Word Sense Disambiguation with THESSOM

Krister Lindén
Helsinki University of Technology, Neural Networks Research Centre,
P.O.Box 9800 (Tammasaarenkatu 3), FIN-02015 HUT, Finland,
Phone: +358-505909014, E-mail: Krister.Linden@hut.fi

Keywords: Word sense disambiguation, Self-organized document maps, Semantic space

April 23, 2003

***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.

In this article we formalize THESSOM, which is an algorithm for word sense disambiguation using self-organized document maps created with WEBSOM. The algorithm is tested on the SENSEVAL-2 benchmark data and shown to be on a par with the top three contenders of the SENSEVAL-2 competition.

We also show that the performance of the algorithm improves when using more advanced linguistic features for creating the WEBSOM maps.

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 handmade rule sets or semantic networks, by knowledge-based methods using dictionaries or thesauri, and by corpus-based methods [1]. For a textbook introduction to word sense disambiguation, see [2]. For recent comparisons of algorithms, see [3, 4, 5, 6], and for results of statistically combining methods, see e.g. [7, 8].

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. For the purpose of this paper, 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. 'War in Iraq'. A discourse is a collection of documents related to a topic. A domain is a collection of topics. The global context of a word sense is the discourse. In [9], Yarowsky noted that there seems to be only one sense per collocation and that words tend to keep the same sense during a discourse. In [10], Leacock & al. pointed out that some words have non-topical senses which may occur in almost any discourse. Magnini & al. [11] 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. In [4], Lee and Ng showed that the disambiguation effect of linguistic features occurring in a local context 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 form analysis, e.g. syntax and part-of-speech labels, accounted for an absolute improvement of 8-9 % of the disambiguation result of the best algorithm.

A mathematical structure for a representation of semantic space is proposed in [12]. 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. In [13], Steyvers and Tenenbaum show that large-scale natural language semantic structures such as thesauri are characterized by sparse connectivity and strong local clustering. Martinetz & al. [14] showed that self-organizing maps tend to preserve the local neighbourhood of the high dimensional space when projecting it onto a low dimensional display. Lindén and Lagus [15] confirmed that a 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 [16, 17], represents semantic space as ordered clusters of documents. In [15], a technique is proposed which calibrates the self-organized document map with a small batch of hand-tagged data and

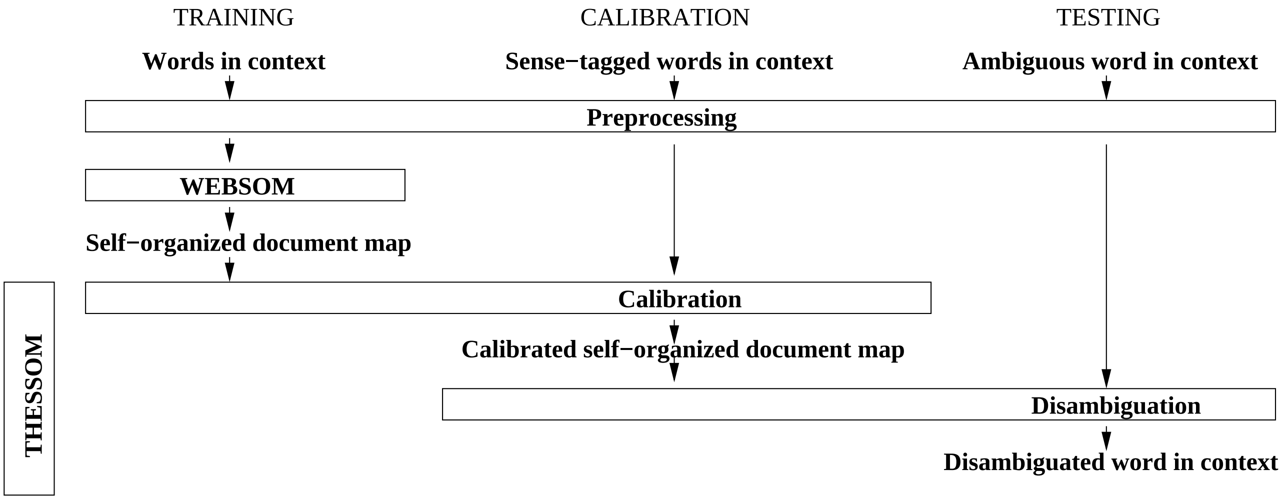
TRAINING
CALIBRATION
TESTING
Words in context
Sense−tagged words in context
Ambiguous word in context
Preprocessing
WEBSOM
Self−organized document map
Calibration
Calibrated self−organized document map
Disambiguation
Disambiguated word in context
THESSOM

Figure 1: Data flow of word sense disambiguation with self-organized document maps

evaluates the map for word sense disambiguation. The technique is called THESSOM[1]. For an overview of the data flow, see Figure 1.

