Learning Document Similarity Using Natural Language Processing

Paola Merlo/James Henderson/Gerold Schneider/Eric Wehrli (Geneva)

Abstract

The recent considerable growth in the amount of easily available on-line text has brought to the foreground the need for large-scale natural language processing tools for text data mining. In this paper we address the problem of organizing documents into meaningful groups according to their content and to visualize a text collection, providing an overview of the range of documents and of their relationships, so that they can be browsed more easily. We use Self-Organizing Maps (SOMs) (Kohonen 1984). Great efficiency challenges arise in creating these maps. We study linguistically-motivated ways of reducing the representation of a document to increase efficiency and ways to disambiguate the words in the documents.

1 Introduction

We live in an information society. Unstructured text represents 80% of the mass of data flowing into information networks, and is becoming the most common data stored on-line. There is urgent need for large-scale NLP tools, to transform this huge mass of data into readily available information. In particular, the problem of extracting novel knowledge out of very large unstructured collections of text documents (text data mining) has attracted a lot of attention. One step towards a solution of this problem is to organize the documents into meaningful groups according to their content and to visualize the collection, providing an overview of the range of documents and of their relationships, so that they can be browsed more easily (Kohonen et al. 2000, Rauber/Merkl 1999). Self-Organizing Maps (SOMs) (Kohonen 1984) are a method for generating a 2-dimensional visual map of a document collection. SOMs produce clusters of documents, which are positioned on the map such that similar clusters are next to each other. These clusters can then be labelled with words describing their most important topics, giving an overview of the major topics covered in the document collection, and of their similarity to each other. The topics of clusters change continuously as one moves across the map, making it easier for a viewer to understand the range of documents in the collection than would be possible with an unstructured list of topics.

For example, figure 3 below is one of the maps produced by our system on the Reuters database. The low left hand corner shows a cluster of documents related to agricultural commodities. The neighbouring clusters contain documents on the related topics of the commodity oil and international trade. Nearby areas of the map cover corporate management,

ISSN 1615-3014

stocks, corporate performance, and international banking. The top of the map is dedicated to very short documents reporting financial data, which in this corpus are characterised by specific format words like "versus" and "DATELINE". The importance of a topical term is indicated by the number displayed to its right. From this map we can see that the corpus is dominated by financial articles, and, for example, that the articles on commodity trading are more similar to those on trade issues than to those on corporate earnings. Clustering the documents facilitates retrieval of the information that the user is looking for, while the spatial organization of the map supports the discovery of unlooked for, but related, pieces of information, like in an ordinary library, where books on similar topics are usually grouped in the same section (Rauber/Merkl 1999).

The main advantages of using this method to create a visualization of the documents is that, compared to other methods, it is computationally feasible and it produces qualitatively better maps. Moreover, SOMs can integrate new incoming documents without recomputing the complete map every time (Kohonen et al. 2000). The main disadvantage in using this method is that, although feasible, it is computationally intensive. SOMs search for the globally optimal map, a very time-consuming process. They achieve good performance, but they take a long time to reach the solution. The large amounts of computing time (weeks) needed to calculate these maps detracts from their usefulness. One of the main factors affecting the efficiency of the algorithm is the size of the document representation.

Typically, SOMs, as well as other text processing and information retrieval applications, represents documents simply as the set of words that occur in the document itself. In our work, we explore linguistically motivated ways of reducing this simple representation, by selecting only the most prominent words in the document, based on their syntactic position. The aim of this selection process is to represent a document only using the most important words, thus reducing its size with only little loss of information. We find that methods based on the semantic prominence of a word in a constituent yield some improvements in efficiency, while other methods degrade the quality of the maps too much. We also try combining these linguistically based methods to a technique that reduces documents by using only the high frequency words in the documents. Combining these two approaches yields satisfactory results. Another problem that affects the quality of the map is the ambiguity of certain words. *Bank* is an ambiguous word that can mean river board or financial institution. We study the impact of word sense disambiguation techniques on the quality of the clusters produced in SOMs.

2 NLP for Document Representation

Using NLP techniques in document processing has been explored before with mixed results (Lewis/Sparck-Jones 1996). In this work, we restrict the use of NLP techniques to simple, robust tasks that can be performed with existing tools and techniques. Our main aim is to reduce the dimensionality of the document.

2.1 Identifying Important Words with Syntactic Analysis

The motivation behind using NLP techniques to select informative words in a text, is that the importance of a word token depends both on its type and on the specific linguistic context in which it appears. Syntactic analysis is a computationally efficient first step to identify which words bear contentful information in the document, under the assumption that there is a regular mapping between the content of a text and its syntactic structure. Because of current NLP technology's limitations, we choose to use those parts of a syntactic analysis that can be performed accurately on a large scale. Therefore, we tag the words, extract heads of phrases, as the identification of phrases is accurate, and identify subjects and objects, a task that can take advantage of the rather fixed word order of English, especially for subjects.

A word out of context does not provide information about its salience in the text. The usual practice to take only content words into account is a form of selection based on salience with respect to the inventory of classes of words: nouns, verbs, adverbs, adjectives and prepositions are salient classes while other classes, such as determiners, are not. However, the semantic and pragmatic prominence of a word is primarily determined in relation to its syntactic salience in a sentence. Words of content classes may appear in more or less salient positions. For example, in the expression *pocket computer*, computer is more salient than *pocket*, since a pocket computer is a kind of computer. Generally speaking, linguistic heads appear in semantically more prominent positions than non-head words. We therefore propose to use a document representation consisting of the linguistic heads of the content words. In order to reduce the dimensionality of the document representation without losing too much content, we thus individuate the head of the principal content phrases. Specifically, we consider the heads of NPs and VPs, in the way they are expressed by the syntactic structure. In the case of proper nouns, all its component words equally contribute to the meaning. For example, whereas *new company* is a type of *company*, *New York* is not a type of *York*; similarly *British* in *British Petrol* constitutes an as important part of the proper noun as the head and is thus included in the document representation.

