Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Document Maps

Krister Lindén
Helsinki University of Technology, Neural Networks Research Centre, P.O.Box 9800, FIN-02015 HUT, Finland, (krister.linden@hut.fi)

Abstract.

Word sense disambiguation automatically determines the appropriate senses of a word in context. We have previously shown that self-organized document maps have properties similar to a large-scale semantic structure that is useful for word sense disambiguation.

This work evaluates the impact of different linguistic features on self-organized document maps for word sense disambiguation. The features evaluated are various qualitative features, e.g. part-of-speech and syntactic labels, and quantitative features, e.g. cut-off levels for word frequency.

It is shown that linguistic features help make contextual information explicit. If the training corpus is large even contextually weak features, such as base forms, will act in concert to produce sense distinctions in a statistically significant way. However, the most important features are syntactic dependency relations and base forms annotated with part of speech or syntactic labels.

We achieve 62.9 %±0.73 % correct results on the fine grained lexical task of the English SENSEVAL-2 data. On the 96.7 % of the test cases which need no back-off to the most frequent sense we achieve 65.7 % correct results.

Keywords: Linguistic features, Self-organized document maps, Semantic space, SENSEVAL-2, Word sense disambiguation

1. Introduction

Word sense disambiguation automatically determines the appropriate senses of a word in context. It is an important and difficult problem with many practical consequences for language-technology applications in information retrieval, document classification, machine translation, spelling correction, parsing, and speech synthesis as well as speech recognition.

The word sense disambiguation problem has been approached by traditional AI methods, such as hand-made rule sets or semantic networks, by knowledge-based methods using dictionaries or thesauri, and by corpus-based methods (Ide and Veronis, 1998). In this work we create a self-organized representation of the high-dimensional semantic space and use the representation for word sense disambiguation. For a textbook introduction to word sense disambiguation, see (Manning and Schütze, 1999). For recent comparisons of algorithms, see (Yarowsky

and Florian, 2002; Lee and Ng, 2002; SENSEVAL-2, 2001; Escudero et al., 2000), and for results of statistically combining methods, see e.g. (Florian et al., 2002; Florian and Yarowsky, 2002).

The methods vary in how different levels of context are selected and encoded. From a linguistic point of view the information included in the representation of context corresponds to approximations of morphological, syntactic and discourse context. The context is encoded by linguistic features. A linguistic feature means a word form or a combination of words and labels resulting from natural language processing. A collocation means linguistic features which co-occur in the same context. A topic is e.g. 'Elections in Iraq'. A domain is a collection of related topics. The global context of a word sense is the discourse. Yarowsky (1995) noted that there seems to be only one sense per collocation and that words tend to keep the same sense during a discourse. Leacock et al. (1998) pointed out that some words have non-topical senses which may occur in almost any discourse. Magnini et al. (2002) manually grouped the word senses for WordNet belonging to the same domain and were able to show that one domain per discourse is a better prediction than one sense per discourse.

Lee and Ng (2002) showed that the disambiguation effect of local linguistic features was considerable regardless of which learning method they chose achieving results between 57.2-65.4 % accuracy on the fine grained lexical task of the English SENSEVAL-2 data. Their analysis showed that adding more complex linguistic features to the base forms, e.g. syntax and part-of-speech labels, accounted for an absolute improvement of 8-9 % of the disambiguation result of the best algorithm. Yarowsky and Florian (2002) and Voorhees et al. (1995) compared several linguistic features and algorithms with the conclusion that major differences in the feature space was a more dominant factor than differences in algorithm architecture.

When studying the impact of different linguistic features on word sense disambiguation, a data structure representing semantic space makes it possible to keep constant the parameters of an algorithm evaluating the semantic space while varying the linguistic features. A mathematical structure for a representation of semantic space is proposed in (Lowe, 2001). Formally it is a quadruple $\langle A, B, S, M \rangle$, where B is the set of basis elements, e.g. linguistic features, A is the mapping between particular basis elements and each word in the language, S is the similarity measure between vectors of basis elements, and M is a transformation between two representations of semantic space, e.g. a dimensionality reduction. Steyvers and Tenenbaum (submitted) show that large-scale natural language semantic structures such as thesauri are characterized by sparse connectivity and strong local clus-

tering. Martinetz and Schulten (1994) showed that self-organizing maps tend to preserve the local neighborhood of the high dimensional space when projecting it onto a low dimensional display. Lindén and Lagus (2002) confirmed that a two-dimensional self-organized document map of a massive document collection has properties similar to a large-scale semantic structure or a thesaurus that is useful for word sense disambiguation.

A self-organized document map, created with the WEBSOM method (Kohonen et al., 2000; Honkela et al., 1996), represents the semantic space as ordered clusters of documents. In (Lindén and Lagus, 2002), a technique is proposed which calibrates the self-organized document map with a small batch of hand-tagged data and evaluates the map for word sense disambiguation. The technique is called THESSOM[1]. For an overview of the dataflow of the semi-supervised procedure, see Figure 1.

