From Predicting Predominant Senses to Local Context for Word Sense Disambiguation

Rob Koeling & Diana McCarthy
Department of Informatics,
University of Sussex
Brighton BN1 9QH, UK
{*robk,dianam*}*@sussex.ac.uk*

Abstract

Recent work on automatically predicting the predominant sense of a word has proven to be promising (McCarthy et al., 2004). It can be applied (as a first sense heuristic) to Word Sense Disambiguation (WSD) tasks, without needing expensive hand-annotated data sets. Due to the big skew in the sense distribution of many words (Yarowsky and Florian, 2002), the First Sense heuristic for WSD is often hard to beat. However, the local context of an ambiguous word can give important clues to which of its senses was intended. The sense ranking method proposed by (McCarthy et al., 2004) uses a distributional similarity thesaurus. The *k* nearest neighbours in the thesaurus are used to establish the predominant sense of a word. In this paper we report on a first investigation on how to use the grammatical relations the target word is involved with, in order to select a subset of the neighbours from the automatically created thesaurus, to take the local context into account. This unsupervised method is quantitatively evaluated on SemCor. We found a slight improvement in precision over using the predicted first sense. Finally, we discuss strengths and weaknesses of the method and suggest ways to improve the results in the future.

1 Introduction

In recent years, a lot of research was done on establishing the predominant sense of ambiguous words automatically using untagged texts (McCarthy et al., 2004; McCarthy et al., 2007). The motivation for that work is twofold: on the one hand it builds on the strength of the *first sense heuristic* in Word Sense Disambiguation (WSD) (i.e. the heuristic of choosing the most commonly used sense of a word, irrespective of the context in which the word occurs) and on the other hand it recognizes that manually created resources for establishing word sense distributions are expensive to create and therefore hard to find. The one resource that is used most widely, SemCor (Miller et al., 1993), is only available for English and only representative for 'general' (non domain specific) text. McCarthy et al's method was successfully applied to a corpus of modern English text (the BNC (Leech, 1992)) and the predicted predominant senses compared well with the gold standard given by SemCor. Other experiments showed that the method can successfully be adapted to domain specific text (Koeling et al., 2005) and other languages (for example, Japanese (Iida et al., 2008)).

Even though the first sense heuristic is powerful, it would be preferable to only use it for WSD, when either the sense distribution is so skewed that the most commonly used sense is by far the most dominant, or as a back-off when few other clues are available to decide otherwise. The use of local context is ultimately necessary to find evidence for the intended sense of an ambiguous word. In this paper we investigate how we can exploit results from intermediate steps taken when calculating the predominant senses to this end.

The work on automatically finding predominant senses[1] was partly inspired by the observation that you can identify word senses by looking at the nearest neighbours of a target word in a distributional thesaurus. For example, consider the following (simplified) entry for the word *plant* in such a thesaurus (omitting the scores for distributional similarity):

plant : factory, industry, facility, business, company, species, tree, crop, engine, flower, farm, leaf, market, garden, field, seed, shrub... (1)

Just by looking at the neighbours you can identify two main groups of neighbours, each pointing at separate senses of the word. First there is the set of words consisting of *factory, industry, facility, business, company, engine* that hint at the 'industrial plant' sense of the word and then there is the set consisting of *tree, crop, flower, leaf, species, garden, field, seed, shrub* that are more closely related to the 'flora' sense of the word. A few words, like *farm* and possibly *market*

[1] (McCarthy et al., 2004) concentrates on evaluating the predominant sense, but the method does in fact rank all the senses in order of frequency of use.

could be associated equally strongly with either sense. The idea behind 'sense ranking' is, that the right mix of

1. number of neighbours with a strong associations with one or more of the senses,
2. the strength of the association (semantic similarity) between neighbour and sense and
3. the strength of the distributional similarity of the contributing neighbour and the target word, will allow us to estimate the relative importance (i.e. frequency of use) of each sense.

