An N-gram frequency database reference to handle MWE extraction in NLP applications

Patrick Watrin
Centre for Natural Language Processing
Institut Langage et Communication
UCLouvain
patrick.watrin@uclouvain.be

Thomas François
Aspirant F.N.R.S.
Centre for Natural Language Processing
Institut Langage et Communication
UCLouvain
thomas.francois@uclouvain.be

Abstract

The identification and extraction of Multiword Expressions (MWEs) currently deliver satisfactory results. However, the integration of these results into a wider application remains an issue. This is mainly due to the fact that the association measures (AMs) used to detect MWEs require a critical amount of data and that the MWE dictionaries cannot account for all the lexical and syntactic variations inherent in MWEs. In this study, we use an alternative technique to overcome these limitations. It consists in defining an n-gram frequency database that can be used to compute AMs on-the-fly, allowing the extraction procedure to efficiently process all the MWEs in a text, even if they have not been previously observed.

1 Introduction

Multiword Expressions (MWEs) are commonly defined as "recurrent combinations of words that co-occur more often than expected by chance and that correspond to arbitrary word usages" (Smadja, 1993, 143). Their importance in the field of natural language processing (NLP) is undeniable. Although composed of several words, these sequences are nonetheless considered as simple units with regard to part-of-speech at the lexical as well as syntactic levels. Their identification is therefore essential to the efficiency of applications such as parsing (Nivre and Nilsson, 2004), machine translation (Ren et al., 2009), information extraction, or information retrieval (Vechtomova, 2005). In these systems, the principle of syntactic or semantic/informational unit is particularly important.

Although the identification and extraction of MWEs now deliver satisfactory results (Evert and Krenn, 2001; Pearce, 2002), their integration into a broader applicative context remains problematic (Sag et al., 2001). The explanations for this situation are twofold.

1. The most effective extraction methods resort to statistical association measures based on the frequency of lexical structures. They, therefore, require a critical amount of data and cannot function properly from a simple phrase or even from a short text.
2. Since the syntactic and lexical variability of MWEs may be high, lexical resources learned from a corpus cannot take it into account. The coverage of these resources is indeed too limited when applied to a new text.

To address these two limitations, this article describes how an n-gram frequency database can be used to compute association measures (AMs) efficiently, even for small texts. The specificity of this new technique is that AMs are computed on-the-fly, freeing it from the coverage limitation that afflicts more simple techniques based on a dictionary.

We start off focussing on our extraction method, and more particularly on the process via which a candidate structure is statistically validated (Section 2). This presentation principally aims to identify the precise needs of a frequency database reference, both in terms of the interrogation process and in the type of information to be kept in the database. Then, we will address various issues of storage and query performance raised by the design of the frequency

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 83–91,
Portland, Oregon, USA, 23 June 2011.

database (Section 3). Finally, Section 4 reports the results of our experiments and Section 5 concludes and open up future perspectives.

2 Extraction process

Our extraction procedure is comparable to those developed by Smadja (1993) and Daille (1995). They use a linguistic filter upstream of the statistical estimation. Unlike purely statistical techniques, this solution provides less coverage but greater accuracy. It also allows us to assign a unique morpho-syntactic category to each extracted unit (as well as a description of its internal structure), which facilitates its integration into a more complex procedure.

Concretely, we first tagged the texts to clear any lexical ambiguities [1]. We then identified all MWE candidates in the tagged text with the help of a library of transducers [2] (or syntactic patterns). Finally, the list of candidates was submitted to the statistical validation module which assigns an AM to each of these.

2.1 Linguistic filters

In this study, we consider four basic types of nominal structures [3] : adjective-noun (*AN*), noun-adjective (*NA*), noun-preposition-noun (*NprepN*), and noun-noun (*NN*), which are likely to undergo three types of variations : modification (mainly adverbial insertion and / or adjectival), coordination, and juxtaposition (*e.g. NprepNprepN*, *NprepNN*, etc). This enables us to identify a wide variety of sequences that are labelled by XML tags which specify :

- the lexical heads of the various components ;
- the adjectival and prepositional dependencies ;
- any possible coordination.

This information can be exploited later to carry out the syntactic decomposition of the extracted structures and also to limit the statistical validation to the content words of each structure.

1. The tagging is done with the *TreeTagger* (Schmid, 1994).
2. To apply our transducers to the tagged text, we use *Unitex* (Paumier, 2003). The output of the process is a file containing only the recognized sequences.
3. As we work in the field of indexation, we limit our extraction to nominal terms.

2.2 Statistical validation

Association measures are conventionally used to automatically determine whether an extracted phrase is an MWE or not. They are mathematical functions that aim to capture the degree of cohesion or association between the constituents. The most frequently used measures are the *log-likelihood ratio* (Dunning, 1993), the *mutual information* (Church and Hanks, 1990) or the ϕ^2 (Church and Gale, 1991), although up to 82 measures have been considered by Pecina and Schlesinger (2006). In this paper, we did not aim to compare AMs, but simply to select some effective ones in order to evaluate the relevance of a reference for MWE extraction.