In this article we introduce a formalization of the THESSOM algorithm and test the algorithm on the English SENSEVAL-2 benchmark corpus achieving 62.8 %±0.73 % correct results on the fine grained lexical task.

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

2.1 Creating document maps with WEBSOM

The WEBSOM method [16, 17] uses the Self-Organizing Map algorithm [18] to organize a large document collection onto a two-dimensional display called the map. The map provides a general view into the document collection visualizing similarity relations between the documents on the map display.

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 documents. Documents similar in content are located near each other on the ordered map display [16].

WEBSOM uses domain-entropy weighting. The entropy weighting of a feature describes how well the feature is focused on some domains [16].

For computational reasons the dimensionality of the representation is reduced by using random projection, which projects each feature onto F randomly chosen encoding features. 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 [19].

2.2 Calibration and interpretation with THESSOM

In [15], a technique is presented which calibrates a self-organized document map and evaluates it for word sense disambiguation. The disambiguation is based on relevant samples of a word in context. The relevance of the samples is decided by the self-organized document map by displaying similar samples near each other. The map is calibrated and interpreted by the algorithm called THESSOM. Below is a formalization of the algorithm.

The two-dimensional document map display $\mathcal{M} \subset \mathbb{R}^2$ is regarded as an instrument for word sense disambiguation. In order to read the indications of the instrument, it needs to be calibrated. The map is calibrated by positioning a set of data samples S with known readings T on the map display. Each data sample $s_i \in S$ is a sense-tagged word in context. The word in context is treated as a small document, from which linguistic features are extracted. The same linguistic features from similar contexts are extracted for calibration as when creating the WEBSOM map. The linguistic features of the sample are encoded as a doc-

[1] THESSOM is an acronym for THESaurus-like reading of a Self-Organized document Map.

ument vector, which is positioned on the map. The document vector matches a map unit with location l_{ij} to a degree indicated by a real number $m_{l_{ij}} \in [0, 1]$. The N locations $l_{i1} \ldots l_{iN}$ with the highest degree of matching are the N best-matching locations.

The word sense disambiguation is based on finding relevant samples of a word in context. The WEBSOM map displays similar samples near each other. The distance between two locations x and y on the map is defined as the map lattice distance $d(x, y)$. The closest locations on the map display are likely to be relevant if they are images of samples from the same data cloud in the original high-dimensional space. If the best matching location is an image of the outskirts of a data cloud, WEBSOM may have projected a portion of some other data cloud representing a different sense onto a neighbouring location. Because distant map locations among the N best-matching map units of a sample s are likely to represent different senses, only the locations l_j within a distance less than $d \in \mathbb{R}$ from a reference location k are considered, see Figure 2. The considered locations create an area K on the map around k, in which the N best-matching map units of s are similar to the reference location k. In order to let a sample with better-matching map units create a larger area than a sample with worse-matching map units, the distance from k to other locations in K is scaled by the degree of matching between the data sample and the map units. The map locations of the area K are $K(k, m_k, d, N, s) = \{l_j \in \mathcal{M} |\ d(k, l_j) < d * m_k * m_{l_j},\ m_k, m_{l_j} \in [0, 1],\ j = 1 \ldots N, \}$.

