



Word Sense Disambiguation by Information Filtering and Extraction

JEREMY ELLMAN, IAN KLINCKE and JOHN TAIT

School of Computing & Information Systems, University of Sunderland, UK
(E-mail: Jeremy.Ellman@sunderland.ac.uk)

Abstract. We describe a simple approach to word sense disambiguation using information filtering and extraction. The method fully exploits and extends the information available in the Hector dictionary. The algorithm proceeds by the application of several filters to prune the candidate set of word senses returning the most frequent if more than one remains. The experimental methodology and its implication are also discussed.

Key words: word sense disambiguation, information filtering, SENSEVAL

1. Introduction

Our interest in word sense disambiguation comes from experiences with “Hesperus”, a research system that clusters Internet web pages based on their similarity to sample texts (Ellman and Tait, 1997). Hesperus uses a development of Morris and Hirst’s (1991) idea of lexical chains—that coherent texts are characterised by sets of words with meanings related to the topics and that these topics may be detected by reference to an external thesaurus.

Of course, many words are ambiguous and have meanings corresponding to different thesaural headwords. This represents a problem for lexical chaining, since the selection of an incorrect sense means that the word may be joined to an inappropriate chain. It will also be excluded from the correct chain, disrupting the apparent topic flow of the text, and degrading the accuracy of the procedure. This is counteracted using a word sense disambiguation pre-processor.

The function of the pre-processor is as much to filter out spurious sense assignments as to wholly provide unique sense identifications. This increases the accuracy of word sense disambiguation which is one of the effects of the lexical chaining process (Okumura and Honda, 1994) by early elimination of inappropriate senses.

The pre-processor follows the “sliding window” approach described in Sussna (1993), where ambiguous words are examined within the several words of their surrounding local context. This is compatible with the Senseval task (based as it is

on lexical samples), and was consequently re-implemented for Senseval, where it competed as SUSS.

SUSS's principal objective in Senseval was to evaluate different disambiguation techniques that could be used to improve the performance of a future version of Hesperus. This excludes both training and deep linguistic analysis. Training, as in machine learning approaches, implies the existence of training corpora. Such corpora tend only to exist in limited subject areas, or are restricted in scope. Machine Learning approaches were consequently excluded since Hesperus is intended to be applicable to any subject area. Indeed, we could argue that the associations found in thesauri contain the most common representations and subsume the associations found in normal text. Deep linguistic analysis is rarely robust, and often slow. This makes it incompatible with Hesperus, which is designed as a real-time system. A derived objective was to maximise the number of successful disambiguations – the essential competition requirement!

SUSS extensively exploited the Hector machine readable dictionary entries used in Senseval. There were two reasons for this: Firstly, Hector dictionary entries are extremely rich, and allowed us to consider disambiguation techniques that would not have been possible using Roget's Thesaurus alone (as used in Hesperus). Secondly, Hector sense definitions were much finer grained than those used in Roget. A system that used Roget would have been at a considerable disadvantage since it would not have been able to propose exact Hector senses in the competition.

One noteworthy technique made possible by the Hector dictionary was the conversion and adaptation of dictionary fields to patterns, as used in Information Extraction (e.g. Onyshkevych, 1993; Riloff, 1994). Where possible, this allowed the unique selection of a candidate word sense, with minimal impact on the performance of the rest of the algorithm.

2. SUSS: The Sunderland University Senseval System

SUSS is a multi-pass system that attempts to reduce the number of candidate word senses by repeated filtering. Following an initialisation phase, different filters are applied to select a preferred sense tag.

The order of filter application is important. Word and sense-specific techniques are applied first; more general techniques are used if these fail. Specific techniques are not likely to affect any other than their prospective targets, whereas general methods introduce probable misinterpretation over the entire corpus. For example, a collocate such as "brass band" uniquely identifies that sense of "band", with no impact on other word senses. Other techniques required careful assessment to ensure that their overall effect was positive. This was part of a structured development strategy.

