Statistical Methods for the Evaluation of Indexing Phrases

Antoine Doucet, Helena Ahonen-Myka

STATISTICAL METHODS FOR THE EVALUATION OF INDEXING PHRASES

Antoine Doucet
Department of Computer Science, University of Caen Lower-Normandy, France
first.last@info.unicaen.fr

Helena Ahonen-Myka
Department of Computer Science, University of Helsinki, Finland
first.last@cs.helsinki.fi

Keywords: Text Mining. Natural Language Processing. Keyphrases. Collocations.

Abstract: In this paper, we review statistical techniques for the direct evaluation of descriptive phrases and introduce a new technique based on mutual information. In the experiments, we apply this technique to different types of frequent sequences, hereby finding mathematical justification of former empirical practice.

1 INTRODUCTION

The growing quantity of textual data has required adapted methods for retrieving relevant information from overwhelming document collections. A method was developed to efficiently extract content descriptors from large document collections. The technique is based on datamining concepts and is extracting the longest frequent sequences. The resulting descriptors are known as the maximal frequent sequences (MFS) (Ahonen-Myka and Doucet, 2005).

These compact descriptors take the form of word sequences. A challenge is to estimate the relevance of these descriptors. *Relevance* is indeed a subjective notion, which naturally implies that many difficulties arise when one seeks for a numerical evaluation. The usual way is to ask to a domain expert to evaluate a random sample of the results. But this is unfortunately extremely time-consuming, and the subjectivity of the domain expert remains fully correlated with the estimate.

Therefore, based on the example of the maximal frequent sequences, we review in this paper different ideas to numerically estimate the "a priori relevance" of content descriptors. A strong motivation is that an expert judgement is usually requested after a project was finished, but there are few alternatives during the development process. Because it is of course impossible to ask for a daily expert evaluation, based on unachieved work. Being able to estimate the descriptors at any time would be very helpful.

Many of the statistical techniques used to acquire collocations have appeared very interesting for this purpose. Thus, the concept of a collocation is explained in Section 2, to be able to relate it to maximal frequent sequences, described in Section 3, and a study of the existing work on collocation acquisition is given in Section 4. We will then present an extension of our descriptors, resulting of an initially empirical postprocessing, of which use we want to evaluate numerically (Section 5). The choice and definition of our estimation technique is made in Section 6. That technique was implemented and tested on a financial news corpus in Section 7. Finally, a brief conclusion is given in Section 8.

2 WHAT ARE COLLOCATIONS ?

Extracting collocations has a variety of applications. Using the likeliness that one word occurs after another can be used for **disambiguation**. **Lexicography** is another evident application: Many dictionaries are aiming to integrate the variations of meaning, induced by combining words. Thereafter, collocations have been extensively used to improve the fluency of language **generation** systems, by using a lexicon of collocations or word phrases during the word selection phase. The other common application is machine **translation**. Since collocations cannot be characterized by using syntactic and semantic regularities,

and thus they cannot be translated on a word-by-word basis, they need to be known from a bilingual collocation lexicon. Such a lexicon can be built semi-automatically by using text alignment techniques, applied to a bilingual corpora.

2.1 The lexicographic and linguistic approaches

Many authors pointed out that collocations are not easy to define. McKeown and Radev describe them as *"covering word pairs and phrases that are commonly used in language, but for which no general syntactic or semantic rules apply"* (McKeown and Radev, 2000). In the linguistic and lexicographic literature, collocations are usually said to lie *somewhere* between two opposite types of word phrases, free word combinations and idioms. From these points of view, these notions differ of that of a collocation:

- A **free word combination** can be described using general rules, respecting a certain syntactic relation. For example: *run+[object]* (i.e., manage), where "object" is an open-ended class.
- An **idiom** is a rigid word combination to which no generalities apply. For example: *foot the bill*, where no word can be interchanged.