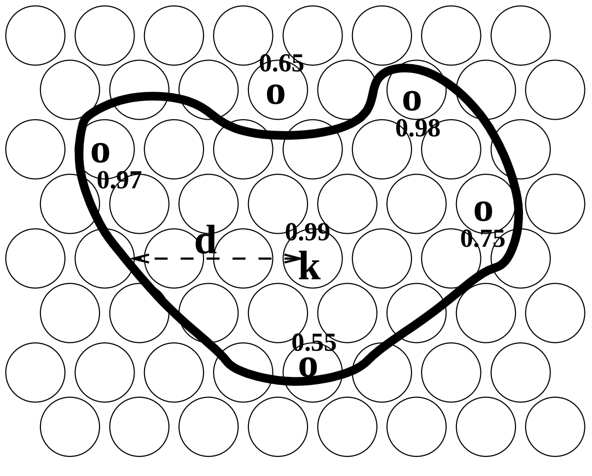

Figure 2: The map locations of the area $K(k, m_k, d, N, s)$, where k is the reference location, $m_k = 0.99$, $d = 4$, $N = 5$, and the data sample s has the label o. The sample s is shown on the 5 best-matching map locations annotated with the corresponding degree of matching.

When a test sample s_0, i.e. an ambiguous word in context, needs disambiguating, s_0 is processed identically to the sense-tagged data samples and the N best-matching locations $l_{01} \ldots l_{0N}$ on the map are found. The N best-matching map locations of the test data sample create the test sample area $K_0 = K(l_{01}, m_{l_01}, \infty, N, s_0)$.

The sense-tag for s_0 is selected by a majority vote according to the following principles:

(I) The N-best locations of the calibration data samples $s_i, i = 1 \ldots |S|$, create the calibration sample areas $K_i = K(l_{i1}, m_{l_{i1}}, d_0, N, s_i)$. The set of data samples with the sense-tag $t \in T$ is denoted $S_t \subset S$. The votes are collected from the sense-tags of the map locations l in the intersections of the sense-tagged areas K_i and the test sample area K_0. The votes for sense-tag t is $v_0(t) = \sum_{s_i \in S_t} \sum_{l \in K_i \cap K_0} 1$.

(II) Additional votes for a sense-tag t are collected from the extrapolation areas $L_{n,i}$: the calibration data samples $s_i, i = 1 \ldots |S|$, whose 1-best locations l_{i1} act as reference locations for the N-best locations of the test data sample s_0 within a distance $d_n, n = 1 \ldots D$, create the extrapolation areas $L_{n,i} = K(l_{i1}, m_{l_{i1}}, d_n, N, s_0)$. The votes are collected from the sense-tags of the map locations l in the intersections of the sense-tagged areas $L_{n,i}$ and the test sample area K_0. The votes for sense-tag t is $v_n(t) = v_{n-1}(t) + \sum_{s_i \in S_t} \sum_{l \in L_{n,i} \cap K_0} 1$.

(III) The sense-tag t for an ambiguous word s_0 in context is determined by the function $\arg\max_{t \in T} v_n(t)$. The function is evaluated for each $n = 0 \ldots D$, until a single winner is found. If no single winner is found within $n \leq D$ in the local map context, the globally most frequent of the winning senses is chosen.

(IV) If no calibration data sample is near enough, i.e. $\leq d_D$, 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.

3 Data set

In order to compare the performance of THESSOM to other systems for word sense disambiguation, the data was taken from the English lexical sample task of the SENSEVAL-2 competition [5].

The training data is used both as training and calibration data for the WEBSOM map. The training data consists of 8611 samples from the same corpora as the test material.

The test data 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. The lexicon used for the sense inventory of the SENSEVAL-2 data is WordNet 1.7.

3.1 Baselines and significance tests

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 [5]. When the base forms are preprocessed as parts of existing WordNet multi-word entries in context, their set of sense ambiguities is constrained so that the most frequent sense baseline is 53.0 %.

The significance of the results is tested against the baselines using the McNemar test [20]. McNemar is a non-parametric test using matched pairs of labels. It is essentially a sign test on nominal data.

4 Feature selection

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 feature weights of 1 or 0. This is referred to as qualitative feature selection. A more nuanced picture of each feature gives weights between 0 and 1. This is referred to as quantitative feature selection and is done with entropy weighting in WEBSOM.

4.1 Qualitative feature selection

Below only brief motivations for the linguistic features used in the experiments are presented. For a textbook introduction to natural language processing, see [2].

Traditionally, a base form is the form of a keyword 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.

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 [21], and for a statistical approach, see [22].

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 consistently turned into lower case, a prominent distinction in English is lost between e.g. ‘Church’ as an institution and ‘church’ as a building.