Among nominal expressions, a further distinction based on grammatical functions can be made: subjects typically express more salient concepts than their object counterparts. Adjuncts, which are nouns which are not logically required by a verb, typically express less relevant circumstances, they are typically even less salient. The salience hierarchy from subject to object to adjuncts has been proposed by Keenan/Comrie (1977) and is used for example by practical anaphora resolution algorithms (Lappin/Leass 1994) and for automatic text summarization (Boguraev/Kennedy 1997).

2.2 Word Sense Disambiguation for Document Classification

Another linguistically-based manipulation of the text that can lead to better clustering and visualization of documents is disambiguation of those words that have several meanings.

At the level of the meaning of words, two main problems must be faced: synonymy (two words that share a very similar meaning) and polysemy (a single word that has several meanings or senses). If synonymy is not recognized, two related documents might not be

clustered together, while if polysemy is not handled, two unrelated documents might erroneously be placed in the same cluster because they share a word. Humans know whether two words are similar or whether a word can have two senses because of their extensive knowledge of the world. In a text classification task, the most obvious source of information about world knowledge are the texts to be classified. Therefore, we use the text itself to define the meaning of words.

3 Natural Language Processing Tools

The two main NLP tasks described above, syntactic analysis and word sense disambiguation, were performed on a large scale on the Reuters document collection by existing tools, or implementation of existing techniques.

3.1 The Underlying Parsing System

Recognizing heads of phrases and subjects and objects benefits from full-fledged syntactic analysis of the sentences in the text. We perform this task using the in-house Fips system (Wehrli 1997).

The Fips parsing system is inspired by the Principles and Parameters framework (Chomsky 1986), which posits existence of abstract, language-independent principles, and of parameters whose values vary according to a specific language. In its implementation, the grammar of each language includes components corresponding to a particular process. Some components generate structure such as the lexical projection process which produces the elementary building blocks based on lexical information, the long-distance chain composition process, used for questions and relative clauses, or the coordination process. Others have a filtering function on these structures, such as the checking of morphological cases or of the valency of the verb.

Algorithmically, the Fips parser is a bottom-up tabular parser, which pursues all alternatives in parallel. It does not use a grammar of context-free rules to build the structure, but rather it proceeds as a licensing parser (Abney 1989), where only three types of actions are supported: projection, attachment to the left (specifiers) and attachment to the right (complement). The parser does not distinguish between complements and modifiers. The advantage of using a licensing technique is that incoming tokens are immediately incorporated in the structure. Alternatives are considered concurrently, and a small number of heuristics are used to restrict the hypothesis set.

From the point of view of parsing, the projection operation turns a lexical entry into the head of a little subtree of a precisely defined form, which can then combine with the already constructed portion of the tree according to the two operations of specifier and complement attachment. In a specifier attachment, the incoming node takes a complete constituent to the left as its subconstituent. In a complement attachment, the incoming node attaches to the right of the active nodes of the existing structure. Figure 1 shows the kind of phrase structure representations which the parser builds. These representations make notions such as subject and

object explicit: the subject is the noun phrase (NP) immediately dominated by the sentence (TP), while the object is the noun phrase immediately dominated by the verb phrase (VP).

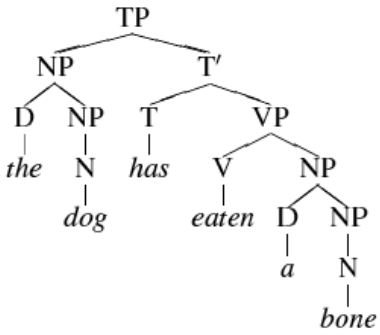


Figure 1: Example of tree constructed by Fips for the sentence *The dog has eaten a bone*

The portion of the Fips parser devoted to parsing English uses a lexical database exceeding 60,000 entries and has a broad grammatical coverage including simple and complex sentences, complex determiners and possessives, yes/no and *wh*-interrogatives, relatives, passive, some comparatives as well as most cases of coordination (excluding gapping). This is a very robust parsing system which was able to parse the Reuters database completely, after limited tuning.

3.2 The Word Sense Disambiguation System

Our word sense disambiguation system is a direct implementation of the context-based method proposed by Schütze (1997). The approach is as follows. A word meaning is represented by its contexts, that is by the set of words which tend to co-occur with it in a given window of words. As the window of words we use the entire document. For example, the word *bank* is ambiguous because it can mean river board or financial institution. These two meanings can often be distinguished by the token's context, as a river board's typical context words are *river*, *water*, *boating*, while a financial institute's context words are *stocks*, *exchange*, *cash*.

These representations of the words which tend to co-occur with a given word are called word vectors. These word vectors partially solve a semantic word categorization problem: words with similar distributional semantics will have similar contexts and therefore get similar word vectors. Synonyms usually do not occur together, but they occur in the context of the same words, as the exchangeability criterion suggests. While the use of word vectors allows us to detect synonyms, sense distinctions of a word are not captured in this representation.