What we want to explore here, is how we can use the local context of an occurrence of the target word, to select a subset of these neighbours. This subset should consist of words that are related more strongly to the sense of the word in the target sentence. For example, consider the word *plant* in a sentence like:

> *'The gardener grows plants from vegetable seeds.'* (2)

Plant is used in this sentence as the 'subject of grow'. A simple way of zooming in on potentially relevant neighbours is by using the most informative contexts shared between neighbours and the word in the target sentence. This is implemented by selecting just those words that occur in *the same grammatical context* (i.e. as subject of the verb 'grow') in a reference corpus[2]. If we apply that to the example in 1, we end up with the following subset: *business, industry, species, tree, crop, flower, seed, shrub*. Even though the first two words are still associated with the 'industrial plant' sense, we can see that the majority of the words in this subset is strongly associated with the intended sense.

In the next section we first give a quick introduction to the sense ranking algorithm introduced in (McCarthy et al., 2004). Then we explain how we can use the database of grammatical relations that we used for creating the thesaurus, for selecting a subset of neighbours in the thesaurus. The following section describes an evaluation performed on the SemCor data. In the last two sections we discuss the results and especially why both recall and precision are lower than we had hoped and what can be done to improve the results.

2 Predominant Senses and Local Context

For a full review of McCarthy et al's ranking method, we refer to (McCarthy et al., 2004) or (McCarthy et al., 2007). Here we give a short description of the method. Since we need the grammatical relations used for building the thesaurus, for selecting a subset of the neighbours, we explain the procedure for building the thesaurus in 2.2. In the last part of this section we explain how we exploit local context for SD.

2.1 Finding Predominant Senses

We use the method described in McCarthy et al. (2004) for finding predominant senses from raw text. It can be applied to all parts of speech, but the experiments in this paper all focus on nouns only. The method uses a thesaurus obtained from the text by parsing, extracting grammatical relations and then listing each word (w) with its top k nearest neighbours, where k is a constant. Like McCarthy et al. (2004) we use $k = 50$ and obtain our thesaurus using the distributional similarity metric described by Lin (1998). We use WordNet (WN) as our sense inventory. The senses of a word w are each assigned a ranking score which sums over the distributional similarity scores of the neighbours and weights each neighbour's score by a WN Similarity score (Patwardhan and Pedersen, 2003) between the sense of w and the sense of the neighbour that maximises the WN Similarity score. This weight is normalised by the sum of such WN similarity scores between all senses of w and and the senses of the neighbour that maximises this score. We use the WN Similarity **jcn** score on nouns (Jiang and Conrath, 1997) since this gave reasonable results for McCarthy et al. and it is efficient at run time given precompilation of frequency information. The **jcn** measure needs word frequency information, which we obtained from the British National Corpus (BNC) (Leech, 1992). The distributional thesaurus was constructed using subject, direct object adjective modifier and noun modifier relations.

Thus we rank each sense $ws_i \in WS_w$ using:
Prevalence Score $ws_i =$

$$\sum_{n_j \in N_w} dss_{n_j} \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in WS_w} wnss(ws_{i'}, n_j)} \quad (3)$$

where the WordNet similarity score (*wnss*) is defined as:

$$wnss(ws_i, n_j) = \max_{ns_x \in NS_{n_j}} (wnss(ws_i, ns_x))$$

2.2 Building the Thesaurus

The thesaurus was acquired using the method described by Lin (1998). For input we used grammatical relation data extracted using an automatic

[2] We use the same corpus used for generating the thesaurus as for the reference corpus (in all our experiments).