In word sense disambiguation the length of the keyword context is important. Nouns often benefit from a longer context than verbs and adjectives, which mostly depend on the local context [23, 1].

5 Experiments

First the parameters selected for the experiments in this work are introduced and then the results of the experiments are presented.

5.1 Parameter selection

We used the training samples of the SENSEVAL-2 data, which were disambiguated words in context, for calibration. For the parameter selection we used 10-fold cross-validation on the training and calibration data to find the best-performing parameter combinations, which were then used for disambiguating the test data.

5.1.1 WEBSOM parameters

The *frequency cut-off value* was varied between 1, 2 and 3, but 1 was found to perform best. The *entropy weighting* of features was calculated separately for each base form data set using the sense groups of the base form. The *random projection* produced a feature vector of length 300 and each feature was projected onto 3 encoding features. The *size of a map* was 720 map units with one map for each of the 73 base form data sets in the training data.

5.1.2 THESSOM parameters

The parameter d_0 was varied between $0 \ldots 3$. As N-best units $1, 5, 11, 15$ were tested. The maximum distance d_D was varied between $4 \ldots 7$. The best performance was achieved, when d_0 was set to 2, the N-best units to 15, and d_D to 5.

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 [24] provided the linguistic analysis used in the feature extraction. For suggestions on feature extraction, see [5, 4].

The global features were the correct base forms in context for all the words in the sentence surrounding the keyword.

The local features were the correct base form in context combined with its part-of-speech in a window of

±3 words around the keyword, bigram collocations of the keyword and the base forms in a ±3-word window, the unnormalized word form of the keyword, and the head syntax label combined with the correct base form in context for the keyword. If a ±3-window extended over the sentence border, it was padded with empty words.

The syntactic features 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 morphosyntactic features of W_1 and W_2, respectively. If M_2 is a preposition, 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.

5.2 Test results

The test results were obtained using a separate test data set, namely the English lexical task test corpus of the SENSEVAL-2 competition. The test results measure the percentage of correctly classified test data samples, a.k.a. the classification accuracy. The test result is 62.8 % correct classifications with a standard deviation of 0.73 %. This was 67.2 % for adjectives, 68.4 % for nouns and 55.6 % for verbs.

In order to estimate the impact of the different feature contexts on the classification accuracy of each part-of-speech, we did a sensitivity analysis as shown in Table 1 with combinations of feature sets keeping the other parameters as specified above.

SENSEVAL-2	all	adj	noun	verb
GLOB	56.0	65.0	61.5	46.8
LOC	59.6	65.4	65.0	51.8
SYN	60.4	65.9	66.7	52.0
LOC+GLOB	60.2	67.2	65.0	52.6
SYN+GLOB	61.3	66.7	67.3	53.1
SYN+LOC	62.0	66.7	67.3	54.8
SYN+LOC+GLOB	62.8	67.2	68.4	55.6

Table 1: Classification accuracy of part-of-speech by feature context

5.3 Importance of test results

Total results above 55.5 % on the SENSEVAL-2 data are significantly above the most frequent sense baselines with a rejection risk of $p < 0.001$ using the McNemar test.

6 Discussion

In [4] 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 competition. The best reported performance of a single approach 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 [4, 5]. Only by combining classifiers has a better overall result of 66.5 % been achieved in [8].

It is interesting to compare our present results to the results for THESSOM reported in [15] using the WEBSOM patent abstract map [16]. The modest 54 % classification accuracy (65.3 % for adjectives, 59.6 % for nouns and 46.9 % for verbs) was statistically significant with a rejection risk of $p < 0.05$ [15]. Even if the patent abstract map is huge, it lacks usage information about many of the word senses included in the SENSEVAL-2 test data. However, if we use only base forms as linguistic features in the current training data, as was the case for the patent abstract map, this seems to bring about approximately the same result in this study for adjectives and verbs. It is only when we apply a more advanced linguistic analysis that we are able to make substantial progress. Our current study shows that, as might be expected, verbs in particular gain in performance by the addition of more complex linguistic features. This is important for applications relying heavily on the semantics of verbs, e.g. machine translation applications.

WEBSOM is a self-organizing method which takes domain information into account. In practice, separate maps for each of the 73 base form data sets in the test data correspond to partitioning the original feature space into distinct subspaces. However, a significant bottle-neck is the small amount of data for each base form. This needs to be addressed in future research.

