Analysis of Interdisciplinary Text Corpora

Matti Pöllä*
Timo Honkela*
*Helsinki University of Technology
Adaptive Informatics Research Centre
P.O. Box 5400 FI-02015 TKK, Finland
{matti.polla, timo.honkela}@tkk.fi

Henrik Bruun†
†University of Turku
Institutions and Social Mechanisms IASM
FI-20014 Turku, Finland
henbru@utu.fi

Ann Russell‡
‡University of Toronto
Knowledge Translation Program at St. Michael's Hospital
russell@smh.toronto.on.ca

Abstract

Computational analysis of natural language is often focused on the syntactic structure of language–often with no regard to the overall context and area of expertise of the analyzed text. In this paper, we present a means of analyzing text documents from various areas of expertise to discover groups of thematically similar texts with no prior information about the topics. The presented results show how a relatively simple keyword analysis combined with a SOM projection can be very descriptive in terms of analyzing the contextual relationships between documents and their authors.

1 Introduction

When reading a text document, humans process information not only by parsing individual sentences and their meanings. Instead, a large collection of prior information about the subject at hand is often necessary–especially in the communication between content experts in highly specialized professions. However, computational processing of natural language is often limited to processing information sentence-by-sentence. While this approach can be sufficient for example in elementary machine translation tasks, a more abstract view about the overall nature of the document would greatly facilitate automated processing of natural language in many applications, such as machine translation and information retrieval. Eventually, this kind of information about the document can be used further for interdisciplinary (within language) translation tasks.

In this paper, a method based on automatically extracted keywords combined with the Self-Organizing Map Kohonen (1995) is presented for text document analysis to visualize the relationships of the documents and their authors. The significance of computational tools for interdisciplinary language analysis has been outlined in Bruun and Laine (2006) referring to previous success in the analysis of text document using the WEBSOM method Honkela et al. (1997).

In section 2 we describe the applied method for keyword selection and in section 3 how the SOM is employed for analysis. In section 4 we present results of an experiment that used corpora from two distinctive areas of expertise. In section 5 some conclusions are made about the applications of the presented method.

2 Selecting descriptive keywords

Our analysis of text documents attempts to extract information about the area of expertise of the document using a set of keywords which are extracted from the documents automatically. To extract relevant keywords for each text document the frequency of each word is examined as an indicator of relevance. As the word frequency distribution for any source of natural language has a Zipf'ian form, special care has to be taken to filter out words, which occur frequently in all documents but are irrelevant in describing the topic of the document (such as 'the', 'a', 'an', 'is' etc.). Figure 1 shows a typical word frequency distribution for a text document in which only a small class of words have a large frequency with the rest of the words having an exponential-like distribution. In Figure 2 the word-use for text documents is visualized in a matrix form. In this figure, each row corresponds to a text document and each column to a specific word with

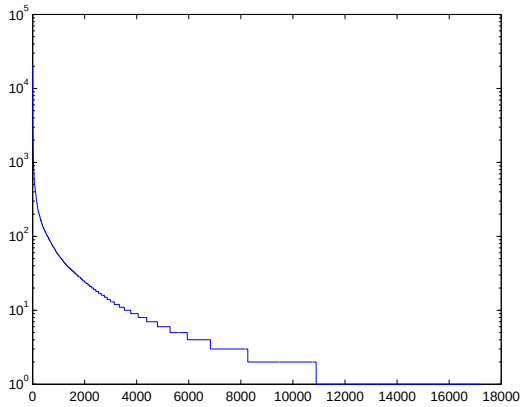


Figure 1: The word-use frequency for natural language resulting in a Zipf'ian distribution.

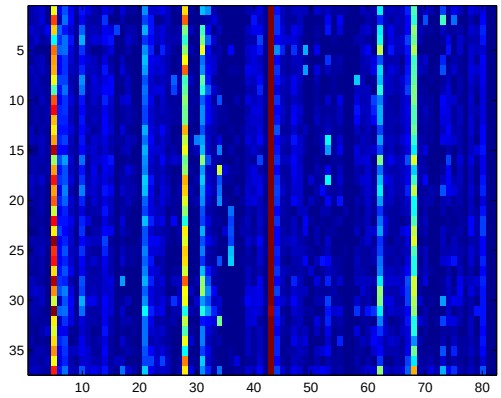


Figure 2: A visualization of the word frequency: each row corresponds to a text document and each column the frequency of a certain word. Note how some of the most common words (which appear in every document) form a light vertical stripe. Other more discipline-specific words can be seen as disconnected light dots.

a light color shade indicating frequently used words. The interesting document-specific frequent words can be seen as individual spots whereas the frequent but irrelevant words can be seen as vertical light stripes.