Collocation fall between these two extremes. An example of a collocation is *to explode a myth*, which falls neither in the free word combination's nor in the idiom's categories. Indeed, *myth* and some other words (e.g. "idea" or "theory"...) can be substituted, but this exchange is not opened to any class (and in this specific case, it is not opened to the class [object] of the verb "to explode"). One can easily guess that, in practice, this categorization can be really difficult. Similar simple combinations can easily trigger two categories. For example, the combination [adjective]+table should be categorized as a free word combination in some cases ("red/blue/wooden table") and as a collocation in some others ("multiplication/tennis table").

However, many of these subtleties are barely dealt with via automated statistical techniques. And our work does not focus on extracting collocations as such, but rather on exploiting a study of known collocation acquisition techniques so as to find means to evaluate the relevance of descriptors. Thus, for our purpose, it is much more appropriate to adopt a slightly different definition of a collocation, that of (Benson, 1990):

"A collocation is an arbitrary and recurrent word combination."

This shift is easily justified by Smadja's observation that, depending on their interests and points of view, researchers have focused on different characteristics of collocation, resulting in no consensus about a global definition (Smadja, 1993). However, some general properties of collocations have been pointed out.

2.2 Some general properties of collocations

Collocations are said to be "Arbitrary". This notion enlightens a more intuitive feature of collocations. If one word of a collocation is substituted by a synonym, the resulting phrase may become "peculiar", or even incorrect. Indeed, one can definitely wish "warm greetings", but "hot greetings" would make the audience more skeptical.

Collocations may rely on a domain. There are numerous domain-specific collocations, that either do only occur in one specific domain, or have a particular meaning in this domain.

The main consequence is that any natural language processing (NLP) application (translation, disambiguation, language generation...) based on a domain-specific corpora requires a specific lexicon for that domain. Building this lexicon consists in the process called *terminology extraction*.

Collocations occur! The best known practice to recognize collocations has been to observe them. This observation is primordial in statistical extraction techniques. It is simply a consequence of the fact, that even if they do not obey any general syntactic or semantic rule, collocations appear in text. Observing regular occurrences of neighboring words is an excellent way to suspect them to form a collocation.

3 MAXIMAL FREQUENT SEQUENCES

The technique of extracting *Maximal Frequent Sequences* (MFS) from a document collection is extensively described, for instance (Doucet and Ahonen-Myka, 2006). We will hereby summarize the main steps of that method and later remind of its specific strengths.

3.1 MFS: Definition and Extraction Technique

The general idea fits the main phases of KDD (Knowledge Discovery in Databases), that is, selection and cleansing of the data, followed by the use of core mining techniques, and a final post-processing step, intending to transform and select the results into an understandable knowledge.

3.1.1 Definition of MFS

Assuming S is a set of documents, and each document consists of a sequence of words...
Definition 1 *A sequence $p = a_1 ... a_k$ is a* subsequence *of a sequence q if all the items a_i, $1 \leq i \leq k$, occur in q and they occur in the same order as in p. If a sequence p is a subsequence of a sequence q, we also say that p* occurs *in q.*

Definition 2 *A sequence p is* frequent *in S if p is a subsequence of at least σ documents of S, where σ is a given frequency threshold.*

Note that only one occurrence of a sequence within a document is counted: whether a sequence occurs once or several times within the same document does not change its frequency.

Definition 3 *A sequence p is a* maximal frequent (sub)sequence *in S if there does not exist any sequence p' in S such that p is a subsequence of p', and p' is frequent in S.*

3.1.2 Preprocessing

We first rely on a stop list to remove the most common words. Typically, the two following text fragments:

```
...President of the United States Bush...
...President George W. Bush...
```

would be resulting in:

```
...President United States Bush...
...President George Bush...
```

3.1.3 The Extraction Technique: An overview

Initial phase: Collecting all Frequent Pairs. In this initial phase, all pairs of words, such that their frequency is greater than a given threshold, σ (10 in the experiment), are being collected. Two words form a pair if they occur in the same document, and if their distance is less than a given maximal gap. A gap of 2 was used in the experiment, which means, that at most 2 other words can appear between the words forming a pair. Also, note that the pairs are ordered, i.e. the pairs (A,B) and (B,A) are different.

