Self-Organized Ordering of Terms and Documents in NSF Awards Data

Mikaela Klami and Timo Honkela
Adaptive Informatics Research Centre
Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland
email: {mikaela.klami, timo.honkela}@tkk.fi

Keywords: text mining, term extraction, self-organizing map

***Abstract*—**
We present the results of an analysis of a text corpus of 129,000 abstracts of NSF-sponsored basic research projects between years 1990 and 2003. The methods used in the analysis include term extraction based on a reference corpus and an entropy measure, and the Self-Organizing Map algorithm for the formation of a term map and a document map. Methodologically, the basic approach is based on earlier developments, such as word category maps and the WEBSOM method, but in the level of details, we report several new aspects and quantitative comparison results between methodological preprocessing variants in this article. The data covers a quite large proportion of US-based scientific research during recent years. The analysis results indicate the basic patterns discernable in the data, both at the level of the awards and at the terminology used in them.

1 Introduction

Text mining aims at extracting relevant, novel or interesting information from text corpora, including tasks such as text categorization, text clustering, modeling relationships between entities, and document summarization. The Self-Organizing Map (SOM) [9] is a classical neurally inspired method for data analysis and visualization that has also been widely used for text mining, through e.g. clustering document collections and the unsupervised detection of topics or conceptual domains. The SOM-based text analyses have ranged from finding emergent structures among a small collection of document titles [13], software documentation [14], newsgroup discussions [3], curriculum vitaes [4], etc. The basic architecture of a text mining system based on the Self-Organizing Map can be divided into three parts: preprocessing, map formation, and map visualization and use.

In preprocessing, various methods have been used to vectorize text documents, including random projection [16, 7] and Latent Semantic Indexing [2]. These methods also help to reduce the dimensionality of the original representation that is usually based on a rather large vocabulary. In this article, we present a dimensionality-reducing method that selects terms in an intelligent manner. In our case, the term extraction is based on a reference corpus used to distinguish scientific terms from other forms of language use [5]. Following this idea, proposed by [1], the domain-specific terms can be identified by comparing a word's rank in a specialized text to its rank in a large, well-balanced corpus covering many aspects and domains of language use. Our term extraction results are further refined using an entropy-based measure in a classical manner.

The map formation is based on the basic SOM algorithm or on one of its many variants [9], e.g. methods developed for the creation of very large maps [11]. The map use covers tasks such as exploration, search and filtering [3]. The visualization may be based, for instance, on a landscape, sea-view, outer space or library metaphor. In this article, the main purpose of the analysis is to explore the structure of one specific database, the 129,000 abstracts of NSF-sponsored basic research projects from years 1990 to 2003. In particular, we wish to reveal the relationship between the conceptual contents of the research descriptions, and the organizational structure according to which the projects have been classified.

2 Methods

In the following, we describe the methods used in the analysis of the NSF data, including term extraction and term map and document map formation.

2.1 Term extraction

In order to be able to train Self-Organizing Maps (SOM) on textual data (in our case, the NSF awards corpus), a list of terms or keywords describing the data as accurately as possible needs to be obtained. In our approach, this process of term extraction has three steps.

2.1.1 Extraction of frequent terms

First, a list of term candidates is extracted from the corpus, based on the terms' frequencies in the corpus. Terms may consist of one or more words. Usually the length is limited to three or four words.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

2.1.2 Using reference corpus

Next, an entirely different reference corpus is utilized to select such terms from the frequency list that are common in the corpus under examination, but rare in the reference corpus; i.e., terms that best distinguish the particular corpus from the reference corpus. The purpose of the reference corpus is to represent a certain language in general, and typically a very large and well-balanced corpus is chosen for the task. Thus, comparing a particular-domain corpus to the reference corpus should reveal terms that are specific to that domain only.

The term lists of both corpora are sorted according to the term frequencies, and all terms are given their rank in the list (all terms that have an equal frequency receive an equal rank). Then, the terms of the actual corpus are processed one by one, calculating for each term the ratio of the ranks of the term in the two corpora. For example, if the term "research" should have a rank of 14 in the first corpus and 372 in the reference corpus, its ratio would be 14/372 = 0.0376. Terms that could not be found in the term list of the reference corpus receive a ratio of their own rank divided by the largest rank in the reference corpus plus one.

