Combining Supervised and Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation

E. Agirre (eneko@si.ehu.es)
LSI saila. Euskal Herriko Unibertsitatea. Donostia, Basque Country.

G. Rigau (g.rigau@lsi.upc.es), L. Padró (padro@lsi.upc.es) and J. Atserias (batalla@lsi.upc.es)
Departament de LSI. Universitat Politècnica de Catalunya. Barcelona, Catalonia.

Abstract. This work combines a set of available techniques –which could be further extended– to perform noun sense disambiguation. We use several unsupervised techniques (Rigau et al., 1997) that draw knowledge from a variety of sources. In addition, we also apply a supervised technique, in order to show that supervised and unsupervised methods can be combined to obtain better results. This paper tries to prove that using an appropriate method to combine those heuristics we can disambiguate words in free running text with reasonable precision.

Keywords: word sense disambiguation, combining knowledge sources.

1. Introduction

The methods used by our sense disambiguating system are mainly unsupervised. Nevertheless, it may incorporate supervised knowledge when available. Although fully supervised systems have been proposed (Ng, 1997), it seems impractical to rely only on these techniques, given the high human labor cost they imply.

The techniques presented in this paper were tried on the Hector corpus in the framework of the SensEval competition. Since most of our techniques disambiguate using WordNet, we had to map WordNet synsets into Hector senses. Although the techniques can be applied to most parts of speech, for the time being we focused on nouns.

This paper is organized as follows: section 2 shows the methods we have applied. Section 3 deals with the lexical knowledge used and section 4 shows the results. Section 5 discusses previous work, and finally, section 6 present some conclusions.

2. Heuristics for Word Sense Disambiguation

The methods described in this paper are to be applied in a combined way. Each one must be seen as a container of part of the knowledge needed to perform a correct sense disambiguation. Each heuristic assigns a weight ranging in $[0, 1]$ to each candidate sense. These *votes* are later joined in a final decision.

Heuristic H_1 (Multi–words) is applied when the word is part of a multi–word term. In this case, the Hector sense corresponding to the multi–word term is assigned. Only H_1 and H_8 yield Hector senses.

Heuristic H_2 (Entry Sense Ordering) assumes that senses are ordered in an entry by frequency of usage. That is, the most used and important senses are placed in the entry before less frequent or less important ones. This heuristic assigns the maximum score to the first candidate sense and linearly decreasing scores to the others. The sense ordering used is that provided by WordNet.

Heuristic H_3 (Topic Domain) selects the WordNet synset belonging to the WN semantic file most frequent among the semantic files for all words in the context, in the style of Liddy & Paik (1992).

Heuristic H_4 (Word Matching) is based on the hypothesis that related concepts are expressed using the same content words, computing the amount of content words shared by the context and the glosses (Lesk, 1986).

Heuristic H_5 (Simple Co–occurrence) uses co–occurrence data collected from a whole dictionary. Thus, given a context and a set of candidate synsets, this method selects the target synset which returns the maximum sum of pairwise co–occurrence weights between a word in the context and a word in the synset. The co–occurrence weight between two words is computed as Association Score (Resnik, 1992).

Heuristic H_6 (Co–occurrence Vectors) is based on the work by Wilks et al. (1993), who also use co–occurrence data collected from a whole dictionary. Given a context and a set of candidate synsets, this method selects the candidate which yields the highest similarity with the context. This similarity can be measured by the dot product, the cosine function or the Euclidean distance between two vectors. The vector for a context or a synset is computed by adding the co–occurrence information vectors of the words it contains. The co–occurrence information vector for a word is collected from the whole dictionary using Association Score (see section 3).

Heuristic H_7 (Conceptual Density) (Agirre & Rigau, 1996; Agirre, 1998) provides a relatedness measure among words and word senses, taking as reference a structured hierarchical net. Conceptual Density captures the closeness of a set of concepts in the hierarchy, using the relation between the weighted amount of word senses and the size of the minimum subtree covering all word senses.

Given the target word and the nouns in the surrounding context, the algorithm chooses the sense of the target word which lies in the sub–hierarchy with highest Conceptual Density, i.e. the sub–hierarchy which contains a larger number of possible senses of context words in a proportionally smaller hierarchy.