2.1. THE SUSS DEVELOPMENT STRATEGY

We used the training data considerably to develop SUSS, not to train the system but to ensure that promising techniques for some types of ambiguity did not adversely influence the overall performance of the system. The strategy was as follows:

1. A basic system was implemented that processed the training data.
2. A statistics module was implemented that displayed disambiguation effectiveness by word, word sense, and percentage precision.
3. As different disambiguation techniques were developed, effectiveness was measured on the whole corpus.
4. Techniques that improved performance (as measured by total percentage successful disambiguations) were further developed. Those that degraded performance were dropped. (Since the competition was time limited, it was not cost effective to pursue interesting but unsuccessful approaches.)

2.2. SUSS INITIALISATION PHASE

SUSS used a preparation phase that included dictionary processing and other preparations that would otherwise be repeated for each lexical sample to be processed. The HECTOR dictionary was loaded into memory using a public domain program that parses SGML instances. This made the definition available as an array of homographs that is further divided into an array of finer sense distinctions. Each of these contained fields, such as the word sense definition, part of speech information, plus examples of usage.

The usage examples were used in the “Example Comparison Filter” and the “Semantic Relations Filter” techniques (described below). They were reduced to narrow windows W words wide centred on the word to be disambiguated from which stopwords (Salton and McGill, 1983) have been eliminated. This facilitated comparison with identically structured text windows produced from the test data. The main SUSS algorithm is as follows.

2.3. SUSS ALGORITHM PROCESSING PHASE

1. For each *sample*:
2. Filter possible entries as collocates.
(**DONE** if there is only one candidate sense.)
3. Filter remaining senses for information extraction pattern.
(**DONE** if there is only one candidate sense.)
4. Filter remaining senses for idiomatic phrases.
(**DONE** if there is only one candidate sense.)
5. Eliminate stopwords from sample.

6. Produce window w words wide centred on word to be disambiguated
 7. For each *example in the Hector dictionary entry*
 - Match the sample window against the example window
 - Select the sense that has the highest example matching score.
 8. If no unique match found, return the most frequent sense of those remaining from the training corpus (or first remaining dictionary entry – note 1).
- We now go on to describe the specific techniques tested.

2.4. COLLOCATION FILTER

Collocations are short, set expressions which have undergone a process of lexicalisation. For example, consider the collocation ‘brass band’. This expression, without context, is understood to refer to a collection of musicians, playing together on a range of brass instruments, rather than a band made of brass to be worn on the wrist. For these reasons it is possible for the Hector dictionary to define such expressions as distinct senses of the word.

Given the set nature of collocations, therefore, it was considered that to look for these senses early in the disambiguation process would be a simple method of identifying or eliminating them from consideration.

The collocation identification module, therefore, worked as a filter using simple string matching. If a word occurrence passing through the module corresponded to one of the collocational senses defined in the dictionary it would be tagged as having that sense. If none of these senses were applicable, however, all senses taking a collocational form were filtered out.

2.5. INFORMATION EXTRACTION PATTERN FILTER

The Information Extraction filter refers exclusively to enhancements to the Hector dictionary entries specifically to support word sense disambiguation. The HECTOR dictionary is primarily intended for human readers. Many entries contain a *clues* field in a restricted language that indicates typical usage. Examples include phrases such as “*learn at mother’s knee, learn at father’s knee, and variants*”, or “*usu on or after*”. Such phrases have long been proposed as an important element of language understanding (Becker, 1975). These phrases were manually converted into string matching patterns and successfully used to identify individual senses. For example, “shake’ contains the following:

```
<idi>shake in one’s shoes, shake in one’s boots</idi>
<clues>v/= prep/in pron-poss prep-obj/(shoes,boots,seat)</clues>
```

This can be used to convert the idiom field (using PERL patterns) as follows:

```
<idi>shake in \w* (shoes/boots/seat)</idi>
```

This may now be used to match against any of the idiomatic expressions “shake in her boots”, “your boots”, etc.