Finally, the terms are sorted in ascending order according to their ratios. Terms that receive a small ratio are the ones we are interested in, since they were more common in the particular corpus than in the reference corpus. On the other hand, the middle ground of the term list is now occupied by general, probably non-specific terms that were common in both corpora, and the end of the list has terms that were more frequent in the reference corpus [5].

2.1.3 Refining with entropy over classes

Finally, the list of terms is further refined by calculating the entropy of the terms over the classes of the corpus documents. The terms that have the lowest class entropy best distinguish the classes from each other, thus making good features for analyzing document categorization.

2.2 Term maps

The Self-Organizing Map (SOM) [9] algorithm can be used to create word category maps that describe the relations of words. Typically, the analysis is based on the textual contexts of the words, and interrelated words with similar contexts will appear close to each other on the map. In its simplest form, the context consists of the neighboring words in a sentence ([6, 16], see also [15]). Each unique word is represented by a vector x_i, whose values (frequencies of feature words in the word's context) are calculated as an average over all of its appearance in the text corpus.

In this article, however, we instead use the categorization information available in the NSF corpus. Each term is encoded based on its frequency in different categories. A term-by-category matrix X is generated. Each extracted term is represented by a row in matrix X, and each NSF category is represented by a column. An individual entry in the matrix, x_{ij}, represents the frequency of the term i in category j. The rows were normalized to unit length, and a term SOM was trained using the normalized matrix X.

2.3 Document maps

The WEBSOM method was developed to facilitate an automatic organization of text collections into visual and browsable document maps [3, 12]. Based on the Self-Organizing Map (SOM) algorithm [9], the system organizes documents into a two-dimensional plane in which two documents tend to be close to each other if their contents are similar. The similarity assessment is based on the full-text contents of the documents.

In the original WEBSOM method [3], the similarity assessment consisted of two phases. In the first phase, a word category map (WCM) [16, 6] was formed to detect similarities of words based on their contexts. Latent Semantic Indexing (LSI) method [2] is nowadays often used for a similar purpose. In the second phase, the document contents were mapped on the WCM. The distribution of the words in a document over the WCM was used as the feature vector for the document SOM. Later, the WEBSOM method was streamlined to facilitate processing of very large document collections [11] and the use of the WCM as a preprocessing step was abandoned. In this work, we follow this direct approach to document map creation.

3 Data

We analyzed the NSF Research Awards Abstracts corpus[1]. The data set consists of 129 000 relatively short abstracts in English, describing awards granted for basic research by the US National Science Foundation during the period 1990–2003. For each abstract, there is a considerable amount of metadata available, including the abbreviation code of the NSF division that processed and granted the award in question. We will use this NSF division code as the class of each document. The data was preprocessed by extracting the actual abstract texts from the documents, removing most non-word characters, and converting the texts into lowercased format.

As explained in Section 2.1.2, we also needed a reference corpus to represent the English language in general. We use the English versions of the texts in the large, multilingual Europarl corpus [8] that was originally developed for the purposes of evaluating machine translation systems. It consists of proceedings of the European Parliament in 11 different European languages, up to 28 million words per language. While the Europarl corpus hardly covers all possible types of English texts, it should still be sufficient for extracting terminology specific to scientific applications.

[1] http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html

4 Experiments

The experiments covered both term-level and document-level analysis. The data provided also a chance to quantitatively measure different methodological variants.

4.1 Term selection

For the purposes of the term selection process, we considered terms consisting of one to three consecutive words. We extracted a list of all the unigrams, bigrams and trigrams that occurred at least 10 times in the NSF corpus. Using the same criteria, we also extracted a similar list of terms from the Europarl corpus used as a reference corpus.

Next, we compared the actual NSF corpus term list to the reference corpus term list as described in Section 2, and selected 2000 terms that were the most characteristic of the NSF texts, i.e. that best distinguished our particular corpus from the Europarl data.