To distinguish between words that appear frequently in a specific document from the words that are frequent in language in general, a straightforward comparison can be made by relating the frequency f_w^D of the word w in the examined document D and the corresponding frequency f_w^{ref} in another corpus that can be considered more neutral in terms of using less discipline-specific vocabulary Damerau (1993).

However, by comparing the word frequencies directly we would need to normalize the frequencies for each document to get comparable results. For this reason we use the rank of occurrence of each word instead of the frequency itself. Using this method we can compute a relevance score for each word w_i

$$\text{score}(w_i) = \frac{\text{rank}_D(w_i)}{\text{rank}_{\text{ref}}(w_i)} \tag{1}$$

where $\text{rank}_D(w_i)$ is the rank of word w_i in the list of most common words in document D and $\text{rank}_{\text{ref}}(w_i)$ is the corresponding rank in the reference corpus.

3 SOM analysis of keyword occurrence in documents

After applying the keyword extraction method to each document we have gathered an overall set of keywords. We can use this result in a further analysis where the objective is to find clusters of similar documents and analyze the relationships between the documents and their authors by using the Self-Organizing Map algorithm Kohonen (1995).

The SOM algorithm is a widely accepted tool for creating a two-dimensional visual map representation of a high dimensional data set. In this case we use the occurrence of the extracted keywords as a vector representation of the text documents. This way each document D_i will be represented by an n-dimensional vector

$$D = [f_{w_1}, f_{w_2}, f_{w_3}, ..., f_{w_n}]$$

in which $f_{w_j}^d$ is the frequency of keyword j in the document. For simplicity, we used a binary value merely to indicate whether the keyword is present in the document or not.

As a result, documents with similar keyword occurrence profiles will locate close to each other in the SOM projection. Correspondingly documents that do not share similar keyword occurrence profiles will be located further away from each other.

4 Experiments

To test the proposed method, we selected two sets of documents from two distinctive fields of expertise: the first corpus A was collected from scientific articles published in the proceedings of the AKRR'05

conference with the topics of the papers mainly in the fields of computer science, cognitive science, language technology and bioinformatics Honkela et al. (2005); Bounsaythip et al. (2005); Russell et al. (2005). Corpus B consists of a collection of articles published by the Laboratory of Environmental Protection at Helsinki University of Technology with the topics ranging from social sciences to environmental managing. Henceforth these will be referred to as corpus A and corpus B.

Table 1: The two corpora used in the experiments

	Corpus A	Corpus B	Reference
Words	190 107	372 836	92 360 430
Files	45	37	308

At first, some preprocessing was done to the text material. This included removing all punctuation symbols and transforming all symbols to lowercase letters. This was done because the downside of adding noise caused by the loss of information due to removing punctuation was considered smaller than the benefits of preprocessing.

For the keyword extraction phase, the Europarl corpus Koehn (a,b) was selected as a reasonably discipline-neutral reference material. In Figure 3 a sample of the extracted keyword set is shown. The selection method has resulted in a fairly good set of keywords which describe the contents of the documents without being too general.

```
architecture bayesian best better
biological clustering computer denmark
distance distributed effect experiments
expression extraction forestry found
four further growth impact
interdisciplinary interviewees
introduction kola learned makes meaning
mediated morphs necessary objects often
organized panel panels paradigm
parameters part personal points present
press product publications
recommendations rho science
selforganizing shows significance
subject test thought together
understand visualization weights words
xml
```

Figure 3: A sample of the keyword set extracted from corpus A and B.

Finally, the result of the SOM projection of the keyword frequency data is seen in Figure 4. In this figure, the documents of corpora A and B are shown.