We call a related method “phrasal patterns”. A phrasal pattern is a non-idiomatic multiple word expression that strongly indicates use of a word in a particular sense. For example, “shaken up” seems to occur only in past passive forms. Adding appropriate phrasal patterns to a dictionary sense was found to increase disambiguation performance for that sense. The majority of phrasal patterns were manually derived from the Hector dictionary entries. Others were identified by observing usage patterns in the dictionary examples, or the training data.

Collocation and other phrasal methods are important since they are tightly focused on one word, and on one sense that word may be used in. They do not affect other word senses, and can not influence the interpretation of other words.

2.6. IDIOMATIC FILTER

Idiomatic forms identify some word senses. Unlike collocations, however, idiomatic expressions are not constant in their precise wording. This made it necessary to search for content words in a given order, rather than looking for a fixed string. An idiom was considered present in the text if a subset of the content words were found exceeding a certain (heuristically determined) threshold value. For example, the meaning of “*too many cooks*” is clear, without giving the precise idiom.

Dictionary entries that contained idiomatic forms were processed as follows. Firstly, two word idioms were checked for specifically. If the idiom was longer, stopwords were removed from the idiomatic form listed, and remaining content words compared in order with words occurring in the text. If 60% of the content words were found in the region of the target word, the idiomatic filter succeeded, and senses containing that idiom selected. Otherwise, senses containing that idiomatic form were excluded from further consideration.

2.7. EXAMPLE COMPARISON FILTER

The Example Comparison Filter tries to match the examples given in the dictionary against the word to be disambiguated, looking at the local usage context. It assigns a score for each sense based on identical words occurring in the text and dictionary examples and their relative positions. We take a window of words surrounding the target word, with a specified width and specified position of the target, in the text and in a similar window from each dictionary example.

For each example in each sense, all the words occurring in each window are compared and, where identical words are found, a score, S , is assigned, where

$$S = \sum_{w \in W} d_S d_E$$

and w is a word in window W , and d_S and d_E are functions of the distance of the word from the target word in the sample and example windows respectively, such that greater distances result in lower scores. The size of the window was determined

empirically. Window sizes of 24, 14, and 10 words were tried. Larger window sizes increased the probability of spurious associations, and a window size of ten words, (which is five words before and five words after the target word) was selected as optimal.

When all the example scores have been calculated for each word sense, the sense with the highest example score is chosen as the correct sense of that occurrence.

In cases where this does not produce a result, the most frequently occurring sense (or first dictionary sense – see note 1) that has not been previously eliminated is chosen.

2.8. OTHER TECHNIQUES EVALUATED

One of the objectives of SUSS was to evaluate different disambiguation techniques. Below we describe two methods that were evaluated, but not used in the final system, since they lead to decreased overall performance.

2.9. PART OF SPEECH FILTER

Wilks and Stevenson (1996) have claimed that much of sense tagging may be reduced to part-of-speech tagging. Consequently, we used the Brill (1992) Tagger on the subset of the training data set that required part-of-speech discrimination. This should have improved disambiguation performance by filtering out possible senses not appropriate to the assigned part of speech. However, due to tagging inaccuracy, this was just as likely to eliminate the correct word sense, too. Consequently, it did not make a positive contribution.

Another routine that used the part-of-speech tags attempted to filter out the senses of words marked as noun modifiers by the dictionary grammar labels where the following word was not marked as a noun by the tagger. This routine also checked words that contained an ‘after’ specification in the grammar tag and eliminated these senses where the occurrence did not follow the word given. However, it gave no overall benefit to the results either. One possible cause of this is in occurrences where there are two modifiers joined by a conjunction so that the first is, legitimately, not followed immediately by a noun.

2.10. SEMANTIC RELATIONS FILTER

The Semantic Relations Filter is an extension of the example comparison filter that uses overlapping categories and groups in Roget’s thesaurus, rather than identical word matching. This should allow us to recognise that “accident” is used in the same sense in “car accident” and “motor-bike accident”, since both are means of transport.

Appropriate scores are allocated for each category in Roget that the test sentence window has in common with the dictionary example window. As in the example

comparison, the sense that contains the highest scoring example is selected as the best.

