Computers and the Humanities **34:** 135–140, 2000.

Large Scale WSD Using Learning Applied to SENSEVAL

PAUL HAWKINS and DAVID NETTLETON
University of Durham, Southampton, UK

Abstract. A word sense disambiguation system which is going to be used as part of a NLP system needs to be large scale, able to be optimised towards a specific task and above all accurate. This paper describes the knowledge sources used in a disambiguation system able to achieve all three of these criteria. It is a hybrid system combining sub-symbolic, stochastic and rule-based learning. The paper reports the results achieved in Senseval and analyses them to show the system's strengths and weaknesses relative to other similar systems.

1. Introduction

The motivation behind this work is to develop a core Word Sense Disambiguation (WSD) module which can be integrated into a NLP system. An NLP system imposes three requirements on any dedicated WSD module it may use:

- To be large scale and disambiguate all words contained in all open class categories.
- To be able to be optimised towards a specific task.
- To be accurate.

Senseval facilitated the evaluation of all three of these requirements. Senseval enabled the comparison of disambiguation accuracy with other state-of-the-art systems. It also provided the first opportunity to test if this system was lexicon independent which enables optimisations towards a specific task.

The main features of this system are the way different knowledge sources are combined, how contextual information is learnt from a corpus and how the disambiguation algorithm eliminates senses. This paper concentrates on the knowledge sources used. A detailed examination of all components of the system can be found in (Hawkins, 1999).

2. Knowledge Sources

Three knowledge sources are used to aid disambiguation: frequency, clue words and contextual information. They are all combined together to produce a hybrid system which takes advantage of stochastic, rule-based and sub-symbolic learning methods. A hybrid system seems appropriate for the WSD task because words differ considerably in the number of different senses, the frequency distribution of those senses, the number of training examples available and the number of collocates which can help disambiguation. This makes the task very different for each word, and affects the amount each of the knowledge sources is able to help disambiguation for that particular word. By combining these knowledge sources the aim is to take the useful information each is able to offer, and not allow them to cause confusion in cases where they are unable to help. Each of the three knowledge sources is now described.

2.1. FREQUENCY

The frequency information is calculated from the Hector training corpus which has been manually sense tagged. The frequency of each sense is calculated for each *word form* rather than the *root form* of each word. In some instances this morphological information greatly increases the frequency baseline.[1] For example, the frequency distribution of senses is very different for word forms *sack* and *sacks* than it is for *sacking*. The results show that using frequency information in this way increases the frequency baseline for *sack* from 50% to 86.6%.

2.2. CLUE WORDS

Clue words are collocates or other words which can appear anywhere in the sentence. The clue words are manually identified, which does pose a scaleability problem. However, given the size of the Senseval task it seemed appropriate to take advantage of human knowledge. On average less than one hour was dedicated by an unskilled lexicographer to identifying clues for each word. This is substantially less than the skilled human effort required to manually sense tag the training data. The success of this knowledge source on this scale may influence the decision to invest resources in clue words on a larger scale.

In general, clues give very reliable information and therefore they can often be used even with words which have a very high frequency baseline. If an infrequent sense has a good clue then it provides strong enough evidence to out-weigh the frequency information. For the ambiguous word *wooden*, *spoon* provides an excellent clue for an infrequently used sense. This enabled the system to achieve 98% accuracy – 4% above the frequency baseline. The learning algorithm was unable to help for this word as it does not suggest senses with a high enough confidence to ever out-weigh the frequency information.

2.3. CONTEXTUAL INFORMATION

This section introduces the notion of a contextual score which represents a measure for the contextual information between two concepts. Whilst it contributes less to the overall accuracy than the frequency or clue words information, contextual information aims to correctly disambiguate the more difficult words. It uses a sub-symbolic learning mechanism and requires training data. As with most sub-symbolic approaches it is difficult to obtain an explanation for why a particular sense is chosen.

The contextual score uses the WordNet hierarchy to make generalisations so that the most is gained from each piece of training data. These scores differ from a semantic similarity score described in Sussna (1993), by representing the likelihood of two concepts appearing in the same sentence rather than a measure of how closely related two concepts are. As WordNet does not attempt to capture contextual similarity which is required for WSD (Karov and Edelman, 1996) this information is learnt. This greatly reduces the dependency on the WordNet hierarchy making the system more domain independent. For example, in WordNet *doctor* and *hospital* would be assigned a very low semantic similarity as one is a type of professional and the other is a type of building. However, the concepts do provide very useful contextual information which would be learnt during training.