parser (Briscoe and Carroll, 2002). For the experiments in this paper we used the 90 million words of written English from the BNC. For each noun we considered the co-occurring verbs in the direct object and subject relation, the modifying nouns in noun-noun relations and the modifying adjectives in adjective-noun relations. This limited set of grammatical relations was chosen since accuracy of the parser is particularly high for these 4 relations. We could easily extend the set of relations to more in the future. A noun, w, is thus described by a set of co-occurrence triples $< w, r, x >$ and associated frequencies, where r is a grammatical relation and x is a possible co-occurrence with w in that relation. For every pair of nouns, where each noun had a total frequency in the triple data of 10 or more, we computed their distributional similarity using the measure given by Lin (1998). If $T(w)$ is the set of co-occurrence types (r, x) such that $I(w, r, x)$ is positive then the similarity between two nouns, w and n, can be computed as:

$$\frac{\sum_{(r,x)\in T(w)\cap T(n)} (I(w,r,x) + I(n,r,x))}{\sum_{(r,x)\in T(w)} I(w,r,x) + \sum_{(r,x)\in T(n)} I(n,r,x)} \quad (4)$$

where:

$$I(w,r,x) = \log \frac{P(x|w \cap r)}{P(x|r)}$$

A thesaurus entry of size k for a target noun w can then be defined as the k most similar nouns to noun w.

2.3 Local Context

The basis for building the distributional similarity thesaurus, is the set of grammatical relations that the target word shares with other words. For example, if we look at the thesaurus entry for the noun *bike*, then we see that the closest neighbours are (the synonym) *bicycle* and the closely related *motorbike* (and *motorcycle*). The next 10 closest neighbours are all other vehicles (*car*, *van*, *boat*, *bus*, etc.). This is something we would expect to see, since all these words do occur in similar grammatical contexts. We *travel* by *bike*, as well as by *motorcycle*, *car* and *bus*. We *park* them, *drive_off* with them, *hire* them, *abandon* them and *repair* them. Many of these relations can be applied to a wide range of vehicles (or even a wider range of objects). However, some relations are more specific to two-wheeled vehicles. For example, it is quite common to *mount* a bike or a motorbike, whereas it is less common to *mount* a *car* or a *van*. *(Motor)bikes* are *chained* to stop people from stealing them and it is probably more common to *ride* a *(motor)bike* as opposed to *driving* a *car* or *truck*. Of course there are many other more general things you can do with these vehicles: *buy*, *sell*, *steal* them; there are *yellow bikes*, *cars* and *boats*, just like other objects. Therefore, we can see many other types of objects lower in the list of neighbours that share these more general grammatical relations, but not those that are specific to, say, vehicles or even the sub-category of two-wheeled vehicles.

Consider the following sentence containing the ambiguous noun *body*:

'Regular exercise keeps the body healthy.' (5)

'The funding body approved the final report.' (6)

We would like our algorithm to be able to recognize that Wordnet's first sense of the word body (*the entire physical structure of an organism (especially an animal or human being)*) is the most appropriate for sentence 5 and the third sense (*a group of persons associated by some common tie or occupation and regarded as an entity*) for sentence 6. If we calculate the most likely sense using all of the first 50 nearest neighbours in the thesaurus, we predict that sense 4 is the most frequently used sense (*the body excluding the head and neck and limbs*).

However, the two uses of the target word in 5 and 6 appear each in a very specific grammatical context. How can we exploit this local context to single out a certain subset of the 50 nearest neighbours, containing those words that are particularly relevant for (or more closely related to) the grammatical relation that the target word is involved in this particular sentence. The idea we pursue here is to look at those neighbours in the thesaurus that occur in the same grammatical relation as our target word and share a high mutual information (i.e. word and grammatical relation do not only occur frequently together, but also when you see one, there is a high probability that you see the other).

While creating the thesaurus we consider all the words that co-occur with a certain target word (where co-occur means that it appears in the same grammatical relation). We also calculate the mutual information of both the target word and the co-occurring word and the grammatical relation. Instead of throwing this information away after finishing an entry in the thesaurus, we now store this information in the grammatical relation database.