However, association measures present two main shortcomings that were troublesome for us : they are designed for bigrams, although longer MWEs are quite frequent in any corpus [4], and they require the definition of a threshold above which an extracted phrase is considered as an MWE. The first aspect is very limiting when dealing with real data where longer units are common. The second may be dealt with some experimental process to obtain the optimal value for a given dataset, but is prone to generalization problems. In the next two sections, we present the strategies we have used to overcome these two limitations.

2.2.1 Beyond bigrams

A common way to go beyond the bigram limitation is to compute the AMs at the bigram level and then use the results as input for the computation of higher order AMs (Seretan et al., 2003). However, our preliminary experimentations have yielded unsatisfactory results for this technique when it is applied to all words and not to heads only. This is probably a side effect of high frequency bigrams such as preposition-determiner (*prep det*) in French.

Another strategy explored by Silva and Lopes (1999) is the fair dispersion point normalization. For a given n-gram, which has $n-1$ dispersion points that define $n-1$ "pseudo-bigrams", they compute the arithmetic mean of the probabilities of the various combinations rather than attempting to pick up the right point. This technique enables the

4. In our test corpus (see Section 4), 2044 MWEs out of 3714 are longer than the bigrams.

authors to generalize various conventional measures beyond the bigram level. Among these, we selected the *fair log-likelihood ratio* as the second AM for our experiments (see Equation 1), given that the classic *log-likelihood ratio* has been found to be one of the best measures (Dunning, 1993; Evert and Krenn, 2001).

$$\begin{aligned} LogLik_f(w_1 \cdots w_n) &= 2 * \log L(pf1, kf1, nf1) \\ &\quad + \log L(pf2, kf2, nf2) \\ &\quad - \log L(pf, kf1, nf1) \\ &\quad - \log L(pf, kf2, nf2) \quad (1) \end{aligned}$$

where

$$\begin{aligned} kf1 &= f(w_1 \cdots w_n) \quad nf1 = Avy \\ kf2 &= Avx - kf1 \quad\quad nf2 = N - nf1 \end{aligned}$$

$$Avx = \frac{1}{n-1} \sum_{i=1}^{i=n-1} f(w_1 \cdots w_i)$$

$$Avy = \frac{1}{n-1} \sum_{i=2}^{i=n} f(w_i \cdots w_n)$$

$$pf = \tfrac{kf1+kf2}{N} \quad pf1 = \tfrac{kf1}{nf1} \quad pf2 = \tfrac{kf2}{nf2}$$

and N is the number of n-grams in the corpus.

Silva and Lopes (1999) also suggested an AM of their own : the *Symmetrical Conditional Probability*, which corresponds to $P(w_1|w_2)P(w_2|w_1)$ for a bigram. They defined the fair dispersion point normalization to extend it to larger n-grams, as shown in Equation 2.

$$SCP_f([w_1 \cdots w_n]) = \frac{p(w_1 \cdots w_n)^2}{Avp} \quad (2)$$

where $w_1 \cdots w_n$ is the n-gram considered and Avp is defined as follows :

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(w_1 \cdots w_i) * p(w_{i+1} \cdots w_n) \quad (3)$$

Finally, we considered a last AM : the Mutual Expectation (Dias et al., 1999) (see Equation 4). Its specificity lies in its ability to take into account non-contiguous MWEs such as "to take __ decision" or "a __ number of", which can also be realized using the heads (see above).

$$ME(w_1 \cdots w_n) = \frac{f(w_1 \cdots w_n) * p(w_1 \cdots w_n)}{FPE} \quad (4)$$

where FPE is defined as follows :

$$FPE = \frac{1}{n}[p(w_2 \cdots w_n) + \sum_{i=2}^{n} p(w_1 \cdots \widehat{w_i} \cdots w_n)] \quad (5)$$

It should be noted that the expression $w_1 \cdots \widehat{w_i} \cdots w_n$, where the $\widehat{\ }$ indicates an omitted term, represents all the n (n-1)-grams the candidate MWE comprises. FPE is then able to estimate the "glue" between all the constituents separated by a gap, but this nevertheless requires a more complex string matching process.

To summarize, we have selected the three following association measures for n-grams : the fair log-likelihood ratio, SCP_f, and ME. Their efficiency is further discussed in Section 4.

2.2.2 Selection of MWEs

The second problem that arises when one wants to locate all the MWEs in a given text is the classification criterion. For the *log-likelihood ratio*, which follows a chi-square distribution once it is transformed as $-2 * log\lambda$, a first solution is to base the decision on the p-value. However, significance tests become highly unreliable for large corpora, since the high frequencies produce high scores for the chi-square and all phenomena then appear significant (Kilgarriff, 2005).