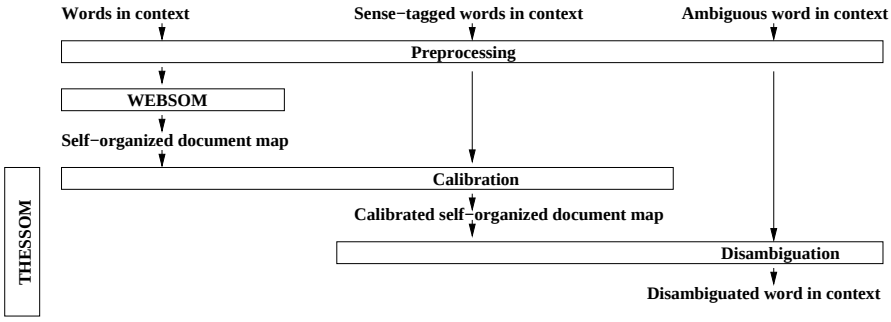

Figure 1. Dataflow of word sense disambiguation with self-organized document maps.

Schütze (1998) presented a related method for clustering data for word sense disambiguation. However, this is the first time the impact of several linguistic features on large-scale clustering is evaluated. The features evaluated are various qualitative features, e.g. part-of-speech and syntactic labels, and quantitative features, e.g. cut-off levels for word frequency. It is shown that using a rich set of linguistic features, such as base forms with part-of-speech or syntactic labels, produces a representation of semantic space currently achieving 62.9 %±0.73 % correct results on the fine grained lexical task of the English SENSEVAL-2 data. On the 96.7 % of the test cases which need no back-off to the most frequent sense we achieve 65.7 % correct results.

The rest of this article is organized as follows. First the WEBSOM and THESSOM methods are presented in Sections 2.1 and 2.2. Then the training, calibration and test data collections are introduced in

Section 3. The feature selection is described in Section 4. The word sense disambiguation experiments and results are presented and evaluated in Section 5. Sections 6 and 7 present the discussion and conclusion, respectively.

2. Methods

For word sense disambiguation it may be useful to know that *house* and *residence* are related and it may also be of interest whether in some context they are more closely related than *house* and *building*. However, it would be sufficient to know that *house* and *zinc mine* are unrelated in most contexts. It is unlikely that we need an accurate measure of whether they are more unrelated than e.g. *house* and *leg*.

The idea that most concepts are closely related to only a few others is supported by the research by Steyvers and Tenenbaum demonstrating that, when e.g. thesauri grow, new concepts are likely to be added to clusters of already locally tightly linked concepts. Conversely, most concepts and their concept clusters are only indirectly related occupying distant parts of the semantic space.

The concepts that are closely related in semantic space can be captured by a low-dimensional local metric. This idea is supported by Lowe in (1997) when he shows that most co-occurrence data is inherently very low-dimensional suggesting that in many cases two dimensions may be sufficient. Using single-value decomposition (SVD) he shows that 80 % of the co-occurrence information could be encoded in only a few SVD components (Lowe, 2001).

SOM is a non-linear projection of high-dimensional space onto a low dimensional display. SOM tends to preserve a fairly accurate image of the local high-dimensional neighborhood, even if similar guarantees cannot be given for distant parts of the high-dimensional space (Martinetz and Schulten, 1994). We call this the local neighborhood preservation property of SOM.

First we present an outline of the WEBSOM method, which is a method for creating large two-dimensional self-organized document maps using the SOM algorithm. Then we present THESSOM which is a method for reading WEBSOM maps for the purpose of word sense disambiguation.

2.1. Creating document maps with WEBSOM

Assume that we have a collection of documents of various lengths from different domains. We also have a domain classification of the

documents. All the words occurring in a document are thereby related to the domain of the document in the context of the other words in the document. If we consider each word or combination of words to be a dimension in semantic space, we wish to create a low-dimensional projection of the high dimensional semantic space such that documents with similar content end up near each other.

The WEBSOM method (Kohonen et al., 2000; Honkela et al., 1996) uses the Self-Organizing Map (SOM) algorithm (Kohonen, 1997; Ritter and Kohonen, 1989) to organize a large document collection in an unsupervised way onto a two-dimensional display called the map. The WEBSOM method is outlined below and the parameters which were used in the experiments for this article are briefly recapitulated.

The SOM map consists of a set of map units ordered on a two-dimensional lattice. By virtue of a model vector stored with each map unit, searches can be performed on the map in order to locate the most similar model vector for a new document or short context (Kohonen, 1997). The map unit of the most similar model vector is called the best-matching unit.

In WEBSOM, documents are encoded by using the bag-of-words vector space model. The features in the document vectors are weighted before the vectors are normalized. The cosine measure (dot product for normalized vectors) is used for measuring similarities between document vectors. Documents similar in content are located near each other on the ordered map display (Kohonen et al., 2000).

WEBSOM uses domain-entropy weighting. The entropy weighting of a feature describes how well the feature is focused on some domains. Let $P_d(w)$ be the probability of a randomly chosen instance of the feature w occurring in domain d and $|D|$ the number of domains. The entropy is $H(w) = -\sum_{d=1}^{|D|} P_d(w) log P_d(w)$ and the weight $W(w)$ of feature w is defined as $W(w) = H_{max} - H(w)$, where $H_{max} = log(|D|)$. (Kohonen et al., 2000)

Rare features have low prediction power and can be discarded with a global frequency cut-off value. High-frequency features with low information content can be placed on a stop word list. The remaining number of features may still be substantial. For computational reasons the dimensionality of the representation is reduced. WEBSOM uses random projection (Kaski, 1998), a well-documented technique which projects each feature onto N randomly chosen encoding features, where N typically is a parameter in the range of $3 \ldots 5$ for an encoding feature vector with $300 \ldots 1000$ elements. The choice of N is motivated by the required mapping accuracy, but values above 5 usually give little or no additional accuracy. The random projection procedure has been shown to retain the distance information of the original high-dimensional

space while introducing only a small amount of random noise (Kaski, 1998). In addition, random projection is much faster than e.g. SVD.