Since this database grows to enormous proportions (in the order of 200GB for the one built up while processing the BNC), we need to reduce its size to be able to work with it. If we only keep those entries in the database that involve the words in the thesaurus

and their 50 neighbours, we can reduce the database to manageable proportions. We experimented with reducing the number of entries in the database even further by limiting the number of entries per grammatical relations to the ones with the highest mutual information scores, but this only had a negative effect on the recall, without improving the precision. As we will see later, data sparseness is a serious issue and it is therefore not advisable to cut-out any usable information that we have at our disposal.

The word sense disambiguation procedure that uses the local context is then straightforward:

1. Parse the sentence with the target word (the word to be disambiguated).

2. If the target word is not involved with any of the 4 grammatical relations we considered for building up the thesaurus, local context can not be used.

3. Otherwise, consult the database to retrieve the co-occurring words:

   - Let GR be the set of triples $< w, r, x >$ from equation 4 in section 2.2 for target word w.
   - Let NGR be the set of triples $< n_j, r, x >$ from equation 4 for any neighbour $n_j \in N_w$
   - For all $w \in T$ and all top 50 $n \in N_w$, keep entries with $< *, r, x >$ in database.
   - Let SGR be the set of relations $< r, x >$ in the target sentence, where $I < w, r, x >$ and $I < n, r, x >$ are both positive (i.e. r, x are both in the target sentence and have high MI in BNC for both w and n.)

4. Compute the ranking score for each sense by applying to a modified version of the ranking equation 7 (compared to the original given in 3), where the k nearest neighbours are replaced by the subset found in the step 3.

Prevalence Score $ws\_lc_i =$

$$\sum_{n_j \in N_w} MI \times dss_{n_j} \times \frac{wnss(ws_i, n_j)}{\sum_{ws_{i'} \in WS_w} wnss(ws_{i'}, n_j)} \quad (7)$$

where the WordNet similarity score (*wnss*) is defined as before and let MI be $I < n, r, x >$, i.e. the Mutual Information given by the events of seeing the grammatical relation in question and seeing the neighbour.

2.4 An example

The fact that a subset of the neighbours in the thesaurus share some specific relations with the target word in a particular sentence is something that we wish to exploit for Word Sense Disambiguation. Let us have a closer look at the two example sentences 5 and 6 that we introduced in the previous section.

The grammatical relations that our target word *body* is involved with are (from sentences 5 and 6 respectively):[3]

'body' object of 'keep' for sentence 5 and (8)

'body' subject of 'approved' and (9)
'body' modified by the noun 'funding' for sentence 6

Since *keep* is a fairly general verb, it is not surprising that quite a few of the neighbours occur as object of *keep*. As a matter of fact, 28 of the first 50 neighbours share this relation. However, the good news is, that pretty much all the words associated with body-parts (such as *arm*, *hand*, *leg*, *face* and *head*) are among them.

The two grammatical relations that *body* is involved with in sentence 6, are more specific. There are just 6 neighbours that share the 'subject of approve' relation with *body* and another 5 that are used to modify the noun *body*. Among these words are the highly relevant words *organisation*, *institution* and *board*.

3 Evaluation on SemCor

The example in the last section shows that in certain cases the method performs the way we envisaged. However, we need a quantitative evaluation to get a proper picture of the method's usefulness. We performed a full evaluation on SemCor. In this experiment we limited our attention to *nouns* only. We further eliminated Proper Names and multi-word units from the test set. Since the nouns in both these categories are mostly monosemous, they are less interesting as test material and apart from that, they introduce problems (mostly parser related) that have little to do with the proposed method. A total of 73918 words were left to evaluate. Table 1 summarizes the results. The figure for recall for the 'First Sense' method is not given in the table, because we want to contrast the local context method with the first sense method. Whilst the first sense method will return an answer in most cases, the local context method proposed in this paper will not. Here we want to focus on how we can

[3] At the moment we only take 4 grammatical relations into account: Verb-Subject, Verb-Object, Adj-Noun and Noun-Noun modifier.