Expanding the frequent pairs to MFSs. For each step k, $Grams_k$ is the number of frequent sets of length k. Hence, the frequent pairs found in the initial phase form $Grams_2$. A straightforward bottom-up approach was not possible because of the size of the data. Therefore, the method combines bottom-up and greedy techniques. Each step k is then compounded of *expansion*, *pruning*, and *junction* stages. Although this is done in a greedy manner, the efficiency profit is still substantial. The interleaving processes of expansion, junction and pruning are detailed in (Ahonen-Myka and Doucet, 2005).

Finally, as a result, an (eventually empty) list of content descriptors is attached to each document of the collection.

3.2 Global Strengths

The method efficiently extracts all the maximal frequent word sequences from the collection. From the definitions above, a sequence is said to be maximal, if and only if no other frequent sequence contains that sequence.

Furthermore, a *gap* between words is allowed: the words do not need to appear continuously. A parameter g tells how many other words two words in a sequence can have between them. The parameter g usually gets values between 1 and 3.

For instance, if g = 2, a phrase "president Bush" will be found in both of the following text fragments:

```
...President of the United States Bush...
...President George W. Bush...
```

Note: The words "of" and "the" were notably removed during the preprocessing step.

This allowance of gaps between words of a sequence is probably the strongest specificity of this method, compared to the other existing methods for extracting text descriptors. This greatly increases the quality of the phrase, since the variety of natural language can be processed. The method is *style tolerant*. Even deficient syntax can be handled (which is fairly common in news wires, for example).

Another specificity is the ability to extract maximal frequent sequences of any length. This allows a very compact description. By example, by restricting the length of phrases to 8, the presence, in the document collection, of a frequent 25 words long phrase,

would result in thousands of phrases representing the same knowledge as the one maximal sequence.

4 RELATED WORK ON COLLOCATION ACQUISITION

The initial work on collocation extraction is that of (Choueka et al., 1983). Their definition of a collocation was "a sequence of adjacent words that frequently appear together". The sequences were theoretically of any length, but were limited to size 6 in practice, due to repeated frequency counting. It was experimented on an 11 million words corpus from the *New York Times* archive and found thousands of common expressions such as *"home run", "fried chicken", "Magic Johnson", etc.* After pointing the limited size of the sequences, one can also regret the impossibility to extract any discontinuous sequence such as *"knock ... door"*, due to the adjacency principle of the definition. Finally, the selection/rejection is simply based on a frequency threshold, which makes the result depend on the size of the corpus.

(Church and Hanks, 1990) described a collocation as a pair of correlated words. That is, as a pair of words that occur together more often than chance. The technique is based on the notion of *mutual information*, as defined in Information Theory (Shannon, 1948; Fano, 1961). This new set of techniques permits to retrieve interrupted sequences of words as well as continuous ones. Unfortunately, the set of the candidate sequences is now restricted to pairs of words. In other words, we can only acquire collocations of size 2, where Choueka's technique was up to 6.

Smadja proposed a more advanced technique, built on Choueka's. It resulted in Xtract (Smadja, 1993), a tool combining a frequency-based metric and several filters based on linguistic properties. The metric used by Smadja was the *z-score*. The *z-score* of a pair is calculated by computing the average-frequency of the words occurring within a 5-words radius of a given word (either forward or backward), and then determining the number of standard deviations above the average frequency for each word pair. Pairs with a z-score under a certain threshold were pruned away. Then, linguistic filters were applied to get rid of those pairs, which are not true lexical collocates. For example, for a same pair "noun-verb", the technique differentiates the case were the noun is the subject or the object of the verb. Semantically related pairs (such as *doctors-hospitals*) were also removed. After the identification of these word pairs, the collocation set was recursively extended to longer phrases, by searching for the words that co-occurred significantly together with an already identified collocation. A lexicographer was asked to estimate Xtract's result. After the full processing, including the statistical stages and linguistic filtering, 80% of the phrases were evaluated as good collocations. The score was only 40% before the syntactic filtering, illustrating the primary importance of combining both linguistic and syntactic information, in order to find accurate lexical collocates.