The document maps in WEBSOM are created in several steps. Initially, a small map is created, on which the data is organized. Then the map is magnified and retrained in several steps to the desired level of magnification indicated by a set of magnification parameters. (Kohonen et al., 2000)

2.2. Calibration and disambiguation with THESSOM

Assume that we have a word in context. The word has a number of possible word senses. We wish to determine which word sense is used in the current context. We also have a collection of sense-tagged samples of the word in various contexts. Assume that we have a representation of semantic space in the form of a document map. The document map decides which of the sense-tagged samples in our sample collection are relevant in a particular context by displaying similar samples near each other on the map. By also displaying the untagged sample on the document map and looking at the nearby sense-tags we may determine which sense is appropriate for the word in the current context.

WEBSOM creates a two-dimensional projection of a document collection called a self-organized document map. The document map is regarded as an instrument for word sense disambiguation. In order to be able to read the indications of the instrument, i.e. the unlabeled WEBSOM map, it needs to be calibrated. In (Lindén and Lagus, 2002), a method is presented which calibrates a self-organized document map and uses it for word sense disambiguation. The method is called THESSOM. In (Lindén, 2003), the method is presented in detail. Here we recapitulate the main ideas of calibration and word sense disambiguation with THESSOM.

When we get a short sample document, we preprocess it in the same way as the training data for the WEBSOM map encoding the sample into a document vector of linguistic features. A document vector containing the word and its context is used. By matching the document vector of a sample document s with each unit on the map we get a similarity reading for each map unit. The best-matching unit for a labeled sample s_l can be labeled with the label l. This is called calibration. The hypothesis is that similarity of meaning equals similarity of word context, which manifests itself in the labels on nearby map units.

We note that in general most of the best readings for a sample s are located on the map around the best-matching map unit, i.e. the N-best-matching map units for s are usually near the best-matching unit. Without much loss of information we may restrict our calcula-

tions to the map units within a radius r of the best-matching map unit. As predicted by the local neighborhood preservation property of SOM, the closest neighboring units on the map are also more likely to represent data that have been part of the same data cloud in the original high-dimensional space. We use this property when we create a sparse indicator array for a sample s on the WEBSOM map. For each map unit m, we set the indicator value $I_{s,r}[m]$ to one, if the unit is among the N-best-matching units of s and within a radius r on the map from the best-matching unit:

$$I_{s,r}[m] = \begin{cases} 1 \text{ if } m \in B_s^N \wedge d(m, B_s^1) < r, \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

where B_s^N is the set of N-best-matching units of s, B_s^1 is the best-matching unit, $d(m, B_s^1)$ is the map lattice distance between the map unit m and the best-matching map unit B_s^1. The map lattice distance is the Euclidean distance between the map units on the map lattice.

We use the indicator arrays to create a similarity function, where s_u is an unlabeled sample, l is a label for which we have calibration samples and r is the neighborhood radius. This gives us the THESSOM function presented in (Lindén, 2003)[2]:

$$\text{THESSOM}(s_u, l, r) = \sum_{i=1}^{m_l} I_{s_{l,i},r} \cdot I_{s_u,\infty} \,. \quad (2)$$

where m_l is the number of sample documents with label l. We calculate the similarity between the unlabeled sample document s_u and a labeled sample document s_l with regard to the WEBSOM map by taking the dot product of their indicator arrays. We allow the unlabeled sample to have more than one possible meaning and consequently it may have more than one cluster of almost equally well-matching units on the map. No restricting radius is therefore set for the N-best-matching units of the unlabeled sample.

The sense-tag l is determined for an unlabeled word s_u in context by the function $\arg\max_{l \in L} \text{THESSOM}(s_u, l, r)$, where L is the set of labels with calibration data. If no single winner is found using r, the globally most frequent of the winning senses can be chosen, if a single winner is needed. If no labeled data sample is near enough, i.e. $\leq r$, on the map display, instead of the local decision strategy, a global strategy is applied: a majority vote is taken among all the sense-tagged samples of that word.

The disambiguation is illustrated in Figure 2 with the THESSOM algorithm. The figure shows a WEBSOM map with the named map

units $p_{0...8}$. The winning label is o for the unlabeled sample providing classification of the unclassified sample.