Finally, we calculated the entropy for the terms over the classes (the NSF division that granted each award) of the abstract documents, and picked the 1000 terms that best distinguished the classes from each other.

In order to verify the quality of the term selection process, we will provide a quantitative comparison of our method to simple alternative term selection and weighting approaches in Section 5.3.

4.2 Term SOM

With the term list ready, we proceeded with training a SOM on the selected terms to see how they relate to each other. In addition to giving an insight into the nature of the NSF corpus in itself, such a term map can also be perceived as a final stage of term selection, i.e. a final examination of the quality and nature of the terms that were extracted.

We calculated the frequencies of the 1000 best terms in each class of documents, and used the values as features for a SOM. The term SOM of a size of 20×30 nodes was trained using the SOM_PAK [10] software package, and all 1000 terms were projected on the resulting map.

4.3 Document SOM

Next, we used the same 1000 highest ranked terms as features to train a document SOM on the corpus, using the popular tf-idf weighting scheme [17]. Before analysis, the documents with too sparse feature vectors were discarded (8471 in total, keeping 120 529 documents).

We then trained a 30×45 document SOM on the remaining documents, again using the SOM_PAK package, and projected them all on the resulting map, using the class code as a label. In the next section, different kinds of illustrations of this large all-documents SOM are provided.

5 Results

In the following, we describe the analysis results at the level of terms and documents.

5.1 Term Map

We trained a term SOM on the selected 1000 terms. The resulting map in Fig. 1 reflects the occurrences of our terms in the different categories of documents of the NSF corpus. Terms that occurred frequently in the same class of documents are organized close to each other on the map, forming clusters of similar terms. On our map, there are many interesting clusters of terms that have similar meanings or that are obviously related to the same field of research. However, due to space limitations, we will only present here two such clusters.

The left-hand side of Fig. 1 presents a part of the term map that displays words involving Neuroscience. There are for example words related to sensory mechanisms and cognition (“perception”, “visual”, “sensory”, “behavioral”), brain physiology (“brain”, “receptor”, “neurons”, “tissue”), and neurology (“nerve”, “nervous”, “nervous system”, “neural”). The cluster of neuroscientific words seems to be situated inside a larger cluster of terms from Biology and other Life Sciences, and also on the border of a cluster involving Chemistry.

On the right-hand side of Fig. 1, there is a cluster of terms related to Geology. We have terms involving formations of rock (“rock”, “magma”, “volcanic”, “stratigraphic”, “sediment”, “seismic”), the structure of our planet (“plate”, “crust”, “mantle”, “continental”, “tectonic”), geochemistry (“geochemistry”, “mineral”, “isotope”, “isotopic”), and also bodies of water (“river”, “basin”, “sea”). Here again, the cluster of Geology seems to be bordered by terms involving other Earth Sciences, like Oceanography and Atmospheric Sciences.

The term SOM does indeed seem to give a good insight to the various research topics involved in the documents of the NSF corpus. Even though we only selected 1000 terms to represent the whole corpus, these terms alone already seem to reveal many interesting features of the nature of the corpus. The approach we have adopted in this paper to the task of term selection thus seems qualitatively reasonable.

5.2 Document Map

The document SOM was trained on the documents of the NSF Awards corpus using the selected 1000 terms as features, and all the over 120 000 documents were projected on the resulting map. In order to extract meaningful information from this document SOM representing our very large document collection, we studied the documents in each of the classes individually instead of looking at all of the documents at once.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