Of course, our technique is not as strict as Smadja's, regarding the definition of a collocation, and most of its linguistic filtering can be regarded as unnecessary for our purpose. Indeed, we are not fundamentally aiming at the discovery of collocations from a document collection, but considering collocation-based techniques to estimate the value of document descriptors. As a matter of fact, and as mentioned earlier, we will rather stick to Benson's definition of a collocation, that is probably the most appropriate to statistical techniques: an arbitrary and recurrent word combination. Based on this approach, we will now compare maximal frequent sequences to other types of descriptors.

4.1 Specificities of MFSs as collocations

Among the most satisfactory aspects of MFS extraction is the possibility to discover phrases of any size. From this point of view, it adds up from both Choueka and Church. Another clear strength, opposed to *Choueka et al.*, is the ability to compose phrases from non-adjacent words. This is due to two reasons. First, the use of a gap, the maximal number of words allowed between two other words, so as to consider them as a pair. Second, the use of a list of stop words, which prunes away most of the less informative words. The negative aspect of this stop word filtering is that most os the collection following the *verb+adverb* (e.g., "take ... off", "turn ... on") pattern will be missed. A solution would be parts-of-speech based preprocessing, so as to make sure we keep the adverbs corresponding to these possibly relevant phrases (and only those). Our technique has also the advantage over that of Smadja, that it does not require the computationally heavy combination of frequency and distance. Indeed, using windows of radius 5 implies, for each word of the corpus, to form 10 pairs of words and to calculate their frequency.

Another difference with Smadja's Xtract, is that our technique does not unite a pair and its inverted form, as a one and same phrase. We consider sequences rather than phrases. For example, noun-verb and verb-noun are different in our view, whereas in Smadja, they are first gathered together and then eventually pruned by the z-score threshold. Given a pair,

the z-score can be seen as a filter based on the statistical distribution of the position of one of these words, relatively to the other one. If they pass the z-score, they may still be pruned by Smadja's second filtering: the differentiation between subject-verb and object-verb. We suspect a good estimate can be obtained by using the fact that in a subject-verb pattern, the noun will very likely appear first, whereas in an object-verb pattern, the noun will rather appear after the verb. Thus, an approximation of these filterings is done at first sight in our work, and in one pass, since we considered relevant collocations to occur mostly in the same order. However, it is important to note that if these observations make much sense for English, they may be totally misleading for some other language (Doucet, 2005). But an essential difference between the suggested method and the ones presented above, is that it is a knowledge discovery method, built on data mining concepts. A summary of the different techniques is shown in Table 1. As such, it implies numerous simplifications and must be considered in conjunction with the previous observations.

5 MORE DESCRIPTORS: THE SUBMAXES

Context. In a practical application of MFSs, a supplementary post-processing has been executed (Ahonen-Myka et al., 1999). This experiment was meant to find co-occurrences of text phrases (the descriptors, i.e., both MFSs and submaxes) by computing association rules. An example association rule is:

```
jersey_guernesey  => channel_islands
                (0.78,0.05)
```

...meaning, that when the word sequence *(jersey guernesey)* occurs in a document, then the sequence *(channel islands)* occurs in the same document with a probability 0.78 (or 78%), this value being called the *confidence*. Both of the phrases occur together in 5% of the documents of the collection(*support*).