For a token-wise sense distinction, the words in the context window of a target token do not always contain enough information to take a reliable disambiguation decision. First, none of the typical context words may be present for a given token. Second, some context words may not be good disambiguators: *flow*, in the context of *bank*, may express flowing water as well as cash flow. Chances that a large number of context words in cooperation still show these problems are low, however. It is thus desirable to have as many context words as possible. One method to obtain more context words is to extend the context of a target token by using the contexts of the context words themselves (Schütze 1997). We thus compute a representation of the context for each individual token as follows: instead of the *words*, the *word*

vectors of the words in the context of the target token are summed to obtain a representation of the target token's context. These representations are called the context vectors.

We use these context vectors to handle sense distinctions by clustering the different contexts of a polysemous word into the word's different senses. After computing the context vectors for each token of the word, we apply the k-means clustering algorithm to this set of contexts. If two word tokens have the same word sense, then they should have similar contexts, and thereby will end up in the same cluster. The centroid of this cluster can be thought of as the prototypical context for word tokens which have that sense. Thus this centroid is called a sense vector. A word token is assigned to the word sense whose sense vector is closest to the token's context vector.

This algorithm does not attempt to discover dictionary-defined word senses, but it discriminates different uses of a word according to its distribution. Some of the relations found express corpus-specific relations, others mirror true world-knowledge sense relations. Table 1 shows in each row the 4 senses whose distributions are closest to the word sense *S*, where the word senses *S* are ordered alphabetically.

WORD SENSE	RELATED WORDS			
air	service	aircraft	damage	schedule
air	fleet	newly	report	Charles
aircraft	air	Boeing	contract	North
aircraft	fly	group	identify	final
airline	flight	route	comment	continental
airline	comparable	Ago	comparison	year-earlier
airport	arm	strategy	operate	separate
airport	Friday	attempt	dispute	remove
allegation	status	departure	independent	remove
allegation	investigation	violate	man	comment
allege	violation	court	count	investigate
allege	computer	comply	charge	General
allowance	fee	Price	Dennis	gas
allowance	current	result	fiscal	industry

Table 1: Some disambiguated words and their closest senses of other disambiguated words

4 Visualization Tools

Our use of NLP techniques for document processing aims to reduce the document sizes to improve the speed of calculating visualizations of document collections. Grouping the documents and organizing them in an easy-to-use interface facilitates retrieval of the known information, and supports the discovery of novel, but related, pieces of information and knowledge, like in an ordinary library, where one might stumble upon an unknown but highly relevant book (Rauber/Merkl 1999). Our system visualizes the documents on interactive maps. The documents are organized on a grid based on their topic. A web-based interface allows the

user to click on the map and access the document. We describe here the different components of the system and the computational steps to build the maps.

4.1 Computing the Document Vectors

As is standard in Information Retrieval (Salton/Buckley 1988), each document is represented by a vector which specifies how many times each word occurs in the document (the word frequencies). These counts are weighted to reflect the importance of each word. The weighting is the inverse of the log of the number of documents each word occurs in (the inverse document frequency). This vector of weighted counts is called a "bag of words" representation. Words from a specific list of "stop words" (such as function words) are not included in the representation. Also, words which occur in three or fewer documents are removed from the document representation, because they are too infrequent to have any impact on the results of the SOM algorithm, and removing them greatly reduces the total number of different words (by 70% in the baseline model).

4.2 Building and Visualizing the Self Organizing Maps

Given a set of document representation vectors, the SOM algorithm finds a partitioning of those documents into clusters and an assignment of these clusters to positions on a 2-dimensional grid. The range of documents in the collection can then be visualized by displaying each cluster's topic at the cluster's position on a 2-dimensional map, as illustrated in Figures 2-3.

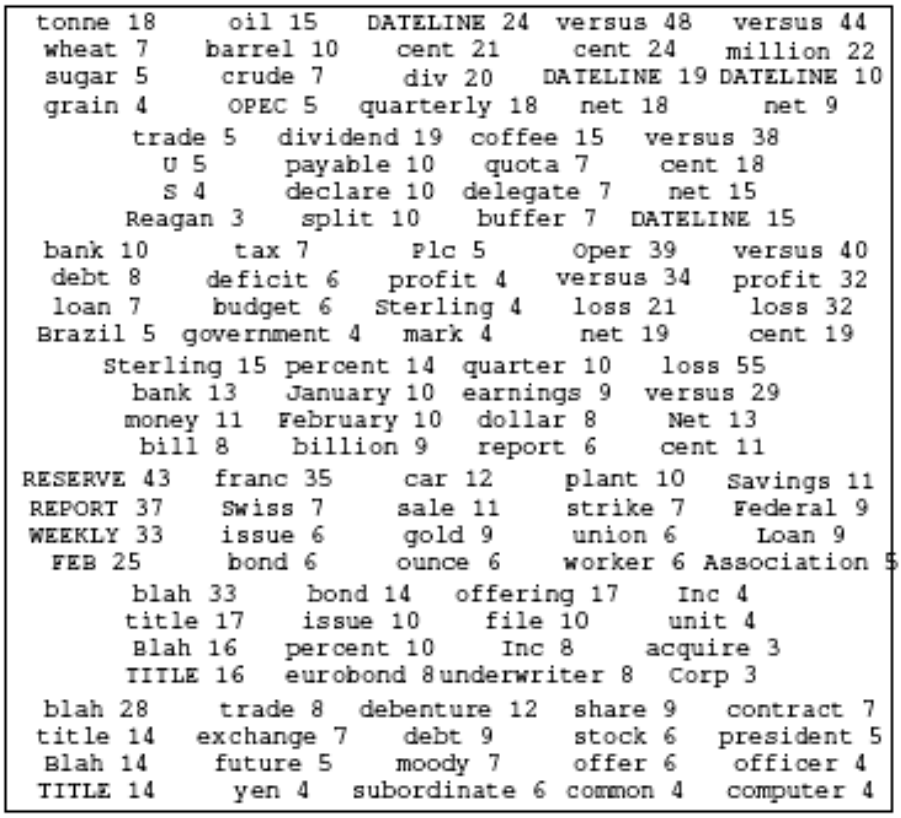


Figure 2: Labelled map for baseline Model.