Method	Attempted	Correct	Wrong	Precision	Recall
Local Context	23235	11904	11331	0.512	0.161
First sense	23235	11795	11440	0.508	-

Table 1: Results of evaluation on the nouns in SemCor.

improve on using the first sense heuristic by taking local context into account, rather than give complete results for a WSD task.

There are several things to say about these results. First of all, even though the results for 'local context' are slightly better than for 'first sense', we expected more from it. We had identified quite a few cases like 5 and 6 above, where the local context seemed to be able to help to identify the right neigbours in order to make the difference. Below, we will discuss a few cases where the grammatical relations involved are so general, that the subset of neighbours is large and most importantly, not discriminative enough. It seems to be reasonable to expect that the latter cases will not influence the precision too much (i.e. a smaller group of neighbours will often give a different result, but some better, some worse).

The recall is also lower than expected. The first thought was that data sparseness was the main problem here, but additional experiments showed us that that is unlike to be the case. In one experiment we took a part of the GigaWord corpus ((Graff, 2003)), similar in size to the written part of the BNC (used in our original experiment) and built our grammatical relation database using the combined corpus. The recall went up a little, but at the price of a slightly lower precision.

3.1 Discussion

The main problem causing the low recall seems to be the small number of grammatical relations that we use for building the thesaurus. The four relations used (verb-subject, verb-object, noun-noun-modifier and adjective-noun-modifier) were chosen because of the parsers' high accuracy for these. For building the thesaurus, these grammatical relations suffice, since every word will occur in one of these relations sooner or later. However, whenever in a sentence the target word occurs outside these four relations, we are not able to look it up in our database. Nouns within prepositional phrases seem to be a major victim here. It should be straightforward to experiment with including prepositional phrase related grammatical relations. We will have to evaluate the influence of the introduced noise on creating the thesaurus. Alternatively, it is possible to use the four relations as before for creating the thesaurus and store the extra relations in our database just for look-up.

A second cause for missing target words is parser errors. Even though RASP will produce partial parses whenever a full parse of a sentence is not available, some loss is inevitable. This is a harder problem to solve. One way of solving this problem might be by using a *proximity thesaurus* instead of a thesaurus build using grammatical relatons. (McCarthy et al., 2007) reported promising results for using proximity based thesaurus for predicting predominant senses, with accuracy figures closely behind those achieved with a dependency based thesaurus.

One plausible reason why the method is not working in many cases, is the fact that the word to be disambiguated in the target sentence often occurs in a very general grammatical relation. For example, 'subject of' or 'direct object of' a verb like *have*. In these cases, most of the neighbors in the thesaurus will be selected. Even though it is clear that that would minimize the *positive* effect, it is not immediately obvious that this would have a *negative* effect. It might therefore be the case that the number of cases where the grammatical relation is a good selection criterion, is just lower than we thought (although this is not the impression that you get when you look at the data). We will need to establish a way of quantitatively evaluating this.

The Mutual Information score gives us a measure of the dependence between the grammatical relation and the word (neighbour of the target word) we are interested it. It gives us a handle on 'generality' of the combination of seeing both events. This means that for a very common grammatical relation, many words will be expected to co-occur with a frequency comparable to their general frequency in texts. The contrast with relation/word combinations for which this is not the case might be usable for identifying the cases that we want to exclude here.

4 Conclusions

In this paper we propose a completely unsupervised method for Word Sense Disambiguation that takes the local context of the target word into account. The starting point for this method is a method for automatically predicting the predominant senses of words. The grammatical relations that were used

to create the distributional similarity thesaurus is exploited to select a subset of the k neighbours in the thesaurus, to focus on those neighbours that are used in the same grammatical context as the word we want to disambiguate in the target sentence.