7 Conclusion

In this work we have introduced a formalization of the THESSOM algorithm. The 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 competition. We also show that adding more advanced linguistic analysis to the feature extraction seems to be essential for improving the classification accuracy.

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 WEBSOM tools for these experiments.

References

[1] Nancy Ide and Jean Veronis, "Introduction to the special issue on word sense disambiguation: The state of the art," *Computational Linguistics*, vol. 24, no. 1, pp. 1–40, March 1998, Special Issue on Word Sense Disambiguation.

[2] Christopher D. Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

[3] David Yarowsky and Radu Florian, "Evaluating sense disambiguation across diverse parameter spaces," *Natural Language Engineering*, vol. 8, no. 4, pp. 293–310, December 2002.

[4] Yoong Keok Lee and Hwee Tou Ng, "An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation," in *Proceedings of EMNLP-2002*, 2002, pp. 41–48.

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

[6] Gerard Escudero, Lluís Màrquez, and German Rigau, "A comparison between supervised learning algorithms for word sense disambiguation," in *Proceedings of CoNLL-2000 and LLL-2000*, Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, Eds. 2000, pp. 31–36, Lisbon, Portugal.

[7] Radu Florian, Silviu Cucerzan, Charles Schafer, and David Yarowsky, "Combining classifiers for word sense disambiguation," *Natural Language Engineering*, vol. 8, no. 4, pp. 327–341, December 2002.

[8] Radu Florian and David Yarowsky, "Modeling consensus: Classifier combination for word sense disambiguation," in *Proceedings of EMNLP-2002*, 2002, pp. 25–32.

[9] David Yarowsky, "Unsupervised word-sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, Cambridge, MA, 1995, pp. 189–196.

[10] Claudia Leacock, Martin Chodorow, and George A. Miller, "Using corpus statistics and wordnet relations for sense identification," *Computational Linguistics*, vol. 24, no. 1, pp. 147–165, March 1998, Special Issue on Word Sense Disambiguation.

[11] Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo, "The role of domain information in word sense disambiguation," *Natural Language Engineering*, vol. 8, no. 4, pp. 359–373, December 2002.

[12] Will Lowe, "Towards a theory of semantic space," in *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, J. D. Moore and K. Stenning, Eds., Mahwah NJ, 2001, pp. 576–581, Lawrence Erlbaum Associates.

[13] M. Steyvers and J. B. Tenenbaum, "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth," *Cognitive Science*, to appear.

[14] Thomas Martinetz and Klaus Schulten, "Topology representing networks," *Neural Networks*, vol. 7, no. 3, pp. 507–522, 1994.

[15] Krister Lindén and Krista Lagus, "Word sense disambiguation in document space," in *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics*, Hammamet, Tunisia, 2002.

[16] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Vesa Paatero, and Antti Saarela, "Organization of a massive document collection," *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 574–585, May 2000.

[17] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen, "Newsgroup exploration with websom method and browsing interface," Tech. Rep. A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[18] Teuvo Kohonen, *Self-Organizing Maps (Second Edition)*, vol. 30 of *Springer Series in Information Sciences*, Springer, Berlin, 1997.

[19] Samuel Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proceedings of IJCNN'98, International Joint Conference on Neural Networks*, vol. 1, pp. 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[20] G. Somes, "Mcnemar test," in *Encyclopedia of statistical sciences*, S. Kotz and N. Johnson, Eds., vol. 5, pp. 361–363. Wiley, New York, 1983.

[21] Pasi Tapanainen and Timo Järvinen, "A non-projective dependency parser," in *Proceedings of 5th Conference on Applied Natural Language Processing*, 1997, pp. 64–71.

[22] Christer Samuelsson, "A statistical theory of dependency syntax," in *Proceedings of COLING-2000. ICCL.*, 2000.

[23] Eneko Agirre and David Martinez, "Knowledge sources for word sense disambiguation," in *TSD 2001, Proceedings of the International Conference on Text, Speech and Dialogue*, V. Matoušek et al., Ed. 2001, LNAI 2166, pp. 1–10, Springer-Verlag Berlin Heidelberg.

[24] Connexor, "Machinese syntax," [http://www.connexor.com/], 2002.