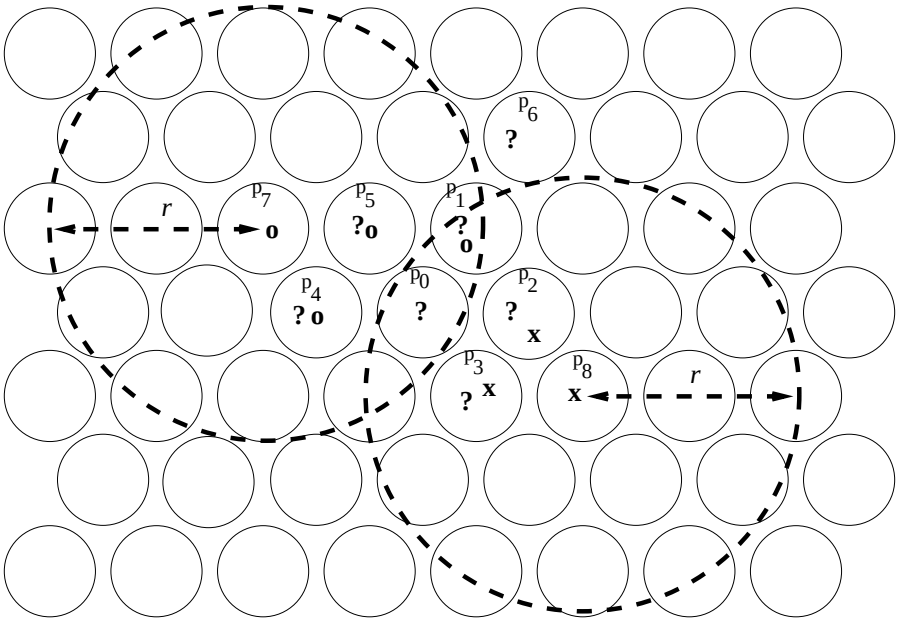

Figure 2. Word sense disambiguation with THESSOM. Each map unit on the WEBSOM map is represented as a small circle. We have one unclassified test sample s_u shown with the label ? on its N-best-matching units $B^1_{s_u} = p_0$, $B^N_{s_u} = \{p_0, p_1, p_2, p_3, p_4, p_5, p_6\}$. We have two labels $L = \{o, x\}$ with one calibration sample each displayed on their N-best-matching units $B^1_{s_o} = p_7$, $B^N_{s_o} = \{p_1, p_4, p_5, p_7\}$ and $Bs_x{}^1 = p_8$, $B^N_{s_x} = \{p_2, p_3, p_8\}$, respectively. When $r = 2$, THESSOM$(s_u, o, 2) = 3$ and THESSOM$(s_u, x, 2) = 2$, which selects o as the sense tag for s_u.

3. Data sets

In order to compare the performance of WEBSOM and THESSOM to other word sense disambiguation systems, we use the training and test data from the English lexical sample task of the SENSEVAL-2 exercise in 2001. First we introduce the training data collection, which is used for training the WEBSOM map. Then we present the calibration and test data collections, which are used by the THESSOM algorithm.

3.1. Training data

The training data for WEBSOM was the English lexical sample task of the SENSEVAL-2 exercise in 2001 without the sense labels. The training data consists of 8611 samples. The samples are instances of 73 base forms in context, i.e. 29 nouns, 29 verbs and 15 adjectives. We call these base forms keywords and the samples keywords in context. The lexicon used for the sense inventory of the SENSEVAL-2 data is WordNet 1.7. WordNet also has multi-word entries. If the 73 base forms are taken as part of an existing WordNet multi-word entry in context, we get an inventory of 323 base forms, i.e. 177 nouns, 131 verbs and 15 adjectives.

3.2. Calibration data

The calibration data is used by THESSOM for associating an explicit word sense to an area of the WEBSOM map. The calibration data is the same 8611 samples of the English lexical sample task of the SENSEVAL-2 exercise as the training data including the sense labels. There are a total of 861 word senses represented by the calibration data samples.

3.3. Test data

The test data in SENSEVAL-2 consists of 4328 samples from the British National Corpus and the Wall Street Journal. The samples are instances of 73 base forms in context, i.e. 29 nouns, 29 verbs and 15 adjectives. If the 73 base forms are seen as parts of existing WordNet multi-word entries in context, we get an inventory of 241 base forms with 125 nouns, 101 verbs and 15 adjectives. All of the multi-word entries of the test data are not represented in the calibration data.

There are a total of 638 word senses represented in the test data. All of the word senses of the test data are not represented by the calibration data samples.

3.4. Baselines and significance tests

As the same calibration and test data is used for all the tests, the baselines remain the same. The most frequent sense baseline, which would be achieved by always selecting the most frequent of the candidate senses of a word, is correct in 47.6 % of the cases. Human inter-annotator-agreement is 85.5 % on the SENSEVAL-2 data (SENSEVAL-2, 2001). When multi-word entries are treated as base forms, the most frequent sense baseline is 53.0 %.

The significance of the results is tested against the baseline and relative to one another using the McNemar test (Somes, 1983). McNemar is a non-parametric test using matched pairs of labels. It is essentially a sign test on nominal data.

4. Feature selection

This work explores the importance of linguistic features for improving the quality of the representation of semantic space. When selecting linguistic features for the word sense disambiguation task we can do this in a binary on/off fashion for each feature. This corresponds to having weights of 1 or 0. This is referred to as qualitative feature selection. A more nuanced picture of each feature gives the features weights between 0 and 1. This is referred to as quantitative feature selection.

4.1. Qualitative feature selection

We briefly motivate the features selected for the experiments, i.e. base forms, parts of speech, head syntactic features, syntactic dependencies and upper/lower case as well as the shape of the context.

Traditionally, a base form is the form found in a dictionary. Some word forms may have several base forms depending on context. In English the base form is often unique. The ambiguity is mainly between parts-of-speech with the same base form. One notable exception is the analysis of participles, e.g. "a *drunk* driver/has *drunk* a lot" with base forms *drunk/drink* or "was *heading* south/the newspaper *heading* is" with base forms *head/heading* etc. The correct base form can be determined in context as a side-effect of part-of-speech tagging.

An intermediate level before full dependency parsing is head syntax, which only indicates in which direction the head word is and what part of speech the head word is. The main advantage of head syntax is that it avoids attachment ambiguities, e.g. in "the man on the hill *with* the telescope" the preposition *with* is tagged as a dependent of some noun to the left, e.g. *with N<*.