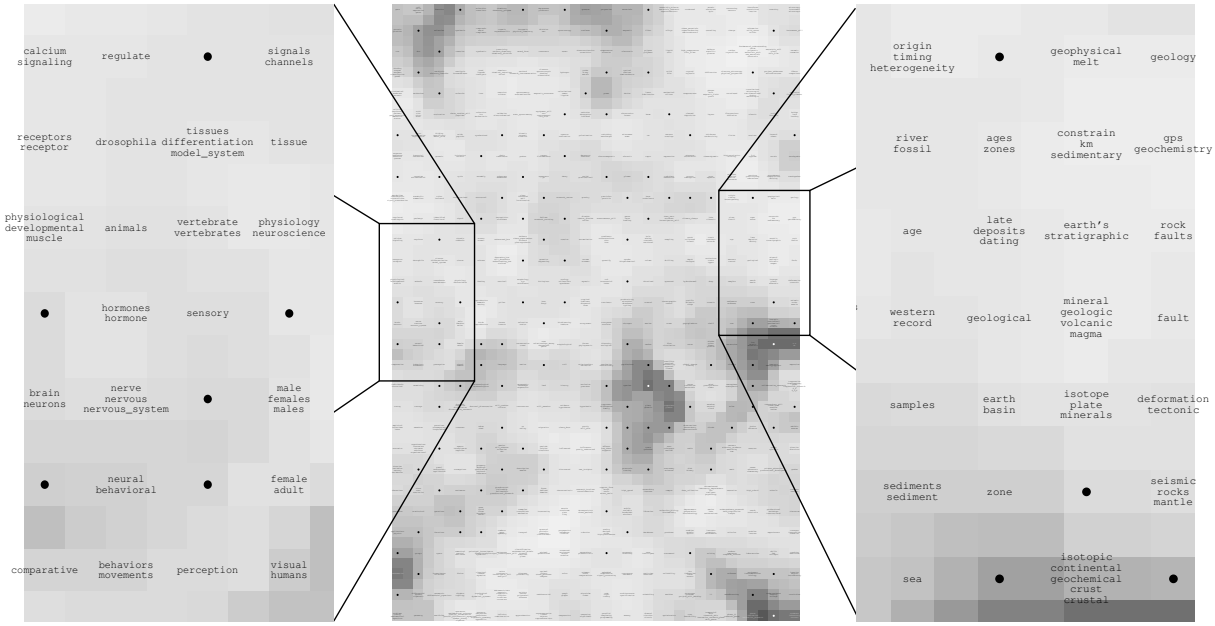

Figure 1: A 20×30 Self-Organizing Map trained with the selected 1000 terms in the NSF corpus. Terms that occurred frequently in the same class of documents are represented close to each other on the map. Two clusters of similar terms have been highlighted for a more detailed inspection. Here, terms consisting of more than one word are represented as compounds that are joined with underscores.

For studying the individual classes, we calculated hit histogram representations of the classes' documents. In a hit histogram map, the area of the "hits" in the nodes of the map is proportional to the number of times those nodes were chosen for the Best Matching Unit (BMU) of a document. From a hit histogram representation, it is easy to see how the documents of a given class are distributed over the map, and the classes can be compared to each other by studying how similar their document hit histograms are.

Fig. 2 depicts the hit histogram maps of the documents in three NSF corpus classes that involve the field of Biology. The acronym IBN stands for "Division of Integrative Biology and Neuroscience", MCB is "Division of Molecular and Cellular Biosciences", and BCS stands for "Division of Biological and Critical Systems". As can be seen from the figure, the class IBN has become divided into two main areas. The larger of these areas coincides quite closely with the area of the class MCB. The smaller separate area of the class IBN is rather close to the area occupied by BCS.

These similarities and disparities between the document distributions of the classes seem to suggest that the award applications handled by the first two divisions of the National Science Foundation have more things in common, whereas the contents of the applications addressed to the last one, the Division of Biological and Critical Systems, are somewhat different. Indeed, it appears that in the official Fields of Science classification of the NSF, the first two fields have been placed under the broader field of "Biological, Behavioral and Social Sciences", but the third one is situated under a different umbrella term, "Engineering". Even though these upper-level fields were not taken into any account in our SOM analysis, the emergent features of the data itself appear to have reflected it into the structure of our document SOM. And, on the other hand, the SOM analysis also reveals the divided structure of the class IBN.

In another case study, we examined the document distributions of classes that involve, in one form or another, the field of Education. The classes were DUE ("Division of Undergraduate Education"), DGE ("Division of Graduate Education"), ESI ("Division of Elementary, Secondary and Informal Education"), HRD ("Division of Human Resources Development"), REC ("Division of Research, Evaluation and Communication"), and ESR ("Education System Reform Programs"). The hit histogram representations of these classes can be found in Fig. 3.