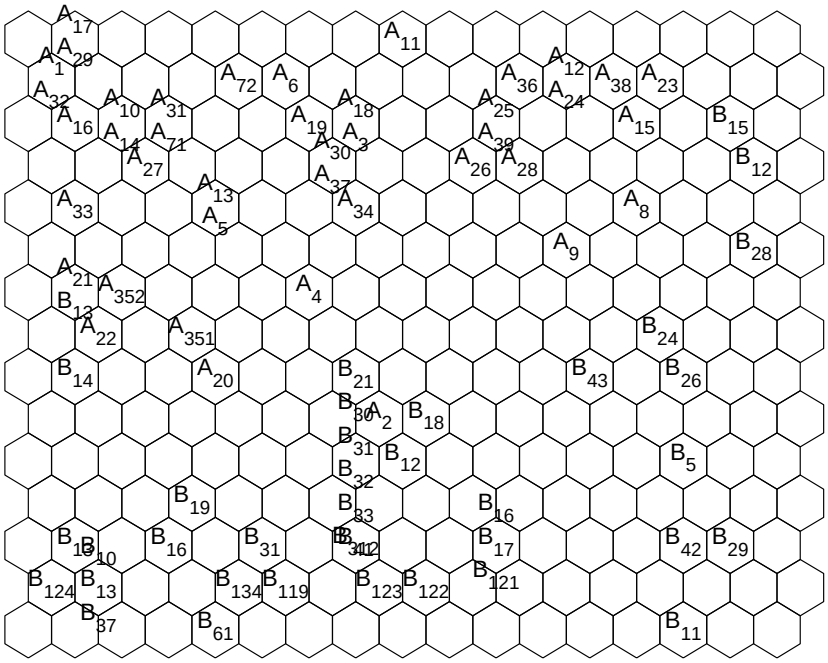
Looking at the distribution of the articles in the U-matrix projection in Figure 4, we can see a clear cluster of papers in corpus A in the top of the map. Most of the map is occupied by the mid-upper part of documents from corpus B. Clusters inside these areas can also bee seen indicating groups of thematically similar documents inside the corpus.

Another result of the experiment is visualized in Figure 5 where the individual component planes of the SOM projection are shown according to several keywords. In this figure a light gray shade corresponds to an area where the keyword in question has been used and a dark area to the areas where it does not occur. By combining this information with the locations of the individual documents in Figure 4 we can see what kind of topics are covered in the documents.

Some of the keywords are common in most of the documents and do not discriminate corpora A and B significantly. For example, the usage of the words 'learning', 'language', 'theory', 'knowledge', and 'complex' is distributed all across the map and are thus used commonly in nearly all documents. This observation is well aligned with our intuition about the overall content of the documents.

By contrast, some of the keywords appear only in a small group of documents are less general. For example the words 'interdisciplinarity', 'ecology', and 'sociobiology' appear almost exclusively in the lower part of the map corresponding to corpus B. Other terms such as 'bayesian', 'algorithm', and 'selforganizing' can be seen mostly in the area occupied by the documents of corpus A.

In addition to grouping documents from different sources, we can see sub-clusters for example in the bottom left corner of the map. In this small area the words 'reindeer', 'herders', and 'forest' occur in the documents B_{11} B_{29} and B_{42} which form a subgroup inside corpus B.



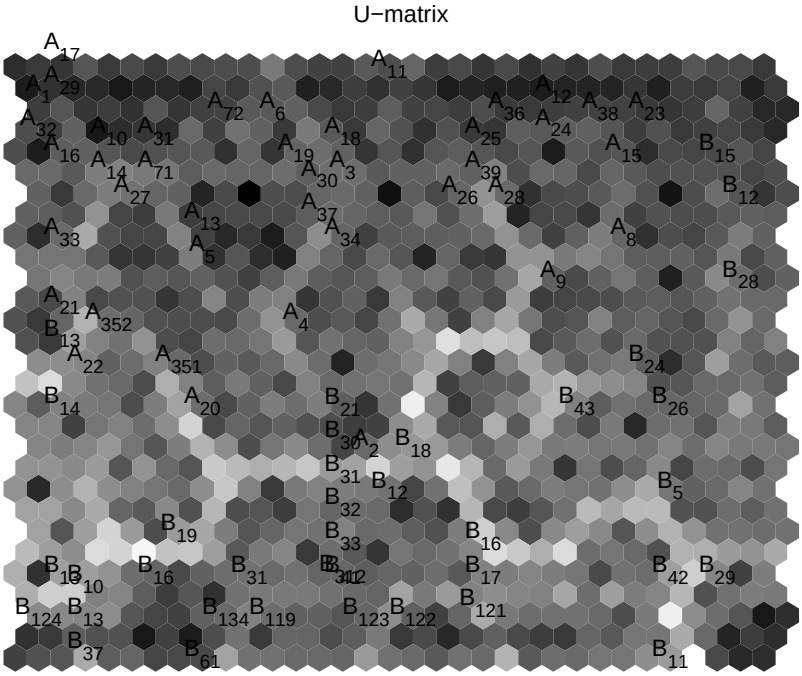


Figure 4: The documents of corpora A and B in a SOM projection (top) and a U-matrix visualization of the document map (bottom). The documents of corpus A have formed a cluster in the upper part of the figure while the lower half of the map is occupied with documents from corpus B. Clusters within the area dominated by documents from corpus B can also be seen (e.g., B_{11} B_{29} and B_{42}).