Dependency syntax builds parse trees with one head word for each word. Each head word may have several words depending on it. For a rule-based approach to dependency syntax, see (Tapanainen and Järvinen, 1997), and for a statistical approach, see (Samuelsson, 2000).

The original case of a word form is an annotation entered by the author of a document. If the word forms are normalized so that capital letters are turned into lower case, at least one prominent distinction is lost between e.g. 'Church' as an institution and 'church' as a building.

In word sense disambiguation the size and shape of the context is important, e.g. nouns often benefit more from a long context than verbs, which often depend on a local context (Agirre and Martinez, 2001; Martinez and Agirre, 2000; Ide and Veronis, 1998).

4.2. Quantitative feature selection

Let us assume that we have a large corpus with some domain classification of individual documents. If we were to see more text from a given domain, the features from the same domain are more likely to co-occur than features from different domains. It follows that features occurring in many domains are less relevant because they have less domain prediction power. We may wish to de-emphasize such generally occurring features in order to guide a clustering algorithm in the right direction. These were the motivations for using entropy weighting in WEBSOM.

The entropy weighting of features can be compared with the results of the tf*idf (= term frequency inverse document frequency) weighting used in information retrieval. In our experiments the idf is taken to be the inverse domain frequency which is similar to the feature weighting suggested by (Cabezas et al., 2001).

5. Experiments

In addition to the linguistic features of the training material, this work also explores what parameters of the training process improve self-organized document maps as representations of semantic space. In this section we first introduce the parameters and the features that were tested. We then study the sensitivity of the disambiguation result to linguistic features of the data and to some of the main parameters of WEBSOM.

5.1. Parameter selection

In order to find the best parameter selection we used 10-fold cross-validation on the training data of the SENSEVAL-2 exercise. The parameters were then used for disambiguating the separate test data of the SENSEVAL-2 exercise.

5.1.1. *WEBSOM parameters*

The *number of maps* was one separate map for each keyword, i.e. 73 maps for the SENSEVAL-2 training material. We also tried using one

general map for the whole training material. Separate maps for each keyword worked best.

The tests were carried out with the *frequency cut-off value* of the features set at different values between 0 and 5 in order to see the impact of a minimum cut-off value. No cut-off value was needed for the keyword maps.

The *weighting* of features was calculated for each map using the the senses of the keywords as domains. Entropy weighting worked better than tf*idf weighting.

The *random projection* produced a feature vector of length 300 and each feature was projected onto 3 encoding features for the keyword maps. For the general map we used a vector length of 1000 with 5 encoding features.

The *size of the map* was at most 720 map units for each of the keyword maps and 12,096 map units for the general SENSEVAL-2 map. On a SOM map each map unit is associated with a model vector. From a statistical point of view a good generalization of model vectors and good predictability of the results should be achieved with a sufficient average of training samples per model vector. This was tested at first, and the disambiguation results were statistically significant already with maps adherent to this principle, but the results improved for all feature combinations when the maps were magnified until there were more map units than training samples. We call this supermagnification. For details of the magnification procedure, see (Kohonen et al., 2000). The supermagnification added granularity and detail to the map surface, because model vectors representing large clusters of training samples tended to attract many different readings through context averaging. These clusters were spread out through magnification adding precision to the map. All maps created for the experiments in this work are supermagnified.

5.1.2. *THESSOM parameters*

The parameter r was varied between $1 \ldots 4$. As N-best units 1, 5, 11 and 15 were tested. The best performance was achieved, when r was set to 3 and the N-best units to 15. These parameters apply to all the experiments below.

5.1.3. *Feature selection*

The extracted features can be divided into global features (GLOB), local features (LOC) and syntactic features (SYN). A dependency syntax parser by Connexor (Connexor, 2002) provided all the linguistic analysis used in the feature extraction. For suggestions on other features, see (SENSEVAL-2, 2001; Lee and Ng, 2002).

The global features were the correct base forms in context (GLOB), which were extracted for all the words in the sentence surrounding the keyword. The sentence was found to be the most useful context for global features in the experiments with WEBSOM in (Lindén and Lagus, 2002).

The local features (LOC) were extracted from a window of ±3 words around the keyword. We used the bigram collocations of the keyword and the base forms, the unnormalized word form of the keyword, and the head syntax and part of speech labels of the base forms. If a ±3-window extended over the sentence border, it was padded with empty words.

The syntactic features (SYN) consisted of the dependency information in the n-tuples $\langle W_1, M_1, R, W_2, M_2 \rangle$, where W_1 and W_2 are base forms in a dependency relation R, and M_1 and M_2 are the word class features of W_1 and W_2, respectively. If M_2 is a preposition or a coordinator, the n-tuple $\langle W_1, W_2, R, W_3, M_3 \rangle$ was also extracted, where W_3 is in a dependency relation to W_2, and M_3 is its set of morphological features.

In Table I, we present a summary of the linguistic features for a sample sentence.

5.2. Test results

The test results measure the percentage of correctly classified test data samples, a.k.a. the classification accuracy. The best test result is 62.9 % correct classifications with a standard deviation of 0.73 %. This was 68.5 % for adjectives, 67.1 % for nouns and 56.5 % for verbs. The best combination of parameters and linguistic features were separate keyword maps using all features (SYN+LOC+GLOB) without frequency cut-off. On the 96.7 % of the test cases which need no back-off to the most frequent sense we achieve 65.7 % correct results.