Submaxes. For this purpose the authors have found it useful to add more descriptors to the maximal frequent sequences. They added some of the frequent subsequences of the MFSs. The rule was the following: For each maximal frequent sequence, any of its subsequences responding to both of the following criteria was selected:

- its frequency is bigger than the corresponding maximal frequent sequence's
- it is not the subsequence of some descriptive sequence having an equal frequency

Goal. The motivation was then, that by computing maximal frequent sequences, the length of the selected sequences was increased, and the corresponding frequencies naturally tended to decrease towards the minimum frequency threshold. Thus sequences that were both shorter and more frequent were not selected, even if they might carry more information. This can be especially true when the frequency gets much higher, by taking a few words out of a sequence. That is how the submaxes post processing was initiated.

Nevertheless, the usefulness of these additional descriptors has never been formally proven. Being able to estimate the relevance of the submaxes and compare it to that of the maximal frequent sequences would be of great interest, and this is one of the goals of the following experiments.

6 Choice of the estimation technique

6.1 Many Alternatives

The fact that this technique does not compute any distance between words, using the concept of windows is an advantage, regarding computational complexity. This also implies that most of the numerous evaluation techniques based on the mean and variance of the distance between the words of a pair cannot be considered. Smadja's z-test is then out of reach.

Another specificity of our descriptors needs to be reminded here, to support the choice of an estimation technique. First, the notion of frequency is slightly different of what one would expect: the frequency of a word (or an n-gram) is not its number of occurrences, but the number of documents in which it appears. Second, the candidate bigrams with their frequency below a certain threshold are ignored. This cut-off ameliorates the efficiency of an estimation based on mutual information, as pointed by (Manning and Schütze, 1999). Indeed, pointwise mutual information has been criticized, because it gives a better score to the lowest-frequency pair, when other things are identical. The frequency threshold mostly solves this, although the underlying problem subsists.

The main other alternatives are hypothesis testing techniques, namely t-test, Pearson's chi-square test, and likelihood ratios. However, it is known, that most of these tests globally give similar results. Our aim is not to compare the different tests, but to get a rough

Table 1: Summary of the collocation acquisition techniques

	Size limit	Adjacency	Corpus size-dependency	Stoplist
(Choueka et al., 1983)	6 words	required	yes	no
(Church and Hanks, 1990)	2 words	not required	yes	no
(Smadja, 1993)	none	not required	no	no
MFS	none	not required	unclear	yes

estimate of the *interestingness* of our document descriptors. Thus, we made the choice of a variation of pointwise mutual information.

6.2 An Information Theoretic Measure

The main inspiration for this measure was the work on collocation acquisition (Church and Hanks, 1990). They ranked all pairs of words, viewing the corpus as a random distribution. Then, they compared the probability that the pair occurs, with the probability that both words occur together independently (i.e., by chance). The pointwise mutual information is the following:

$$I(w_1, w_2) = log_2 \frac{P(w_1 and w_2)}{P(w_1)P(w_2)}.$$

If I(w_1, w_2) is positive, and thus P(w_1 and w_2) is greater than P(w_1)P(w_2), it means than the words w_1 and w_2 occur together more frequently than chance. In practice, Church et al. have found, that the mutual information of almost each pair was greater than zero, due to the fact that natural language is not made of random sequences of words. Thus, the threshold needed to be raised. As a rule of thumb, they observed that the pairs with a pointwise of mutual information above 3 tended to be interesting, and pruned the others away.