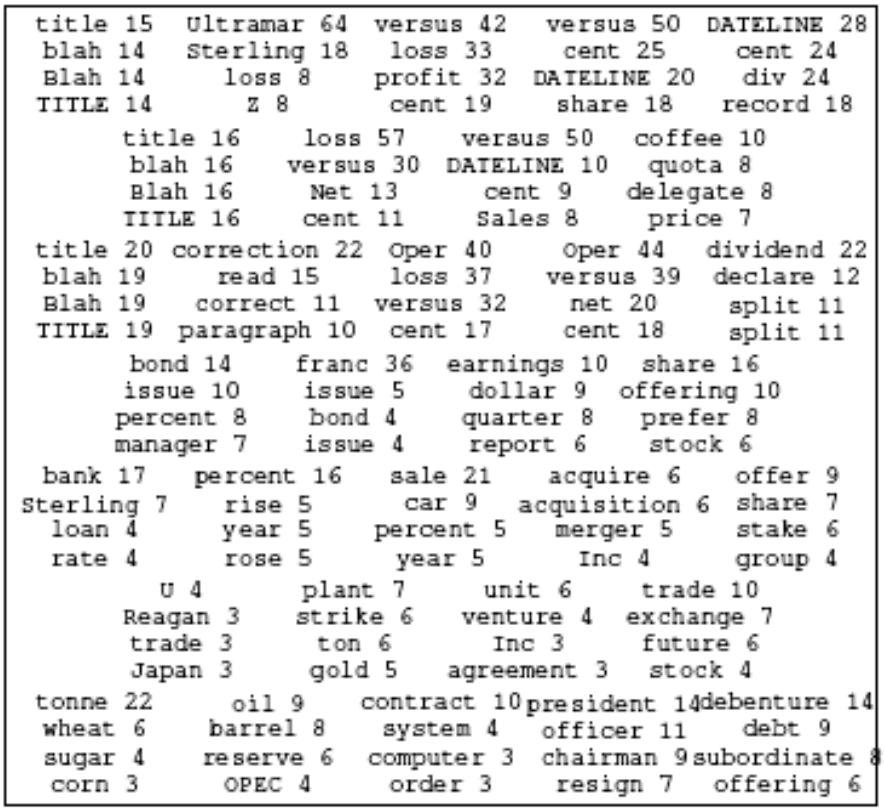


Figure 3: Labelled map for Model 1.

The algorithm searches the space of clustering and the space of position assignments simultaneously, trying to find a global optimum for two criteria. The first criterion is that the documents within a given cluster are similar to each other. This property means that each cluster has a coherent topic. The second criterion is that clusters which have positions next to each other on the map (called "neighbours") have similar documents. This property means that the topics of clusters change continuously as one moves across the map, making it easier for a viewer to understand the range of documents in the collection than would be possible with an unstructured list of topics.

The SOM's 2-dimensional grid of map positions lends itself naturally to a visual display, each cluster being assigned a position on the display according to its position in the grid. To summarize the topics of the documents in a cluster, we display a short list of the most important terms for characterizing that cluster, as illustrated in Figures 2 and 3. The importance of a term (indicated by the number displayed to its right) reflects how influential the term is in determining what documents are assigned to that cluster and that region of the map.

4.3 System's Interface

In addition to producing a labelled map, the resulting SOMs have been incorporated into a web-based interface which allows interactive exploration of the document collection.[1] The top level of this interface is the labelled map, shown in figure 4. From here the user can access each individual cluster, as shown in figure 5.