A sensitivity analysis in Table II with combinations of features shows that the SYN and LOC feature types perform almost equally well separately. However, together SYN+LOC perform statistically significantly better than either of them separately. By adding the GLOB feature type to SYN and LOC or their combination, we observe a small increase in the overall performance.

We repeated the experiment using a frequency cut-off value of 3. The results in Table III show that in general the result degrades when adding a cut-off value to the feature frequency. This is perhaps contrary to expectations of more frequent features having better prediction power. It may, however, be due to the small training material when we create one map for each keyword.

Table I. The linguistic features of a test sample for the keyword *church*. The bag of base forms is abbreviated as BoBF, head syntax label as HS, and part of speech label as PoS.

Sample	... most have quaint churches and other features of interest.		
GLOB	*correct base forms in sentence context*		
BoBF	... many have quaint church and other feature of interest		
LOC	*collection of features in window context*		
-3 bigram	< -3, many >	-3 HS, PoS	< -3, NH, PRON-SUP-PL >
-2 bigram	< -2, have >	-2 HS, PoS	< -2, VA, V-PRES >
-1 bigram	< -1, quaint >	-1 HS, PoS	< -1, >N, A-ABS >
0 string	< 0, churches >	0 HS, PoS	< 0, NH, N-PL >
+1 bigram	< +1, and >	+1 HS, PoS	< +1, CC, CC >
+2 bigram	< +2, other >	+2 HS, PoS	< +2, >N, DET >
+3 bigram	< +3, feature >	+3 HS, PoS	< +3, NH, N-PL >
SYN	*collection of dependency relation features*		
With attr	< church, N, attr, A, quaint >		
Coord by	< church, N, cc, CC, and >		
Coord with	< church, and, cc, N, feature >		
Object of	< church, N, obj, V, have >		

Table II. Classification accuracy by part-of-speech in different feature contexts without frequency cut-off using keyword maps and entropy weighting.

SENSEVAL-2	all	adj	noun	verb
SYN+LOC+GLOB	62.9	68.5	67.1	56.5
SYN+LOC	62.0	66.8	67.0	55.0
SYN+GLOB	61.6	67.5	67.2	53.5
LOC+GLOB	60.1	67.8	65.4	51.6
SYN	59.8	64.2	65.3	52.5
LOC	59.9	64.8	65.6	52.3
GLOB	56.0	65.8	60.9	47.1

We repeated the first experiment using a tf*idf weighting scheme. The results in Table IV show that the tf*idf scheme is statistically significantly worse than the entropy weighting for all feature combinations except when using only GLOB features.

We also repeated the first experiment using one general map with entropy weighting and frequency cut-off value 0. The results in Table V

Table III. Classification accuracy by part-of-speech in different feature contexts with frequency cut-off 3, keyword maps and entropy weighting.

SENSEVAL-2	all	adj	noun	verb
SYN+LOC+GLOB	62.0	67.1	67.6	54.3
SYN+LOC	60.4	63.8	66.7	52.9
SYN+GLOB	60.8	64.6	66.6	53.5
LOC+GLOB	60.7	67.5	66.3	52.5
SYN	58.1	63.7	64.1	49.9
LOC	59.9	65.2	65.4	52.3
GLOB	55.4	62.0	62.1	46.1

Table IV. Classification accuracy by part-of-speech in different feature contexts without frequency cut-off using keyword maps and tf*idf weighting.

SENSEVAL-2	all	adj	noun	verb
SYN+LOC+GLOB	59.7	66.8	66.0	50.5
SYN+LOC	59.4	65.9	64.7	51.4
SYN+GLOB	58.9	65.5	64.9	50.2
LOC+GLOB	56.9	64.7	61.6	49.1
SYN	58.0	63.8	64.3	49.3
LOC	56.5	62.2	62.3	48.4
GLOB	55.3	65.2	60.6	45.9

show that for the general map the results are statistically significantly worse than using separate maps for each keyword. In particular one may notice that using any combination with the GLOB features seems to detract from the impact of the SYN and LOC features on a general map. The SYN feature alone has low coverage, so in the SYN+GLOB combination the GLOB feature improves coverage achieving a positive but insignificant contribution to the performance.

5.3. Importance of test results

Overall results of more than 54.1 % on the SENSEVAL-2 data are statistically significantly above the baseline with a rejection risk $p < 0.05$ using the McNemar test. Results above 55.5 % on the SENSEVAL-2 data are significant with a rejection risk of $p < 0.001$.

Table V. Classification accuracy by part-of-speech in different feature contexts without cut-off frequency using a general map with entropy weighting.

SENSEVAL-2	all	adj	noun	verb
SYN+LOC+GLOB	58.9	66.7	64.5	50.1
SYN+LOC	59.5	65.6	65.7	50.8
SYN+GLOB	58.6	67.6	64.0	49.6
LOC+GLOB	57.0	66.2	61.7	48.5
SYN	58.0	63.5	64.7	49.1
LOC	58.1	65.8	64.4	48.7
GLOB	54.4	66.2	59.9	44.0

6. Discussion

In (Lee and Ng, 2002) the impact of different feature combinations extracted from the SENSEVAL-2 material is evaluated on several supervised learning systems and compared to the three best systems in the SENSEVAL-2 exercise. The best reported performance without combining classifiers on the English SENSEVAL-2 data for a fine-grained lexical task is 65.4 % with the best results being in the range 62.9–65.4 %, i.e. 66.8–73.2 % for adjectives, 66.8–69.5 % for nouns and 56.3–61.1 % for verbs (Lee and Ng, 2002; SENSEVAL-2, 2001), see Table VI. Only by combining classifiers has a better overall result of 66.5 % been achieved in (Florian and Yarowsky, 2002).