Disappointingly, this technique finds many spurious relations where words in the local context are interpreted ambiguously. This led to an overall performance degradation over the test set, and so the technique was not part of the final SUSS algorithm.

3. Discussion and Conclusion

SUSS consisted of a number of individual disambiguation techniques that were applied to the data sequentially. Each of these techniques were designed to have one of two effects; either to attempt to assign a unique dictionary sense for the occurrence, or to eliminate one or more invalid senses from consideration.

During development a range of techniques were tested to determine whether they were effective in increasing the disambiguation accuracy of the algorithm. Details of the different techniques applied, and their relative effectiveness are given in Ellman (forthcoming). The testing procedures utilised the training data, with the algorithm being applied both with and without the technique activated. The results of these applications were compared over different senses and different words, against the human tagged training data.

The statistics produced were used to determine whether the technique improved the overall accuracy of the disambiguation and, hence, whether it was a useful technique. Some techniques, for example, produced a great improvement in accuracy on particular words or specific senses,² yet the overall effect was a reduction in accuracy over all words.

This result reflects the interaction between word specific and generic sense disambiguation methods. A generic disambiguation technique needs to have better accuracy than that which would be given by selecting a default sense. For example, in the training corpus, “wooden” means “made of wood” in 95% of the samples. Thus, a generic technique applied to “wooden” needs to exceed this level of accuracy or it will degrade overall performance.

Regular Information Extraction patterns provided a particularly effective sense specific disambiguation. However, it was necessary to convert each pattern by hand. A clear next step would be the development of a module to automatically produce the patterns from the relevant dictionary fields.

SUSS performed surprisingly well considering its lack of sophistication, with above average performance compared to other systems. It is particularly interesting to note that it was placed in the first three systems where no training data was supplied.

Notes

¹ The calculation of sense occurrence statistics was designed to counter a perceived deficiency in Hector, where the ordering of senses did not appear to match that of sense frequency in the corpus. This was considered to be a training technique, and SUSS was classified as a learning system. The SUSS-Dictionary system did not use this technique and was considered as an “all words” system.

² A future variation of SUSS could use different settings for each word to be disambiguated. These could be determined automatically using machine learning algorithms.

References

- Brill, E. “A Simple Rule-based Part-of-speech Tagger”. *Proceeding of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992.
- Becker, J. D. “The Phrasal Lexicon”. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing*, Cambridge, MA, 1975, pp. 70–77.
- Ellman, J. (forthcoming). “Using Roget’s Thesaurus to determine the similarity of Texts”. PhD Thesis University of Sunderland, UK.
- Ellman, J. and J. Tait. “Using Information Density to Navigate the Web”. UK IEE Colloquium on Intelligent World Wide Web Agents, March 1997.
- Okumura, M. and T. Honda. “Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion”. *Proc. COLING 1994* vol. 2, pp. 755–761.
- Onyshkevych, B. “Template Design for Information Extraction”. *Proceeding of the Fifth Message Understanding Conference (MUC-5)*, 1993.
- Morris, J. and G. Hirst. “Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text”. *Computational Linguistics*, 17(1) (1991), 21–48.
- Riloff, E. and W. Lehnert. “Information Extraction as a Basis for High-precision Text Classification”. *ACM Transactions on Information Systems*, 12(3) (July 1994), 296–333.
- Salton, G. and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill.
- Stairmand, M. “A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval”. PhD Thesis. Dept of Computational Linguistics UMIST UK, 1996.
- St-Onge, D. *Detecting and Correcting Malapropisms with Lexical Chains*. MSc Thesis, University of Toronto, 1995.
- Sussna, M. “Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network”. *Proceedings of the Second International Conference on Information and Knowledge Base Management*, 1993, pp. 67–74.
- Wilks, Y. and M. Stevenson The Grammar of Sense: Is Word-sense Tagging Much More Than Part-of-speech Tagging? Technical Report CS-96-05, University of Sheffield, 1996.