Again, the hit histogram representations display similarities between the classes. All six classes share in their document distribution a concentration of hits in the upper right quarter of the map. Rather than being identically distributed, however, the classes instead seem to have a tendency of complementing each other's distributions. This indeed appears to be the case when performing a more detailed comparison between the classes DUE and DGE, or HRD and REC. For example, in the first case, the nodes with the highest number of documents in class DGE are close to the dense areas of class DUE, but the exact nodes have little or no documents from class DUE. However, as the areas are highly intertwined and as also the contents of

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

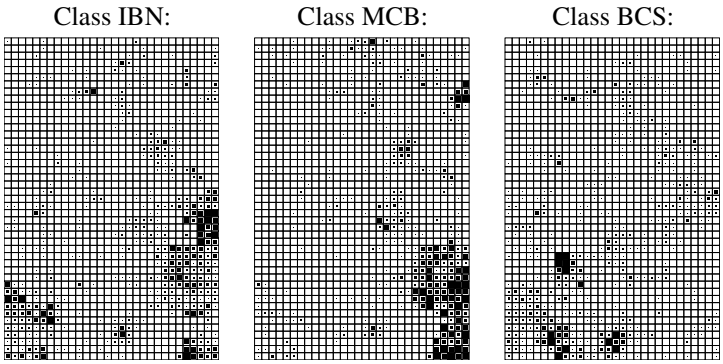

Figure 2: The hit histogram representations of the documents in three Biology-related classes of the NSF corpus. The area of the "hits" on the map is proportional to the number of times those nodes were chosen for the Best Matching Unit of a document. IBN = Division of Integrative Biology and Neuroscience; MCB = Division of Molecular and Cellular Biosciences; BCS = Division of Biological and Critical Systems.

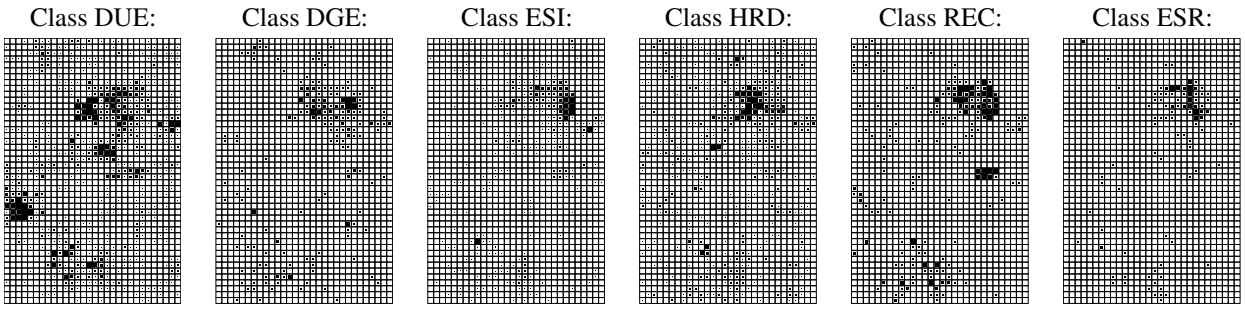

Figure 3: The hit histogram representations of the documents in six Education-related classes of the NSF corpus. The area of the "hits" on the map is proportional to the number of times those nodes were chosen for the Best Matching Unit of a document. DUE = Division of Undergraduate Education; DGE = Division of Graduate Education; ESI = Division of Elementary, Secondary and Informal Education; HRD = Division of Human Resources Development; REC = Division of Research, Evaluation and Communication; ESR = Education System Reform Programs.

class DUE are divided into very separate areas in the map, some restructuring of the NSF classes could be considered based on this analysis. On the other hand, a distributional pattern of the class DUE may also indicate a degree of interdisciplinarity.

5.3 Quantitative Comparison

The proposed term selection process is compared here to a few alternative approaches in the document categorization task. In a good document map the overlap between the documents of different categories is small, assuming that the categories are properly formed. We therefore compare the term selection procedures in their ability to provide maps with a low amount of overlap.