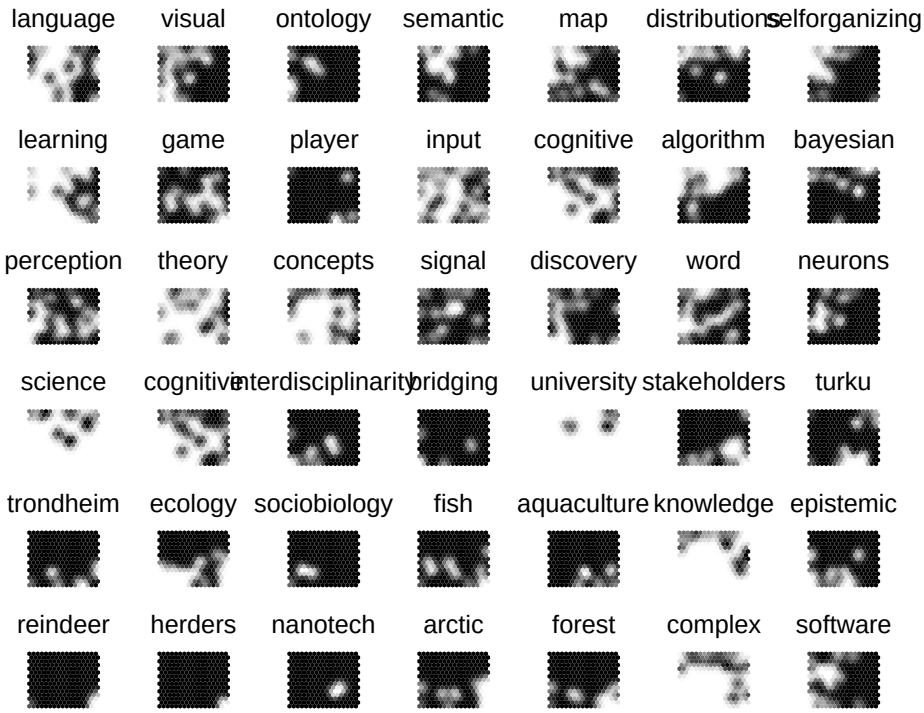


Figure 5: Component plane visualizations of 42 keywords used in the analysis. In the figures, light shades correspond to occurrence of the keyword in the specific region of the map.

5 Discussion

In this paper, a novel method was presented for content analysis of text documents and their authors using automatically extracted keyword selection combined with a Self-Organizing Map analysis.

Our results show that the adopted word frequency analysis can be effective in describing the contents of a document. A SOM analysis of the keyword occurrence was successful in clustering not only documents from different sources but also similar documents in the same corpus.

However, our current efforts have been restricted to isolated single words ignoring the potential of analyzing variable-length phrases. An analysis of keyphrases consisting of several words would evidently be beneficial, since the topics of scientific articles often consist of two-word concepts such as 'neural network' or 'signal processing'.

Despite the limitations of the current implementation, our experiments have shown that a combination of an automatic keyword extraction scheme combined with a clustering algorithm can be effective in describing the mutual similarities of text documents. Our statistical approach to the task has the additional benefit of being independent of the language that is being processed as no prior information about the processed language syntax is encoded into the algorithm.

The possible application areas of the proposed method are diverse. For example, in machine translation the acquired prior information about the context of the information could be used to disambiguate words or acronyms that have different meanings depending on the context. In information retrieval the method could be used to help in navigating through document collections as the search engine could point out keywords that are considered relevant but are omitted in the search. Finally, automated methods that determine similarities and dissimilarities across diverse and vast knowledge bases have the potential to advance scientific discoveries both within and across disciplines.

References

Catherine Bounsaythip, Jaakko Hollmén, Samuel Kaski, and Matej Orešic, editors. *Proceedings of KRBIO'05, International Symposium of the Knowledge Representation in Bioinformatics*. Helsinki University of Technology, Espoo, Finland, June 2005.

Henrik Bruun and Sampsa Laine. Using the self-organizing map for measuring interdisciplinary research. Manuscript, 2006.

Fred Damerau. Generating and evaluating domain-oriented multi-word terms from texts. *Inf. Process. Manage.*, 29(4):433–448, 1993.

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. WEBSOM—self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pages 310–315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

Timo Honkela, Ville Könönen, Matti Pöllä, and Olli Simula, editors. *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Helsinki University of Technology, Espoo, Finland, June 2005.

Philipp Koehn. Europarl parallel corpus, a. http://people.csail.mit.edu/koehn/publications/europarl/.

Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished manuscript, b.

Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

Ann Russell, Timo Honkela, Krista Lagus, and Matti Pöllä, editors. *Proceedings of AMKLC'05, International Symposium on Adaptive Models of Knowledge, Language and Cognition*. Helsinki University of Technology, Espoo, Finland, June 2005.