In our case, pointwise mutual information, as is, cannot be used, due to our biased definition of frequency. Thus, we need to adapt all concepts and use as the probability of occurrence of a phrase P, the number of documents in which that phrase occurs, divided by the number of different words occurring in the document collection. For Church, this probability was the number of occurrences of the phrase divided by the total number of word units in the collection. Furthermore, we will extend the formula to n-grams:

$$Info(w_1, w_2, \ldots, w_n) = log_2 \frac{P(w_1, w_2, \ldots, w_n)}{P(w_1)P(w_2)\ldots P(w_n)}.$$

This is opposed to the intrinsic definition of mutual information, as enounced in (Fano, 1961). And it results in high scores for longer phrases, due to the iterative multiplication by the total number of items in the collection. However, in our case, the only incidence is that the longest phrases will get the best rankings, but if one compares phrases of same size, this concern is irrelevant. Also, it is important to realize that we do not want to use this estimate as an intermediate filter for pairs, priorily to an expansion to longer phrases, as was the case in Church. We use it as a post-processing technique, where we want to estimate the quality of a set of descriptors, which we are given as input.

7 EXPERIMENTS

7.1 General Results

Experiments have been implemented in Perl, using the publicly available Reuters-21578 financial news collection[1]. It contains about 19,000 documents, totalizing 2.56 millions of words. The pruning of the most common words (articles, abbreviations,...) reduced this number to 1.29 millions. This stoplist contained only 386 words out of 48,419 word types in the collection. This means that less than one percent of the word types represent one half of the total number of word tokens. The MFSs have been extracted from this document collection using a frequency threshold of 10. This resulted in 22,663 maximal sequences. The size distribution is shown in Table 2. The longest phrase is composed of 25 words, it occurred in 11 documents of the collection: ``federal reserve entered u.s. government securities market arrange customer repurchase agreements fed spokesman dealers federal funds trading fed began temporary indirect supply reserves banking system".

7.2 Ranking

The *"informativeness"* of the phrases has been computed. As suspected, the longest sequences tend to get the best ranking, as shown in Table 3. This correlation

[1] http://www.research.att.com/~lewis/reuters21578.html

Table 2: Number of maximal phrases of various length

Length	2	3	4	5	6	7	8	9	10	11
$\sigma = 10$	19,421	2,165	618	260	87	41	15	13	11	7

Length	12	13	14	15	16	17	18	19	20	21	...	25
$\sigma = 10$	5	5	171	59	32	7	10	7	7	1	...	1

Table 3: Average score per sequence length

Length	2	3	4
Average Score	0.29	6.28	13.79
Number of phrases	19,421	2,165	618

Length	5	6-10	11-25
Average Score	19.57	31.33	80.05
Number of phrases	260	166	32

between size and score is a serious concern, because it prevents us to compare phrases of different length.

Another interesting fact to be observed, is that, given a fixed sequence length, as the frequency rises, so does the average score, contradicting the weakness mentioned earlier, that when other features are identical, mutual information tends to advantage the lowest frequency. This is due to the fact, that the link between our estimate and mutual information is not that tight, because of our notion of frequency, and the use of a threshold. Table 4 shows the frequency distribution and corresponding average score for the maximal frequent pairs.

But the main point is that, given a sequence length, the score appears to be an excellent indicator of the interestingness of a descriptor. Among the 19,421 maximal frequent pairs, 8,981 occur more often than chance (46%). The best ranked are clear good descriptors: city names (*as "kuala lumpur"*), company names (*as "rolls royce"*), person names (*as "zhao ziyang"*), pairs adjective-noun *(as "chinese-made missiles")*... Among the top-ranked are also latin locutions (*"pro rata", "pro forma", and "ad hoc"*).

At size 3, many names are found again (*"javier perez cuellar", "rio de janeiro"*), but in some cases, supplementary information is also given (*"communist hu yaobang", "chancellor helmut kohl"*), as well as full entities (*"labour centrist alliance", "frozen concentrated orange"*).

At size 4, we still get names and titles (*"minister arturo hernandez grisanti"*), but some phrases are even more meaningful (*"refined bleached deodorised palm", "tax vegetable oil fat", "paid form commodities inventory"*). With longer phrases, the value of the best ranked extracted phrases is even clearer: (*"supply indirectly customer repurchase", "currencies ranges broadly consistent economic fundamentals", "commodity credit corporation ccc accepted bid export bonus cover sale"*).