Heuristic H_8 (Decision Lists). Given a training corpus where the target word has been tagged with the corresponding sense, frequencies are collected for: appearances of each word sense, bigrams of each word sense (form, lemma, and POS tag for left and right words), trigrams of each word sense, and window of surrounding lemmas. Frequencies are filtered, converted to association scores and organized in decreasing order as decision lists. In the test part, the features found in the context are used to select the word sense, going through the decision list until a matching feature is found (Yarowsky, 1994). As the training corpus is tagged with Hector senses, it also outputs Hector senses.

Combination. Finally, the ensemble of the heuristics is also taken into account. The way to combine all the heuristics in a single decision is simple. The weights assigned to the competing senses by each heuristic are normalized dividing them by the highest weight. The votes collected from each heuristic are added up for each competing sense.

3. Derived lexical knowledge resources

According to Wilks et al. (1993), two words co–occur in a dictionary if they appear in the same definition. In our case, a lexicon of 500,413 content word pairs of 41,955 different word forms was derived from *Collins English Dictionary*.

Table I. Words frequently co–occurring with *wine*.

word	AS	word	AS	word	AS
grapes	10.5267	bottle	8.1675	eucharist	7.1267
bottles	8.3157	Burgundy	7.2882	cider	6.9273
bread	8.2815	drink	7.2498	Bordeaux	6.6316

Table I shows the words co–occurring with *wine* with the highest Association Scores. The lexicon produced in this way from the dictionary is used by heuristics H_5 and H_6.

4. Results

Our system tries to disambiguate all nouns except those tagged as proper nouns. The results submitted to the SENSEVAL workshop are shown in table II (July columns). At that stage of development simple co–occurrence (H_5) and co–occurence vector (H_6) were not yet integrated. Small bugs were found and a revised version was re–submitted in October. Finally, we included the simple co–occurrence and co–occurrence vector techniques (November columns). The system is still evolving (see section 6).

Table II. Results obtained at each stage of development.

	Unsupervised (H_1 to H_7)			Supervised (H_1 to H_8)		
	July	October	November	July	October	November
recall	38.8%	38.8%	40.4%	60.7%	63.9%	66.9%
precision	41.6%	41.8%	43.5%	62.0%	65.3%	68.3%
coverage	93.0%	93.0%	93.0%	98.0%	98.0%	98.0%

Two combinations have been tried: an unsupervised system only using lexical knowledge, and a supervised system which includes also knowledge extracted from the training corpora.

Table III. Overall results for isolated heuristics.

	random	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8
recall	16.6%	5.7%	38.4%	30.1%	34.7%	27.6%	32.8%	29.5%	51.3%
precision	16.6%	84.4%	45.4%	35.6%	41.5%	32.7%	38.8%	37.3%	71.6%
coverage	100%	6.8%	84.5%	84.5%	84.5%	84.5%	84.5%	79.0%	71.6%

Table III shows the performance of each heuristic in isolation. Combining them all (table II) has the best recall in both the supervised and the unsupervised system. Our systems perform well in both the supervised and unsupervised categories of the SENSEVAL competition, specially considering that nearly all our techniques –except H_1 and H_8– disambiguate to WordNet senses. In order to yield Hector senses, we used a mapping provided by the SENSEVAL organization. The WordNet to Hector mapping adds a yet unevaluated handicap. Concerns were raised in the SENSEVAL workshop regarding the quality (gaps in either direction, arguable mappings, etc.) of this mapping. Also, the used POS tagger was very simple,

5. Comparison with previous work

Several approaches have been proposed for attaching the correct sense to a word in context. Some of them are only models for simple systems such as connectionist methods (Cottrell & Small, 1992) or bayesian networks (Eizirik et al., 1993), while others have been fully tested in real size texts, like statistical methods (Yarowsky, 1992; Yarowsky, 1994; Miller et al., 1994), knowledge based methods (Sussna, 1993; Agirre & Rigau, 1996), or mixed methods (Richarson, 1994; Resnik, 1995). The performance of WSD is reaching a high stance, although usually only small sets of words with clear sense distinctions are selected for disambiguation. For instance, Yarowsky (1995) reports a success rate of 96% disambiguating twelve words with two clear sense distinctions each, and Wilks et al. (1993) report a success rate of 45% disambiguating the word *bank* (thirteen senses from *Longman Dictionary of Contemporary English*) using a technique similar to heuristic H_6.