We compare our method to two alternative selection procedures. The first uses only the reference corpus ordering, and the other chooses terms solely based on their frequencies in the corpus. In addition, we test each selection approach with and without weighting the terms using the popular tf-idf weighting scheme [17]. The quality is measured as the average class entropy in SOM nodes, weighted by the number of documents mapped to each node. Small entropy indicates good class separation. The results are computed over ten randomly initialized maps of each type, using maps of size 20×30.

The results (Table 1) indicate that the proposed term selection approach clearly outperforms the naive selection by frequency. Using class entropy in term selection improves the separation of classes, as is expected since the same class information is utilized. However, the difference is small compared to using only the reference corpus, revealing that good class separability is achieved even without using the classes in term selection. We use the entropy, since it drops terms that would be uninteresting for interpretation.

6 Conclusions and Discussion

In this paper, we analyzed the NSF Research Awards abstracts text corpus using the Self-Organizing Map. We de-

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

	Term selection method:		
	with entropy	ref. corp.	freq.
unweighted	2.59 ± 0.02	2.94 ± 0.03	4.02 ± 0.01
tf-idf	2.42 ± 0.03	2.55 ± 0.03	2.73 ± 0.05

Table 1: Average class entropies (in bits; error margin of two standard deviations) of maps trained with the three alternative term selection processes, with and without tf-idf weighting. The class entropy of the whole document collection would be 4.80, and small values indicate success in sensible document organization. Using the reference corpus brings a clear advantage over selecting terms solely based on frequency, and adding the entropy criterion further improves the results. Weighting with tf-idf always seems to be of benefit, but the difference shrinks for better term selection methods.

scribed a process of selecting terms that aims to represent the nature of the particular corpus as well as possible, and that is based on using a reference corpus and the term entropies in document categories to refine the list of terms. Then, we validated the term selection by examining the extracted terms using a SOM.

Finally, we utilized these terms to analyze the documents of the NSF corpus that were handled by different NSF divisions. The results of our preliminary analysis suggest that this kind of analysis could be of use in e.g. examining or refining the divisions of the NSF.

Acknowledgements

This work was supported by the Academy of Finland through the Adaptive Informatics Research Centre that is a part of the Finnish Centre of Excellence Programme, and is also closely related to a project funded by the Academy of Finland in which similar analysis is conducted for Finnish research grant applications.

References

[1] F.J. Damerau. Evaluating domain-oriented multiword terms from texts. *Information Processing and Management*, 29:433–447, 1993.

[2] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.

[3] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[4] T. Honkela, R. Nordfors, and R. Tuuli. Document maps for competence management. In *Proceedings of the Symposium on Professional Practice in AI*, pages 31–39. IFIP, 2004.

[5] T. Honkela, M. Pöllä, M.-S. Paukkeri, I. Nieminen, and J.J. Väyrynen. Terminology extraction based on reference corpora. Technical Report E series (manuscript), Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2007.

[6] T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing map. In *Proceedings of ICANN-95*, volume 2, pages 3–7, Paris, France, 1995. EC2 et Cie.

[7] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proceedings of IJCNN'98*, volume 1, pages 413–418. IEEE Service Center, Piscataway, NJ, 1998.

[8] P. Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished draft, 2002.

[9] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 2001.

[10] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen. SOM_PAK: The Self-Organizing Map program package. Technical Report Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[11] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive text document collection. In *Kohonen Maps*, pages 171–182. Elsevier, Amsterdam, 1999.

[12] K. Lagus, S. Kaski, and T. Kohonen. Mining massive document collections by the WEBSOM method. *Information Sciences*, 163:135–156, 2004.

[13] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of 14th Ann. International ACM/SIGIR Conference on Research & Development in Information Retrieval*, pages 262–269, 1991.

[14] D. Merkl. Structuring software for reuse - the case of self-organizing maps. In *Proceedings of IJCNN'93*, volume III, pages 2468–2471, Piscataway, NJ, 1993. IEEE Service Center.

[15] R. Miikkulainen. Self-organizing feature map model of the lexicon. *Brain and Language*, 59:334–366, 1997.

[16] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254, 1989.

[17] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing and management*, 24(5):513–523, 1988.