Table 5: Best and worst ranked maximal frequent pairs

Phrase	Frequency	Score
kuala(11) lumpur(10)	10	12.1
piper(12) jaffray(10)	10	11.98
hoare(13) govett(12)	12	11.86
zhao(13) ziyang(11)	11	11.86
paz(13) estenssoro(10)	10	11.86
boone(10) pickens(13)	10	11.86
bettino(12) craxi(11)	10	11.84
makoto(15) kuroda(15)	15	11.66
paine(13) webber(15)	13	11.66
peat(15) marwick(10)	10	11.66
told(2393) year(5194)	10	-4.68
share(2666) inc(4608)	10	-4.67
inc(4608) after(2567)	10	-4.61
inc(4608) year(5194)	21	-4.56
share(2666) corp(4211)	10	-4.54
corp(4211) market(2839)	12	-4.36
bank(2727) inc(4608)	13	-4.32
net(3220) company(5031)	17	-4.3
u.s.(3530) inc(4608)	17	-4.3
trade(1841) company(5031)	10	-4.26

At the bottom of this ranking, the accidental aspect of the co-occurrence of the words involved is easily noticeable. They are actually enlightening the fact that some words should better have been included in the stoplist, at preprocessing time. This may actually become an application of this evaluation process: refining the stopword list by adding the words involved in too many "thrash-scored" phrases. The 10 best and worst ranked 2-grams are shown on Table 5, while the 5 best and worst ranked 3-grams and 4-grams are on Table 6. The number in parentheses after each word is its frequency.

Table 4: Frequent pairs distribution and average score per frequency

Frequency	10	11	12	13	14-15	16-20	21-25	26-50	51-171
Average Score	0.33	0.32	0.20	0.22	0.22	0.24	0.30	0.39	0.84
Number	3,593	2,672	2,064	1,654	2,496	3,367	1,438	2,141	296

Table 6: Best and worst ranked MFSs of size 3 and 4

Phrase	Frequency	Size	Score
javier(21) perez(12) cuellar(11)	10	3	23.01
denis(24) bra(12) kanon(12)	12	3	22.96
ibc(22) jorio(17) dauster(18)	11	3	21.87
philips(35) gloeilampenfabrieken(15) pglo.as(18)	13	3	21.62
communist(58) hu(14) yaobang(10)	10	3	21.46
inc(4608) inc(4608) company(5031)	10	3	-2.19
inc(4608) new(3731) company(5031)	10	3	-1.88
year(5194) after(2567) year(5194)	10	3	-1.56
co(2824) inc(4608) inc(4608)	10	3	-1.35
co(2824) inc(4608) corp(4211)	10	3	-1.22
refined(78) bleached(13) deodorised(11) palm(70)	11	4	30.57
energy(516) arturo(14) hernandez(18) grisanti(16)	13	4	29.4
energy(516) fernando(15) santos(35) alvite(12)	10	4	28.38
minister(1175) arturo(14) hernandez(18) grisanti(16)	14	4	28.32
barclays(38) de(361) zoete(20) wedd(21)	15	4	28.14
stock(2809) new(3731) stock(2809) exchange(2158)	12	4	4.42
shares(2348) new(3731) stock(2809) exchange(2158)	10	4	4.42
inc(4608) shares(2348) common(1557) stock(2809)	11	4	4.72
sales(1986) note(1668) year(5194) net(3220)	13	4	4.74
corp(4211) shares(2348) common(1557) stock(2809)	11	4	4.85