This paper has presented a general technique for WSD which is a combination of statistical and knowledge based methods, and which has been applied to disambiguate all nouns in a free running text.

6. Conclusions and Further Work

Our system disambiguates to WordNet 1.6 senses (the only exception being heuristics H_1 and H_8). In order to yield Hector senses, the results were automatically converted using a mapping provided by the SENSEVAL organization. It is clear that precision, to an as yet unquantified degree, is reduced if a sense mapping is used.

We have shown that the ensemble of heuristics is a useful way to combine knowledge from several lexical knowledge methods, outperforming each technique in isolation (coverage and/or precision). Better results can be expected from adding new heuristics with different methodologies and different knowledge sources (e.g. from corpora). More sophisticated methods to weight the contribution of each heuristic should also improve the results. Another possible improvement –after Wilks & Stevenson (1998)– would be to use a supervised learning process to establish the best policy for combining the heuristics.

In order to get a fair evaluation, we plan to test our system on a corpus tagged with WordNet senses, such as SemCor. We believe that an all–word task provides a more realistic setting for evaluation. If we want to get an idea of the performance that can be expected from a running system, we cannot depend on the availability of training data for all content words.

References

Agirre, E. and Rigau, G.: 1996, Word Sense Disambiguation using Conceptual Density, In *Proceedings of COLING'96*, pages 16-22. Copenhagen, Denmark.

Agirre, E.: 1998, *Formalization Of Concept-Relatedness Using Ontologies: Conceptual Density*, Ph.D. thesis, LSI saila, University of the Basque Country.

Cottrell, G. and Small, S.: 1992, A Connectionist Scheme for Modeling Word Sense Disambiguation, *Cognition and Brain Theory*, 6(1), pages 89-120.

Eizirik, L., Barbosa, V. and Mendes, S.: 1993, A Bayesian-Network Approach to Lexical Disambiguation, *Cognitive Science* 17, pages 257-283.

Lesk, M.: 1986 Automatic Sense Disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC'86 Conference, ACM.*

Liddy, E. and Paik, W.: 1992 Statistically–Guided Word Sense Disambiguation. In *AAAI Fall Symposium on Statistically Based NLP Techniques.*

Miller, G., Chodorow, M., Landes, S., Leacock, C. and Thomas, R.: 1994, Using a Semantic Concordance for sense Identification, In *Proceedings of ARPA Workshop on Human Language Technology.*

Ng, H.T.: 1997, Getting Serious about Word Sense Disambiguation, In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How?*, Washington DC, USA.

Resnik, P.: 1992, Wordnet and distributional analysis: A class–based approach to lexical discovery. In *AAAI Spring Symposium on Statistically Based NLP Techniques.*

Resnik, P.: 1995, Disambiguating Noun Groupings with Respect to WordNet Senses, In *Proceedings of the Third Workshop on Very Large Corpora*, MIT.

Richarson, R., Smeaton, A.F. and Murphy, J.: 1994, Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words, Working Paper CA-1294, School of Computer Applications, Dublin City University.

Rigau, G., Atserias, J. & Agirre, E.:1997, Combining Unsupervised Lexical Knowledge Methods for WSD, In *Proceedings of joint ACL-EACL'97.* Madrid, Spain.

Sussna, M.: 1993, Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, In *Proceedings of the Second International Conference on Information and knowledge Management.* Arlington, Virginia USA.

Wilks, Y., Fass, D., Guo, C., McDonal, J., Plate, T. and Slator, B.: 1993, Providing Machine Tractable Dictionary Tools, in Pustejowsky, J. (ed) *Semantics and the Lexicon*, pages 341-401. Kluwer Academic Publishers.

Wilks, Y., Stevenson, M.: 1998, Word Sense Disambiguation Using Optimized Combinations of Knowledge Sources, In *Proceedings of joint COLING-ACL'98.* Montreal, Canada.

Yarowsky, D.: 1992, Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, In *Proceedings of COLING'92*, pages 454-460. Nantes, France.

Yarowsky, D.: 1994, Decision Lists for Lexical Ambiguity Resolution, In *Proceedings of ACL'94*, Las Cruces, New Mexico.

Yarowsky, D.: 1995, Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, In *Proceedings of ACL'95*, Cambridge, Massachussets.