Even though the precision of our proposed method is slightly higher than for the predominant sense method, we are disappointed by the current results. We do believe that there is more mileage to be had from the method we suggest. Improvement of both recall and precision is on the agenda for future research. As we stated in the previous section, we believe that the lower than expected recall can be addressed fairly easily, by considering more grammatical relations. This is straightforward to implement and results can be expected in the near future.

A second approach, involving a thesaurus built on proximity, rather than grammatical relations will also be investigated. Considering the expected lower precision for this approach, we plan to use the proximity-based thesaurus as a 'back off' solution in case we fail to produce an answer with the dependency-based thesaurus. When the proximity-based thesaurus is in place, we plan to perform a full evaluation of the dependency versus the proximity approach.

Before we can deal with improving the local context method's precision, we need to have a better idea of the circumstances in which the method gets it wrong. We have identified a large group of examples, where it is unlikely that the method will be successful. A first step will be to develop a method to identify these cases automatically and eliminate those from the targets that we are attempting to try. In the previous section, we sketched how we think that we can achieve this by applying a Pointwise Mutual Information threshold. If we are successful, this will at least give us the opportunity to focus on the strengths and weaknesses of the method. At the moment, the virtues of the method seem to be obscured too much by dealing with cases that should not be considered.

More insight in the method can also be gained from trying to identify in which situations the method is more likely to get it right. At the moment we haven't broken down the results yet in terms of the target word's polysemy and/or frequency of use. Some grammatical relations might be more useful for identifying the intended sense than other. A detailed analysis could give us these insights.

We do believe there is a strong case to be made for using unsupervised methods for Word Sense Disambiguation (apart from (McCarthy et al., 2004)'s predominant sense method, other approaches include e.g. (Basili et al., 2006)). The predominant sense method has proven to be successful. However, applying the first sense heuristic should be limited to certain cases. We can think of the cases where the dominance of the predominant sense is so strong, that there is little to gain from doing a proper attempt to disambiguation or to the cases where 'everything else fails'. Ultimately, our goal is to find a balance between the dominance of the predominant sense and the strength of the evidence from the supporting context. If we are able to recognize the correct clues from the local context and use these clues to focus on those words with a high distributional similarity to the target word in the context in which the word is actually used, we can build on work on predicting predominant senses, to rely less on the first sense heuristic. This would be a good step forward for unsupervised WSD.

Acknowledgments

This work was funded by UK EPSRC project EP/C537262 "Ranking Word Senses for Disambiguation: Models and Applications", and by a UK Royal Society Dorothy Hodgkin Fellowship to the second author. We would like to thank Siddharth Patwardhan and Ted Pedersen for making the WN Similarity package available and Julie Weeds for the thesaurus software.

References

Basili, Roberto, Marco Cammisa, and Alfio Gliozzo. 2006. Integrating domain and paradigmatic similarity for unsupervised sense tagging. In *Proceedings of 7th European Conference on Artificial Intelligence (ECAI06)*.

Briscoe, Edward and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1499–1504, Las Palmas, Canary Islands, Spain.

Graff, David. 2003. English gigaword. Linguistic Data Consortium, Philadelphia.

Iida, R., D. McCarthy, and R. Koeling. 2008. Gloss-based semantic similarity metrics for predominant sense acquisition. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 561–568, Hyderabad, India.

Jiang, Jay and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *10th International Conference on Research in Computational Linguistics*, pages 19–33, Taiwan.

Koeling, Rob, Diana McCarthy, , and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Human Language Technology Conference and EMNLP*, pages 419–426, Vancouver, Canada.

Leech, Geoffrey. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL'98*, pages 768–774, Montreal, Canada.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590, December.

Miller, George A., Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.

Patwardhan, Siddharth and Ted Pedersen. 2003. The CPAN WordNet::Similarity Package. http://search.cpan.org/~sid/ WordNet-Similarity-0.05/.

Yarowsky, David and Radu Florian. 2002. Evaluating sense disambiguation performance across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.