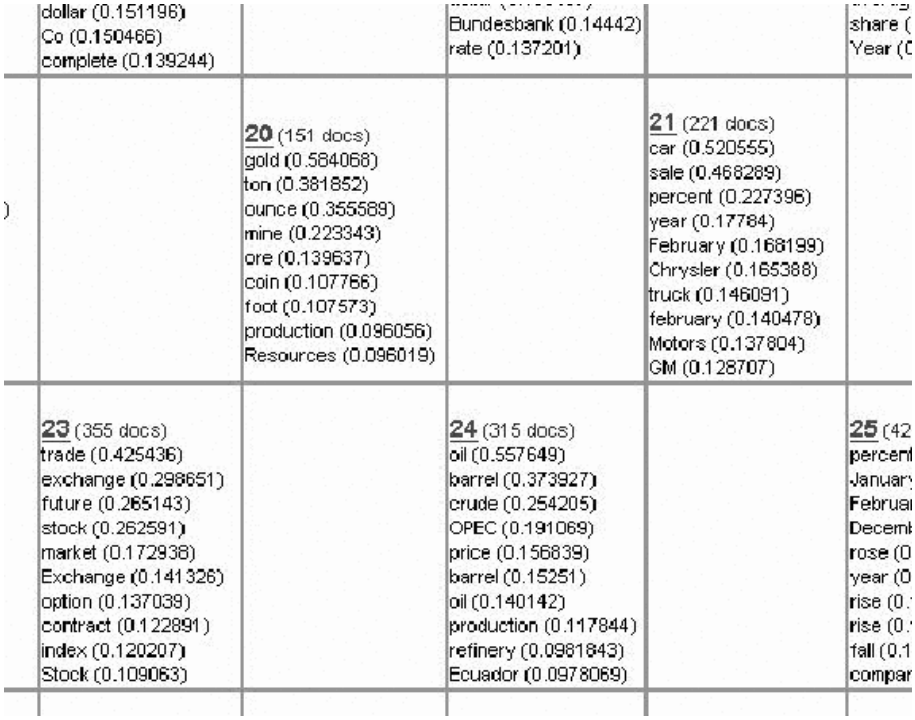
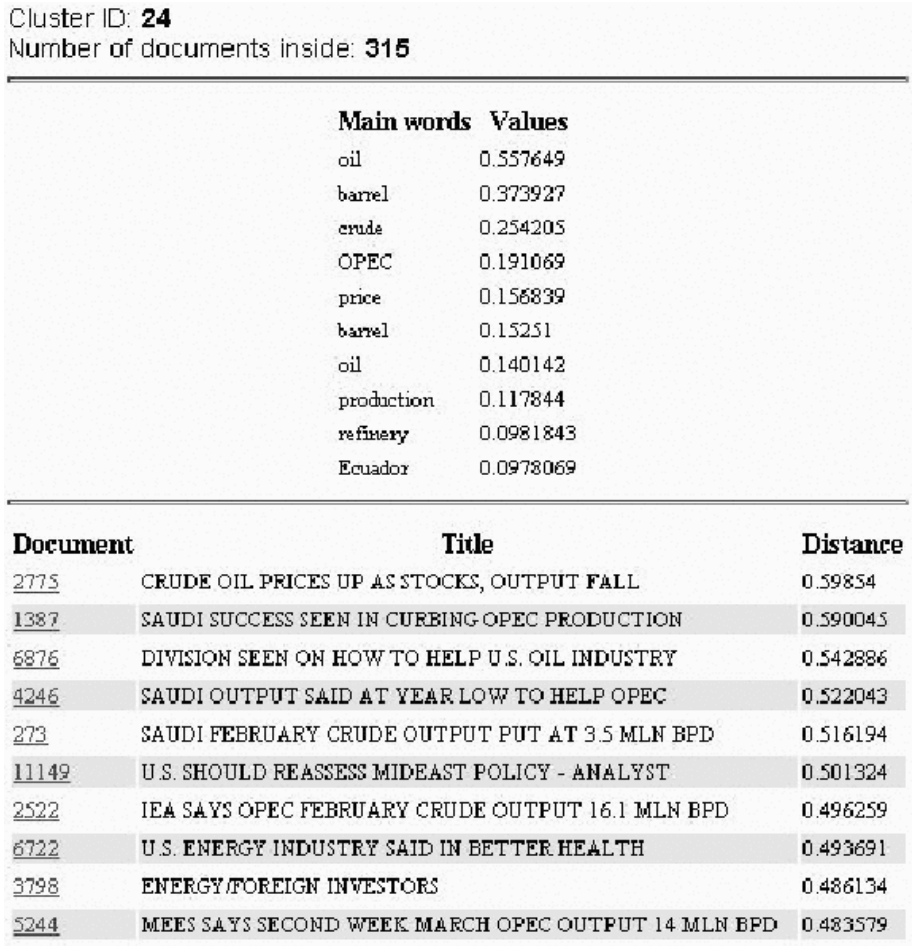


Figure 4: Interface to the map of document clusters

[1] This interface can be accessed via the internet through the web page http://129.194.71.202/~datamining/textmining/index.php. When asked for a project, enter "bow".



Cluster ID: **24**
Number of documents inside: **315**

Main words	Values
oil	0.557649
barrel	0.373927
crude	0.254205
OPEC	0.191069
price	0.156839
barrel	0.15251
oil	0.140142
production	0.117844
refinery	0.0981843
Ecuador	0.0978069

Document	Title	Distance
2775	CRUDE OIL PRICES UP AS STOCKS, OUTPUT FALL	0.59854
1387	SAUDI SUCCESS SEEN IN CURBING OPEC PRODUCTION	0.590045
6876	DIVISION SEEN ON HOW TO HELP U.S. OIL INDUSTRY	0.542886
4246	SAUDI OUTPUT SAID AT YEAR LOW TO HELP OPEC	0.522043
273	SAUDI FEBRUARY CRUDE OUTPUT PUT AT 3.5 MLN BPD	0.516194
11149	U.S. SHOULD REASSESS MIDEAST POLICY - ANALYST	0.501324
2522	IEA SAYS OPEC FEBRUARY CRUDE OUTPUT 16.1 MLN BPD	0.496259
6722	U.S. ENERGY INDUSTRY SAID IN BETTER HEALTH	0.493691
3798	ENERGY/FOREIGN INVESTORS	0.486134
5244	MEES SAYS SECOND WEEK MARCH OPEC OUTPUT 14 MLN BPD	0.483579

Figure 5: Interface to one of the document clusters

In addition to the most important words for the cluster, all the documents for the cluster are listed, ranked from the most typical for that cluster to the least typical. From this list the user can access the texts of individual documents, as well as a ranked list of the document's words. Each cluster's page also provides a list of the other clusters which are the most similar to it. These similar clusters are usually the neighbours on the map, but this list can expose new relationships between clusters which could not be captured in the map due to the requirement that it be 2-dimensional. By exploring the different clusters and reading some of the typical documents in them, the user can quickly get an understanding of the nature of the collection and the range of topics it covers, without the need to look at the total set of documents or a large subset of them.