Table VI. Comparison of accuracy by part-of-speech for different algorithms on SENSEVAL-2 data without statistically combining classifiers.

SENSEVAL-2	all	adj	noun	verb
Lee and Ng	65.4	68.0	68.8	61.1
SENSEVAL-2/1	64.2	73.2	68.2	56.6
SENSEVAL-2/2	63.8	68.8	69.5	56.3
SENSEVAL-2/3	62.9	66.8	66.8	57.6
THESSOM	62.9	68.5	67.1	56.5

WEBSOM is a self-organizing method, i.e. the organization of the map is caused by the interaction of the data elements and the self-organizing principle. It is interesting that the organization of a good map for word sense disambiguation is crucially due to a rich linguistic feature set. Without this the impact of the other parameters is

negligible. A rich linguistic feature set is a way to explicitly describe the function of each word in its current context reducing the need to consider very long contexts. A local keyword context reduces the amount of noisy features, which is important if the corpus is as small as the SENSEVAL-2 training material.

In order to see the impact of a very large general semantic space, we can compare our present results to the results for THESSOM reported in (Lindén and Lagus, 2002) using the WEBSOM patent abstract map of approximately 7,000,000 patent abstracts (Kohonen et al., 2000). With the patent abstract map as a single unified semantic space we achieved the modest 54.1 % classification accuracy (65.3 % for adjectives, 59.6 % for nouns and 46.9 % for verbs), which was statistically significant with a rejection risk of $p < 0.05$ (Lindén and Lagus, 2002). Even if the patent abstract map is huge, it lacks usage information for many of the word senses included in the SENSEVAL-2 test data. However, if we use only GLOB features in the SENSEVAL-2 training data and one general map, see Table V, the results are statistically on a par with the patent abstract map. From this we can conclude that a very large unspecialized corpus like the patent abstract collection is comparable to a specialized corpus like the SENSEVAL-2 training data, if the linguistic analysis is shallow.

Only when we apply a more advanced linguistic analysis, do we make substantial progress. Our current study indicates that verbs in particular gain in performance by the addition of more complex linguistic features. This is important for applications relying heavily on word sense information related to verbs, e.g. machine translation applications.

Another crucial improvement in the SOM environment comes from having separate maps for each keyword. Separate maps correspond to partitioning the original high-dimensional semantic space. The WEBSOM map of each partition gives a more accurate picture of the semantic distinctions we are interested in.

When the word senses of a word are created by a lexicographer, all the usages of the word are inspected and assigned to a sense according to the context of the word. By extracting the features that lexicographers observe and by using them when creating separate self-organized maps for each keyword, we get closer to the word senses identified by lexicographers.

The advantage of the present architecture is that we can use the feature extraction procedure and the WEBSOM map creation on unlabeled samples of the keyword in context. By adding more unlabeled samples we are likely to improve the precision of the keyword map. Unlabeled samples are abundant so it remains to be seen if larger more

fine-grained keyword maps can be even better calibrated for word sense disambiguation.

Another advantage is that heterogeneous information sources, such as different document collections, can be made into WEBSOM maps and used as separate representations of semantic space. The different WEBSOM maps can be combined for word sense disambiguation by using rank-ordered classification of the results of the THESSOM algorithm.

7. Conclusion

This work explores what linguistic features of the training material and parameters of the training process improve self-organized document maps as representations of semantic space. Linguistic features make contextual information explicit. If the corpus is large enough even contextually weak features will act in concert to produce sense distinctions in a statistically significant way.

The THESSOM algorithm is tested on the SENSEVAL-2 benchmark data and shown to perform on a par with the top three contenders of the SENSEVAL-2 exercise. We also show that adding more advanced linguistic analysis to the feature extraction seems to be essential for improving the classification accuracy.

We achieve 62.9 %±0.73 % correct results on the fine grained lexical task of the English SENSEVAL-2 data using the THESSOM algorithm on the best WEBSOM map. On the 96.7 % of the test cases which need no back-off to the most frequent sense we achieve 65.7 % correct results.

Acknowledgements

I am grateful to the WEBSOM team at the Neural Networks Research Centre of the Helsinki University of Technology for the permission to use the patent abstract map and the WEBSOM tools for these experiments. I am also indebted to Krista Lagus for many fruitful discussions as well as to Lauri Carlson, Kalervo Järvelin, Sami Kaski, Mathias Creutz, and Timo Honkela for commenting on earlier versions of this paper. I am also grateful to three anonymous reviewers for their valuable comments.

Notes

[1] THESSOM is an acronym for THESaurus-like Self-Organized document Map. In Old Norse *af thessom* means *among these*.

[2] In (Lindén, 2003), we also include a component in THESSOM for extrapolating in case the set of calibration samples is very small in comparison to the number of map units. This component is not statistically significant on the SENSEVAL-2 data and is therefore omitted in this article.

References

Agirre, E. and D. Martinez: 2001, 'Knowledge Sources for Word Sense Disambiguation'. In: V. M. et al. (ed.): *TSD 2001, Proceedings of the International Conference on Text, Speech and Dialogue*. pp. 1–10, Springer-Verlag Berlin Heidelberg.