Contextual scores are learnt by increasing scores between the correct sense and the contextual words and decreasing scores between the incorrectly chosen sense and the contextual words. The mechanism by which this is performed is beyond the scope of this paper.

The contextual scores between concepts are stored in a large matrix. Only the nodes and their hypernyms which have occurred more than 20 times in the SemCor training data are included in the matrix which comprises about 2000 nodes. Whilst it would be possible to include all WordNet nodes in the matrix, the amount of training data required to train such a matrix is currently not available. Also the purpose of the matrix is to learn scores between more general concepts in the higher parts of the hierarchy and to accept the WordNet structure in the lower parts. To find the contextual score between two nodes they are looked up to see if they are contained in the matrix; if they are not their hypernyms are moved up until a node is found which is in the matrix.

The contextual scores between nodes in the matrix are learnt during training. Given a training sentence such as *"I hit the board with my hammer"*, where *board* is manually sense tagged to the *Board(plank)* sense, *Hit* and *Hammer* are contextual words, but only *Hammer* will be considered in this example. Figure 1 shows how scores are changed between nodes. Let us assume that the system incorrectly assigns the *Circuit Board* sense to *board*. *Hammer* is represented by *Device* in the contextual matrix, the correct sense of *board* is represented by *Building Material* and the incorrectly chosen sense is represented by *Electrical Device*. The training process increases the contextual score between *Device* and *Building Material* and

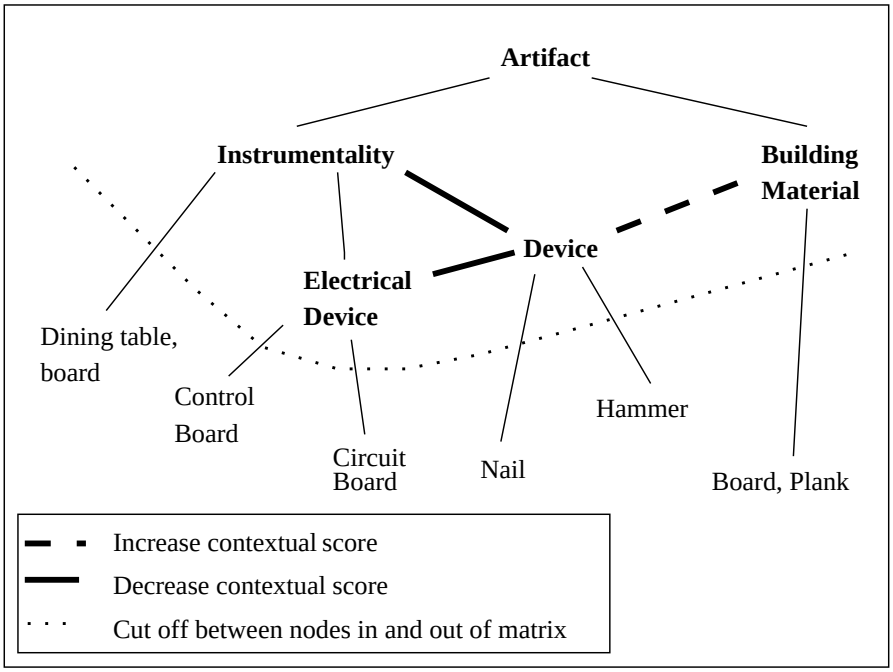

Figure 1. Diagram showing the changes in contextual scores if 'hammer' and the 'board, plank' sense of board appear in a training sentence.

decreases the score between *Electrical Device* and *Device*, Thus making *hammer* a better contextual clue for *Board (plank)* and a worse contextual clue for *Circuit board*. The diagram highlights the benefit of the contextual matrix operating above the word level. The training sentence also enables *Nail* to obtain a higher contextual score with *Board(plank)*.

The greatest benefit of the contextual score has proved to be for words which are difficult to disambiguate. Typically these words have a low frequency baseline and clue words are unable to improve accuracy.

Contextual scores can be learnt for concepts with different POS. This vastly increases the amount of contextual information available for each ambiguous word and also enables all words of all POS to be disambiguated. This is important in order to meet the large-scale requirement imposed on the system.