5 Experiments

In order to measure the impact of the different document representations we use in creating the SOMs, we ran a set of controlled experiments. In each experiment we use a different model of document representation. Each successive model is increasingly reduced, as illustrated in Table 2. We measure the loss in map quality of each reduction. We also experiment with word sense disambiguation and its impact on the quality of the maps.

	the	US	president	delivered	a	long	speech	yesterday
baseline		US	president	deliver		long	speech	yesterday
Model 1		US	president	deliver			speech	yesterday
Model 2		US	president					
Model 3		US US_1	president president_2	deliver deliver_2		long long_1	speech speech_1	yesterday yesterday_2

Table 2: Sample document representations of the 4 models.

5.1 Corpus and Tools

Our data collection consists of the training portion of the Lewis Split of the Reuters-21578 database, for a total of 13,625 documents, varying from one sentence to several pages in length (Lewis 1997). Reuters newswire articles usually report on one coherent event, their size is often very short and never exceeds a couple of pages. As discussed in section 3.1, the syntactic analysis was performed using the Fips system, a large-scale grammar-based parser that outputs very richly annotated structures (Wehrli 1997). We use only a small portion of this annotation in our document representation models. The word sense disambiguation was performed using Schütze's method (Schütze 1997), as discussed in section 3.2.

5.2 The Document Representations

The **baseline** model is a tagged lemmatized bag of words representation (a bag is a set where repetitions are allowed). It utilizes the part of speech tags output by the parser to disambiguate word senses that can be detected by POS tag alone. A small hand evaluation over 882 words has revealed a tagging error of 6.3%.

Model 1 is based on the full syntactic analysis of the text produced by the Fips system. Model 1 reduces the document representation because only nouns and verbs that are heads of phrases are kept, while functional words and modifiers and words that are not heads are discarded. We expect this representation to still capture the denotational and predicative content of the document, but to be considerably smaller in size, because the descriptive and qualitative aspects of it are discarded. Specifically, we extract the head of all NPs and VPs in the document. Proper nouns are treated as multi-head phrases: we keep all their component words, as they all equally contribute to the meaning of the phrase. Fips hypothesizes proper nouns based on lexical information and on orthography and filters out many incorrect hypotheses while parsing. A small hand evaluation on 721 heads (4 articles) yields 94.3% precision and 98.1% recall for this step and 94% precision and 87.8% recall for recognition of proper nouns, on a sample of 100 items.

Model 2 is also based on a full parse. In a structure-based syntactic analysis, different grammatical functions are defined by structural positions. The subject is the nominal phrase attached directly under the main sentential node, while objects occur directly inside the verb phrase, as a sister to the verb. Since proper nouns have been found to be particularly decisive topic indicators (Strzalkowski/Marinescu 1995) we have again decided to include them disregarding their grammatical function. A small hand evaluation on 101 reported subjects (12

articles) yields 51.4% precision and 62% recall. For 92 reported objects, it yields 47.8% precision and 53% recall.

We have explored increasingly drastic reductions to the set of words used in the representation. These models are motivated by a salience hierarchy based on grammatical function (Keenan/Comrie 1977), which has been used successfully before for text summarization (Boguraev/Kennedy 1997). According to this hierarchy, subjects are more salient than objects, which are more salient than other noun phrases. We have tried a model which differs from Model 1 in that nouns which are not in either subject or object position are not included in the document representation, and another model which reduces the document representation further by also removing verbs. In this paper we report results for the most severe reduction, Model 2, which represents documents as a bag of noun heads in subject positions. We found that the results for other intermediate models simply formed a continuum between Model 1 and Model 2, so we do not report them here. They are described in detail in Henderson et al. (2002).

From a linguistic point of view, our work is similar to Hatzivassiloglou/Gravano/Maganti (2000), who explore the use of noun phrase heads and proper names to enrich the feature set input to a hierarchical clustering algorithm. They *add* these features to the bag-of-words document representation, with the expectation that it will facilitate the algorithm in finding relevant terms. Their results are mixed: they find that the additional features improve overall clustering performance if used in combination with the initial words, but they also find an unexpected negative correlation between the head nouns and the topic clusters, which requires further investigation. Most other uses of NLP techniques in document processing and in particular in information retrieval, have aimed at enriching the document representation or the set of indexing terms, with mixed results (Lewis/Sparck-Jones 1996, Strzalkowski 1999). Differently from these pieces of work, we pursue here an application more aimed at visualizing documents than at ranking them, where NLP is used to *reduce* the complexity of the representation of the document, and to focus only on the important words for efficiency reasons. Therefore, we do not enrich the baseline representation, but we substitute it with more compressed models.

Model 3 is built to investigate the impact of word sense disambiguation. Model 3 is an augmentation of the bag-of-words representation to which the disambiguated words have been added. We have also tried a model where the disambiguated words replace their ambiguous counterparts, but we found that removing the original words resulted in a loss of information due to the fact that the majority of senses of polysemous words are semantically related. We follow the hypothesis that words are used in only one sense within the same discourse unit (Yarowsky 1995). We take a discourse unit to be a document, because Reuters newswire articles usually report on one coherent event. Therefore, an individual occurrence of a word is defined by the other words in the same document, its sense context. From a practical point of view, in these experiments, we only disambiguate words which occur in at least 30 documents, and in each case we assume that there are two senses to the word (this means that we run the k-means clustering, described below, with two clusters).

5.3 Experimental Evaluations

To measure the effects of the reduced representation models on the SOM algorithm, we trained SOMs on the above document representations and evaluated both their training efficiency and the quality of the resulting maps.

