptual closeness" or "semantic affinity." Methods such as marker passing (Charniak, 1983) and spreading activation (Waltz and Pollack, 1985) have been developed to search for optimal paths that connect different word meanings represented in such a semantic network. These methods work best on small-size networks with sparse connections among the nodes in the network.
- Lexical semantics: Selectional constraints are often encoded in individual lexicon entries along syntactic dependencies or sometimes through other forms of "licensing" or "expectations." Systems that take this approach include that of Wilks (1975) and other conceptual analyzers such as CA (Birnbaum and Selfridge, 1981) and Word Expert Parser (Small and Rieger, 1982). These methods work best on single-language systems where any world knowledge can be combined with lexical and linguistic knowledge and compiled into individual lexical entries or "word experts."
- Scripts: Various forms of pre-packaged contexts encoded in complex knowledge structures have been proposed to provide context-specific expectations to the analyzer (Cullingford, 1978; Schank and Abelson, 1977). In a sense, scripts are "world experts" instead of "word experts." Although such methods simplify word sense resolution and other problems remarkably, they demand detailed knowledge of possible scenarios. It is often prohibitively expensive to acquire such knowledge for general purpose, domain independent NLP.

None of these methods used a large scale ontology (because none was available). Nor did they show that they can resolve sense ambiguities in entire texts and at the same time produce complete meaning representations for the texts. We believe that Mikrokosmos is the first successful application of a knowledge-based method for large scale word sense disambiguation and text meaning representation.

Statistical methods (e.g., Yarowsky (1992)), work well on carefully chosen domains and training corpora. However, they are not as effective for processing texts from a wide variety of domains in general. Moreover, statistical methods are attractive for solving individual problems such as word sense disambiguation or part of speech tagging. They do not explain why certain meanings were chosen or how the chosen meanings together provide a meaning for a whole sentence or text, something that is often required to carry out further processing (e.g., to generate the meaning in a target language for machine translation).

6 Conclusions

Resolving word sense ambiguities is a central problem for machine translation. Previous attempts with knowledge-based methods have failed to show that the methods can be scaled up to perform well on real-world texts. In this paper, we have described how the Mikrokosmos analyzer applied knowledge from a large ontology and a large computational lexicon to select correct word senses 97% of the time for all the open-class words in real texts. In addition, the same analyzer also deals with several other semantic problems and constructs a text meaning representation for entire texts. We believe that this is a significant step in applying knowledge-based methods for machine translation.

Acknowledgements

Research reported in this paper was supported in part by Contract MDA904-92-C-5189 from the U.S. Department of Defense.

References

Beale, S., Nirenburg, S., and Mahesh, K. 1996. Hunter-Gatherer: Three Search Techniques Integrated for Natural Language Semantics. In *Proc. AAAI-96,* Portland, OR.

Beale, S., Nirenburg, S., and Mahesh, K. 1995. Semantic Analysis in the Mikrokosmos Machine Translation Project. In Proceedings of the 2nd Symposium on Natural Language Processing, 297-307. Bangkok, Thailand.

Birnbaum, L. and Selfridge, M. 1981. Conceptual analysis of natural language. In *Inside Computer Understanding,* ed. R. Schank and C. Riesbeck, p. 318-353. Lawrence Erlbaum Associates.

Charniak, E. 1983. Passing Markers: A Theory of Contextual Influence in Language Comprehension. *Cognitive Science* 7:171-190.

Cullingford, R. 1978. *Script application: Computer understanding of newspaper stories.* PhD thesis, Yale University, Department of Computer Science, New Haven, CT. Research Report #116.

Dijkstra, E. 1959. A Note on Two Problems in Connection with Graphs, in *Numerische Mathematik.* vol. 1, pp. 269-271.

Gibbons, A. 1985. *Algorithmic Graph Theory.* New York: Cambridge University Press.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B. M. 1996. The role of lexicons in natural language processing. *Communications of the ACM,* 39(1): 63-72.

Katz, J. J. and Fodor, J. A. 1963. The structure of semantic theory. *Language* 39, 170-210. Also in J. A. Fodor et al (eds.). *The Structure of Language: Readings in the Philosophy of Language.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. 1983. Optimization by Simulated Annealing. *Science,* vol. 220:671-680.

Mahesh, K. 1996. *Ontology Development: Ideology and Methodology.* Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University.

Mahesh, K. and Nirenburg, S. 1995. A situated ontology for practical NLP. In Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, August 1995.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. Equation of State Calculations by Fast Computing Machines, in *Journal of Chem. Physics,* vol 21:6, pp. 1087-1092.

Nirenburg, S., editor. 1994. The PANGLOSS Mark III Machine Translation System. A Joint Technical Report by NMSU CRL, USC ISI and CMU CMT, Jan. 1994.

Onyshkevych, B. and Nirenburg, S. 1995. A Lexicon for Knowledge-based MT. *Machine Translation,* 10: 1-2.

Schank, R. C. and Abelson, R. P. 1977. *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Lawrence Erlbaum.

Small, S. L. and Rieger, C. 1982. Parsing and comprehending with word experts. In *Strategies for natural language processing,* ed. W. G. Lehnert and M. H. Ringle. Lawrence Erlbaum.

Tsang, E. and Foster, N. 1990. Solution Synthesis in the Constraint Satisfaction Problem, Tech Report, CSM-142, Dept. of Computer Science, Univ. of Essex.

Viegas, E., Onyshkevych, B., Raskin, V. and Nirenburg, S. 1996. From *Submit* to *Submitted* via *Submission:* on Lexical Rules in Large-scale Lexicon Acquisition. In *Proceedings of the 84th Annual Conference of the Association for Computational Linguistics,* Santa Cruz, June 23-28, 1996.

Viegas, E. and Raskin, V. (in preparation). *Lexical Acquisition: Ideology and Methodology.* Memoranda in Computer and Cognitive Science. Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Waltz, D. L. and Pollack, J. B. 1985. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science,* 9:51-74.

Wilks, Y. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence* 6(l):53-74.

Wilks, Y., Slator, B., and Guthrie, L. 1995. *Electric Words: Dictionaries, Computers and Meanings.* Cambridge, MA: MIT Press.

Yarowsky, D. 1992. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proc. 14th International Conference on Computational Linguistics.* pp. 454-460, Nantes, France.