As contextual scores are learnt there is a reliance on training data. However, as the system is not dependant on the WordNet hierarchy, a system trained on SemCor should be able to be used on a different lexicon without re-learning. Using the Hector lexicon during Senseval was the first opportunity to test this feature. Analysis of the results in section 3 shows that the learning aspects of the system do exhibit lexicon independent features.

Table I. The effect of each knowledge source on overall accuracy

	Onion	Generous	Shake	All words
(1) Root Form Frequency	84.6	39.6	23.9	57.3
(2) Word Form Frequency	85	37	30.6	61.6
(3) Clue words + 2	92.5	44.9	71.1	73.7
(4) Contextual scores + 2	85	50.1	61.8	69.8
(5) Full System 2 + 3 + 4	92.5	50.7	69.9	**77.1**
(6) Coarse Grained 2 + 3 + 4	92.5	50.7	72.5	**81.4**

3. Results

Table I shows the contribution frequency, clue words and contextual scores have made to the overall accuracy of the system. Apart from the final row all scores quoted are 'fine-grained' results. Precision and recall values are the same as this system attempted every sentence.

Row (2) shows that the overall accuracy is increased by 4.3% by using word form rather than root form frequencies. Row (4) shows that this system performs quite well even without the use of manually identified clue words; such a system would have no scaleability problems. Out of the three words identified, *generous* benefits the most from the contextual scores. This is because it has a low frequency baseline and there are very few clues words which are able to help. Row (5) shows that the overall system achieves much higher accuracy than any sub-section of it. This shows that the clue words and contextual scores are useful for disambiguating different types of words and so can be successfully combined.

4. Conclusion and Comparison

The real benefits of the Senseval evaluation are now briefly exploited by comparing different systems' results.

Figure 2 uses Kappa to analyse results of the four systems which achieved the highest overall precision, all of which used supervised learning. Kappa gives a measure of how well the system performed relative to the frequency baseline. This enables the relative difficulty of disambiguating different categories of words to be examined.

The graph shows that all systems found that nouns were the easiest POS to disambiguate and adjectives proved slightly more difficult than verbs. Relative to other systems Durham did well for nouns and least well for verbs. Possible reasons for this are that the Durham system only uses semantic information in the context, and gives equal weight to all words in the sentence. Other systems also use syntactic clues and often concentrate on the words immediately surrounding

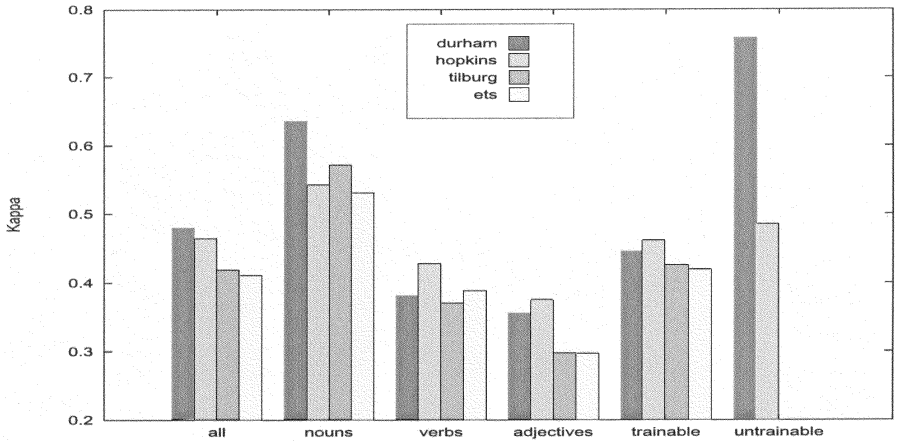

Figure 2. Graph showing comparison between 4 learning systems in Senseval.

the ambiguous word which may be more beneficial for discriminating between verb senses.

The Durham system performed very well on the words where no training data was given. This highlights its lexicon independence feature, as it was able to take advantage of training performed using SemCor and the WordNet lexicon.

Note

[1] The accuracy achieved by a system which always chooses the most frequent sense.

References

Hawkins, P. "DURHAM: A Word Sense Disambiguation System". Ph.D. thesis, Durham University, 1999.

Karov, Y. and S. Edelman. "Similarity-based Word Sense Disambiguation". *Computational Linguistics*, 24(1) (1996), 41–59.

Sussna, M. "Word Sense Disambiguation for Free-Text Indexing Using a Massive Semantic Network". In *Proceedings of the 2nd International Conference on Information and Knowledge Management*, pp. 67–74.