5.3.1 Efficiency Comparisons

The objective of the first two models is to increase the speed of SOM training. To estimate the effects on computation time of these models, we used a timing program to run the SOM implementation on each model for ten iterations. As shown in Table 3, the models result in significant speed-ups over the baseline model, particularly considering the long computation times involved. These increases in speed are directly proportional to the reduction in document representation size, as indicated by the number of word types (column labelled Number of Words) and the number of times a word occurs in at least one document (Nonzero Values).

	Timing and Complexity (% reduction)		
	Sec/Iteration	Number of Words	Nonzero Values
baseline	59.338	11450	510586
Model 1	48.032(19.1)	9413(17.8)	401276(21.4)
Model 2	22.634(61.9)	5526(51.7)	139666(72.6)

Table 3: Comparison of the models.

5.3.2 Quality Comparisons

Measuring the effect of our changes to the document representation on the quality of the maps produced by the SOM algorithm is a difficult task. The SOM algorithm is an *unsupervised* learning algorithm. This means that we are not provided with the correct answers which the algorithm tries to learn. A consequence of this is that there is no gold-standard against which to compare the results. Since we are primarily concerned with achieving a reduction in the document representation, without degrading the quality of the map, our assumption will be that the best map is obtained by the richest representation, that is our baseline model, and we will compare the other maps to this one.

First, we observe the similarity of the two maps produced by the reduced models compared to the baseline map. We see that the quality of the Model 1 map (Figure 3) is not degraded, as indicated by the fact that almost all clusters in Model 1 have a correspondence in the baseline map (Figure 2). For example, the upper left corner of the baseline map is almost identical to the lower left corner of the Model 1 map. Moreover, the labels suggest that the Model 1 clusters are fairly coherent. On the contrary, the map produced by Model 2 is not as similar to the baseline (with about a third of the clusters not having an obvious match in the baseline map). The coherence of their clusters is also slightly worse.

Second, we calculate several quantitative indices of the quality of the map, reported in Table 4. The first results column (WCS) indicates the quality of the individual clusters, measuring the average similarity between the centre of a cluster and each of its documents. As can be

seen, Model 1 does not decrease in quality compared to the baseline, while there is a degradation for Model 2.

	Measures of Quality		
	WCS	BNS	RTR
baseline	0.342	0.305	72.6
Model 1	0.339	0.326	74.0
Model 2	0.308	0.384	60.0

Table 4: Comparison of the models. (WCS: Within Cluster Similarity, BNS: Between Neighbour Similarity, RTR: Reuters Topic Recall.)

The second column of quality measures (BNS) reflects the quality of the positioning of clusters on the map, measuring the average similarity between neighbouring clusters on the map. This measure shows no clear trend across the four models, but all the reduced representations do better than the baseline.

The values shown in the third column of quality measures (RTR) compare our clustering to the original labels of topic in the Reuters collection. The Reuters corpus comes with a set of predefined topic labels. While it cannot be expected that an unsupervised clustering method would discover such predefined topics, these topics do give us an indication of which documents are considered similar by human judges. Two documents are judged similar if they are both assigned the same Reuters topic. A good map is one which places similar documents close to each other, preferably assigning them to the same cluster. The RTR measure indicates the extent to which the model's map has this property for the notion of similarity defined by the Reuters topics. Model 1 performs better than the baseline, while there is a degradation for Model 2.

5.4 Impact of Sense Disambiguation

Table 5 shows the evaluation measures applied to Model 3 and the baseline. Model 3 shows some improvement in the quality of the placement of clusters on the map (BNS), with almost no degradation in the quality of the clusters themselves (WCS). This may be because documents which include words that have not been disambiguated, shown in the baseline, tend to be outliers, not fitting in any specific cluster. The baseline might find clusters which are close to these outliers, and thus are farther apart from each other. In contrast, in Model 3 documents have had their words disambiguated, so there are fewer outlier documents and the clusters can be placed closer to their neighbours on the map. However, this improvement is rather small. Model 3 also shows some improvement in agreement with the Reuters topic labels (RTR).

	Measures of Quality		
	WCS	BNS	RTR
baseline	0.342	0.305	72.6
Model 3	0.335	0.378	74.3

Table 5: Comparison of the baseline and WSD models. (WCS: Within Cluster Similarity, BNS: Between Neighbour Similarity, RTR: Reuters Topic Recall.)

5.5 Comparison with Term Selection Methods

The results of the previous section show that the size of the document representation can be reduced without harming the map quality by selecting only the heads of major phrases (Model 1). In this section, we compare this method for selecting word tokens to a methods for selecting word types. The alternative model is the same as the baseline model except terms which occur in 42 or fewer documents are removed. This frequency threshold was chosen because it produces a document representation with the same number of nonzero values as Model 1, as shown in table 6. The frequency based model is faster than Model 1, due to its fewer terms. As can be seen in table 6, its map quality is equivalent to that of the baseline model, and it is also equivalent to Model 1, except for a slight reduction in the quality of the placement of clusters on the map.

	% Reduction from Baseline			Measures of Quality		
	Seconds	Terms	Values	WCS	BNS	RTR
baseline	-	-	-	0.342	0.305	72.6
Model 1	19.1	17.8	21.4	0.339	0.326	74.0
frequency	37.0	81.8	21.4	0.340	0.304	73.5
combined	48.4	84.5	38.1	0.340	0.304	69.7

Table 6: Comparison of Model 1, frequency based term selection and a combination of the two models. (WCS: Within Cluster Similarity, BNS: Between Neighbour Similarity, RTR: Reuters Topic Recall.)