Table 7: Number of extracted subsequences (*submaxes*) of various length

Length	2	3	4	5	6	7	8	9
Submaxes	3,813	235	30	8		3		4

7.3 SubMaxes Estimation

The justification for the submaxes is that even though MFSs carry much value, the longer sequences can be hardly understandable (remember our sequence of size 25). In this section, we will try to estimate the interestingness of the submaxes, compared to that of the MFSs. Out of the 22,663 maximal frequent sequences, 4,093 submaxes were extracted. Our set of descriptors therefore contains 26,756 elements. The size distribution is shown in Table 7. One can easily observe that many of these new descriptors are good complements of the previous ones: *"mcdonnell douglas", "alan greenspan", "goldman sachs", "saudi arabia", "dow jones industrial average", "issuing australian eurobond", "sinking fond debentures",...* These phrases were not maximal sequences by their selves, because they were clearly above the frequency threshold, and thus, more words were added to them. Taking them apart, by extracting the submaxes creates very cohesive units. On average, their scores per size are indeed always better than the scores per size of the MFSs, as shown by Table 8. Also, among the submax pairs, 2,895 have a positive score (76% against 46% for the MFSs).

8 CONCLUSION

After overviewing collocation acquisition techniques, a way to estimate the interestingness of sequences describing documents has been presented and implemented. It has proven to be a good indicator of whether a sequence should be kept or pruned away. This estimate can then be used as a post-processing technique to cleanse a set of descriptors. It is, how-

Table 8: Average score of phrases: MFS vs. Submaxes

Length	2	3	4	5	6-10	11-25
MFS: Average Score	0.29	6.28	13.79	19.57	31.33	80.05
MFS: Number of phrases	19,421	2,165	618	260	166	32
Submaxes: Average Score	1.72	9.85	16.00	22.58	37.25	X
Submaxes: Number of phrases	3,813	235	30	8	7	none

ever, regrettable that we have not been able to find a technique to compare descriptors of different sizes. This is a well-known problem that we are to address in the future.

The possible applications of these descriptors are numerous. They can be used to create terminology lexicons. Many sequences found in the experiments are clear collocations of the financial domain. Text alignment of bilingual documents can be another field of application. In fact, every application of collocations is concerned. However, we are likely to focus on information retrieval. Each document of the collection will be linked to a set of descriptors. These descriptors can then be used as clusters for dynamic browsing or for indexing.

The MFS extraction technique was intended for very large document collections and, thanks to the gap feature, is especially adapted to incomplete sentences, or sentences with an incorrect syntax. These are fairly common, for example in financial news wires. To filter the descriptors, it would now be interesting to tag them with their part of speech, so as to find grammatical patterns, or just as a means to filter the patterns, as done by (Justeson and Katz, 1995), who kept only those patterns "that are likely to be phrases".

REFERENCES

Ahonen-Myka, H. and Doucet, A. (2005). Data mining meets collocations discovery. In *Inquiries into Words, Constraints and Contexts*, pages 194–203. CSLI Publications, Center for the Study of Language and Information, University of Stanford.

Ahonen-Myka, H., Heinonen, O., Klemettinen, M., and Verkamo, A. I. (1999). Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery. In *Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications*, pages 1–9.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–35.

Choueka, Y., Klein, S. T., and Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic computing*, 4:34–38.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Doucet, A. (2005). *Advanced Document Description, a Sequential Approach*. PhD thesis, University of Helsinki.

Doucet, A. and Ahonen-Myka, H. (2006). Fast extraction of discontiguous sequences in text: a new approach based on maximal frequent sequences. In *Proceedings of IS-LTC 2006, Information Society - Language Technologies Conference*, pages 186–191.

Fano, R. M. (1961). *Transmission of Information: A statistical Theory of Information*. MIT Press, Cambridge MA.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm fo identification in text. *Natural Language Engineering*, 1:9–27.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA, second edition.

McKeown, K. R. and Radev, D. R. (2000). *A Handbook of Natural Language Processing*, chapter 5: Collocations. Marcel Dekker.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech*, 27:379–423, 623–656.

Smadja, F. (March 1993). Retrieving collocations from text: Xtract. *Journal of Computational Linguistics*, 19(1):143–177.