Cabezas, C., P. Resnik, and J. Stevens: 2001, 'Supervised Sense Tagging using Support Vector Machines'. In: *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*. Toulouse, France.

Connexor: 2002, 'Machinese Syntax'. [http://www.connexor.com/].

Escudero, G., L. Màrquez, and G. Rigau: 2000, 'A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation'. In: C. Cardie, W. Daelemans, C. Nedellec, and E. Tjong Kim Sang (eds.): *Proceedings of CoNLL-2000 and LLL-2000*. pp. 31–36, Lisbon, Portugal.

Florian, R., S. Cucerzan, C. Schafer, and D. Yarowsky: 2002, 'Combining Classifiers for Word Sense Disambiguation'. *Natural Language Engineering* **8**(4), 327–341.

Florian, R. and D. Yarowsky: 2002, 'Modeling Consensus: Classifier Combination for Word Sense Disambiguation'. In: *Proceedings of EMNLP-2002*. pp. 25–32.

Honkela, T., S. Kaski, K. Lagus, and T. Kohonen: 1996, 'Newsgroup exploration with WEBSOM method and browsing interface'. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Ide, N. and J. Veronis: 1998, 'Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art'. *Computational Linguistics* **24**(1), 1–40. Special Issue on Word Sense Disambiguation.

Kaski, S.: 1998, 'Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering'. In: *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, Vol. 1. Piscataway, NJ: IEEE Service Center, pp. 413–418.

Kohonen, T.: 1997, *Self-Organizing Maps (Second Edition)*, Vol. 30 of *Springer Series in Information Sciences*. Berlin: Springer.

Kohonen, T., S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela: 2000, 'Organization of a Massive Document Collection'. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery* **11**(3), 574–585.

Leacock, C., M. Chodorow, and G. A. Miller: 1998, 'Using Corpus Statistics and WordNet Relations for Sense Identification'. *Computational Linguistics* **24**(1), 147–165. Special Issue on Word Sense Disambiguation.

Lee, Y. K. and H. T. Ng: 2002, 'An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation'. In: *Proceedings of EMNLP-2002*. pp. 41–48.

Lindén, K.: 2003, 'Word Sense Disambiguation with THESSOM'. In: *Proceedings of the WSOM'03 – Intelligent Systems and Innovational Computing*. Kitakuyshu, Japan.

Lindén, K. and K. Lagus: 2002, 'Word Sense Disambiguation in Document Space'. In: *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics*. Hammamet, Tunisia.

Lowe, W.: 1997, 'Semantic representation and priming in a self-organizing lexicon'. In: J. A. Bullinaria, D. W. Glasspool, and G. Houghton (eds.): *Proceedings of the Fourth Neural Computation and Psychology Workshop: Connectionist Representations*. London, pp. 227–239, Springer-Verlag.

Lowe, W.: 2001, 'Towards a theory of semantic space'. In: J. D. Moore and K. Stenning (eds.): *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Mahwah NJ, pp. 576–581, Lawrence Erlbaum Associates.

Magnini, B., C. Strapparava, G. Pezzulo, and A. Gliozzo: 2002, 'The Role of Domain Information in Word Sense Disambiguation'. *Natural Language Engineering* **8**(4), 359–373.

Manning, C. D. and H. Schütze: 1999, *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Martinetz, T. and K. Schulten: 1994, 'Topology Representing Networks'. *Neural Networks* **7**(3), 507–522.

Martinez, D. and E. Agirre: 2000, 'One Sense per Collocation and Genre/Topic Variations'. In: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong.

Ritter, H. and T. Kohonen: 1989, 'Self-Organizing Semantic Maps'. *Biological Cybernetics* **61**, 241–254.

Samuelsson, C.: 2000, 'A statistical theory of dependency syntax'. In: *Proceedings of COLING-2000. ICCL.*

Schütze, H.: 1998, 'Automatic Word Sense Discrimination'. *Computational Linguistics* **24**(1), 97–123. Special Issue on Word Sense Disambiguation.

SENSEVAL-2: 2001, 'Training and testing corpora'. [http://www.cis.upenn.edu/∼cotton/senseval/corpora.tgz].

Somes, G.: 1983, 'McNemar test'. In: S. Kotz and N. Johnson (eds.): *Encyclopedia of statistical sciences*, Vol. 5. New York: Wiley, pp. 361–363.

Steyvers, M. and J. B. Tenenbaum: submitted, 'The large-scale structure of semantic networks: statistical analyses and a model of semantic growth'. *Cognitive Science*.

Tapanainen, P. and T. Järvinen: 1997, 'A non-projective dependency parser'. In: *Proceedings of 5th Conference on Applied Natural Language Processing*. pp. 64–71.

Voorhees, E. M., C. Leacock, and G. Towell: 1995, *Computational Learning Theory and Natural Language Learning Systems 3: Selecting Good Models*, Chapt. Learning context to disambiguate word senses, pp. 279–305. Cambridge: MIT Press.

Yarowsky, D.: 1995, 'Unsupervised word-sense disambiguation rivaling supervised methods'. In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*. Cambridge, MA, pp. 189–196.

Yarowsky, D. and R. Florian: 2002, 'Evaluating Sense Disambiguation Across Diverse Parameter Spaces'. *Natural Language Engineering* **8**(4), 293–310.