These two methods for reducing the document representation size are very different, and yet they result in roughly equivalent performance of the SOM algorithm. It is thus natural to consider combining them. We derived a new model by taking Model 1 and removing all those terms which were not included in the frequency-based model. This resulted in a much smaller document representation, and a computation time which is almost half compared to those of the baseline model, as indicated in the first panel of table 6. The quality of the maps produced from this model is also equivalent to the baseline, except for some reduction in the correspondence between the clusters found and those defined by the Reuters topics. This indicates that the combination of term selection methods with linguistically-based word token selection methods is an interesting direction for future investigation.

6 Conclusions

These experiments show that we can achieve a significant increase in efficiency in visualizing text collections, without degradation of the maps, by representing documents with the heads of the more important parts of speech (Model 1). This confirms our initial intuition that denotational and predicative information is sufficient to characterize a document. The comparison with a frequency based model shows that the results of linguistically-based token reduction are equivalent to a drastic document frequency cut-offs and that a combination of token and term reduction methods yields very promising initial results. On the other hand, the degradation observed in the model that focuses only on salient words (Model 2) indicate that the reduction in these models is too drastic.

Unlike the first two models, the objective in using word sense disambiguation in Model 3 was to improve the quality of the maps over those of the baseline system. This is a very challenging task, particularly given the difficulty of quantitatively evaluating any improvement achieved. We were not able to demonstrate a significant overall improvement using word sense disambiguation. This may be largely due to the disambiguation method used, which is based on distributional information about word co-occurrences. The SOM is already using this type of information when it decides how to cluster documents. The information added by word sense disambiguation may not be sufficiently different from the information the SOM is already using to make a significant difference to its results. However, future work may show a benefit with other document collections, or with other aspects of the task, such as labelling the clusters.

The results from Model 1 and its combination with frequency based term selection demonstrate that Natural Language Processing can play a useful role in training Self-Organizing Maps. These maps provide an overview of the range of topics covered by a document collection, and allow more effective browsing of the documents. They provide an important tool for text data mining applications. In current work, we are using NLP to fundamentally change the nature of the document representations, so that they capture more of the complete meaning of a text and not just its topic. This change will lead to new types of text data mining tools, such as multi-document text summarization systems. By helping make such systems both efficient and accurate, Natural Language Processing will play a crucial role in transforming the huge amount of available text data into readily available information.

Acknowledgments

This research was supported by the Swiss NSF, grant 21-59416.99. Thanks to our colleagues, Abderrahim Labbi, Christian Pellegrini, and Ivan Petroff.

References

Abney, Steven (1989): "A computational model of human parsing". *Journal of Psycholinguistic Research*, 18: 129-144.

Boguraev, Branimir/Kennedy, Christopher (1997): "Salience-based content characterisation of text documents". In: Mani, Inderjeet/Maybury, Mark (eds.): *Proceedings of the ACL'97/ EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain*. Somerset, NJ: 2-9.

Chomsky, Noam (1986): *Knowledge of Language: Its Nature, Origin, and Use*. New York.

Hatzivassiloglou, Vasileios/Gravano, Luis/Maganti, Ankineedu (2000): "An investigation of linguistic features and clustering algorithms for topical document clustering". *SIGIR 2000*: 224-231.

Henderson, James/Merlo, Paola/Petroff, Ivan/Schneider, Gerold (2002): "Using syntactic analysis to increase efficiency in visualising text collections". In: Tseng, Shu-Chuan (ed.): *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*. Taipei, Taiwan: 335-341.

Keenan, Edward L./Comrie, Bernard (1977): "Noun phrase accessibility and universal grammar". *Linguistic Inquiry*, 8: 62-100.

Kohonen, Teuvo (1984): *Self-Organization and Associative Memory*. Springer-Verlag, Berlin.

Kohonen, Teuvo/Kaski, Samuel/Lagus, Krista/Salojarvi, Jarkko/Honkela, Jukka/Paatero, Vesa/Saarela, Anti (2000): "Self organisation of a massive document collection". *IEEE Transactions on Neural Networks*, 11(3): 574-585.

Lappin, Shalom/Leass, Herbert J. (1994): "An algorithm for pronominal anaphora resolution". *Computational Linguistics*, 20(4): 535-561.

Lewis, David D. (1997): *Reuters-21578 text categorization test collection*, distribution 1.0.

Lewis, David D./Sparck-Jones, Karen (1996): "Natural language processing for information retrieval". *Communications of the ACM*, 39(1): 92-101.

Rauber, Andreas/Merkl, Dieter (1999): "The SOMLib digital library system". In: Abiteboul, Serge/Vercoustre, Anne-Marie (eds.): *Proceedings of the 3rd Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL'99), Paris, France*. Berlin: 323-341.

Salton, Gerard/Buckley, Chris (1988): "Term-weighting approaches in automatic text retrieval". *Information Processing and Management*, 24(5): 513-523.

Schütze, Hinrich (1997): *Ambiguity Resolution in Language Learning*. Stanford, California.

Strzalkowski, Tomek (ed.)(1999): *Natural Language Information Retrieval*. Dordrecht.

Strzalkowski, Tomek/Carballo, Jose Perez/Marinescu, Mihnea (1995): "Natural language information retrieval: Trec-3 report. overview of the third text retrieval conference (trec-3)". In: Harman, Donna K. (ed.): *NIST Special Publication 500-225*. National Institute of Standards and Technology. Gaithersburg, MD: 39-53

Wehrli, Eric (1997): *L'analyse syntaxique des langues naturelles*. Paris.

Yarowsky, David (1995): "Unsupervised word sense disambiguation rivaling supervised methods". In: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistic*s. Cambridge, MA: 189-196.