A second technique commonly used in the MWE literature is to select a threshold for the AM above which an analyzed phrase is considered as an MWE. Again, this threshold depends on the size of the corpus used and cannot be fixed once and for all for a specific AM. It must be obtained empirically for each application of an MWE extractor to a new text or to a new domain. In order not to resort to a threshold, (Silva et al., 1999) suggested the *LocalMax* algorithm that selects MWEs whose AMs are higher than those of their neighborhood. In other words, a given unit is classified as an MWE if $g(w_1 \cdots w_n)$, the associative function, is a local maximum.

In our case, since the notion of reference implies a large corpus and high frequencies, we rejected the first of these three approaches. We experimented with the second and third and show in Section 5 how the use of a reference could partially solve the threshold issues.

3 Reference Building

The integration of MWEs in an NLP system is usually done via a dictionary. MWEs are then regarded as a sequence of simple words separated by spaces (Sag et al., 2001). As a result, their lexical and syntactic structure is fixed and cannot be used to take into account variation at this level.

Several methods have been proposed to overcome this limitation. Nerima et al. (2006) and Sag et al. (2001) associate each MWE with a feature structure specifying the nature of units and the type of fixedness. This approach requires a manual validation of the features when inserting them into the dictionary. Watrin (2007) considers a simpler technique that consists in identifying, for each type of structure, all the possible insertion points and specifying the lexical and syntactic nature of possible modifiers. In this case, each MWE takes the form of a regular expression formalizing all possible variations from the canonical form.

Both solutions enable to consider more MWEs but fail to express all possible variations. For instance, phenomena such as coordination or juxtaposition do not seem to be taken into account by the authors mentioned above including Nerima et al. (2006). Moreover, they limit lexical variations to a finite set of canonical structures that have been encountered and are therefore unable to recognize new candidates.

The notion of reference which we define in this article aims to overcome these two limitations. Rather than providing a list of MWEs that are precomputed on a corpus, we suggest storing the information needed to calculate various AMs within a database. Hence, we no longer restrict MWEs to a finite set of lexical entries but allow the on-the-fly computation of AMs for any MWE candidate, whatever the size of the input text.

3.1 Implementation details

From a computational point of view, this idea involves the compression of a large number of lexical structures of order N as well as their absolute frequency. Moreover, the calculation of the various AMs considered in this study also requires the frequencies of all structures of order n, strictly lower than N ($0 < n < N$). The second type of information can however be inferred from the frequency of the structures of order N, provided the storage and questioning system is efficient enough for real-time applications. The need for efficiency also applies to queries related to the ME measure or the LocalMax algorithm that partly involve the use of wildcards.

This type of search tool can be efficiently implemented with a PATRICIA tree (Morrison, 1968). This data structure enables the compression of n-grams that share a common prefix and of the nodes that have only one child. The latter compression is even more effective as most of the n-grams have a unique suffix (Sekine, 2008). Beyond the compression that this structure allows, it also guarantees a very fast access to data insofar as a query is a simple tree traversal that can be done in constant time.

In order to further optimize the final data structure, we store the vocabulary in a table and associate an integer as a unique identifier for every word. In this way, we avoid the word repetition (whose size in memory far exceeds that of an integer) in the tree. Moreover, this technique also enables to speed up the query mechanism, since the keys are smaller.

We derived two different implementations of this structure. The first stores the data directly in memory. While it enables easy access to data, the number of n-grams that can be stored is limited by the capacity of the RAM. Therefore, in order to take a huge number of n-grams into account, we also implemented a ”disk” version of the tree.

Finally, in order to treat wildcard queries needed by the ME and the LocalMax, we enhanced our structure with a set of indexes to improve access to each word, whatever its depth within the tree. Obviously, this mechanism might not be robust enough for a system multiplying the number of wildcards, but it is perfectly suited to the needs of an MWEs extraction process.

3.2 References used

Once the computational aspects of reference building have been dealt with, a corpus from which to populate the database needs to be selected. This aspect raises two issues : the size and the nature of the corpus used. Dunning (1993) has demonstrated that the size of the corpus from which MWEs are extracted matters. On the other hand, common characteristics of a corpus, such as its register, the contempora-

Reference	# 5-Grams	# Nodes
500 K	500,648	600,536
1000 K	1,001,080	1,183,346
5000 K	5,004,987	5,588,793
Google	1,117,140,444	62,159,203

TABLE 1: Number of 5-grams and nodes in the references used

neity of its language or the nature of the topics covered, may impact the performances of a reference when used on a text with different characteristics.

Given these issues, four corpora were selected (*cf.* Table 1). The first three are made up of articles published in the Belgian daily newspaper *Le Soir* in 2009, with 500K, 1000K and 5000K words respectively. They share many characteristics with our test corpus. The last corpus is made up of the largest amount of n-grams publicly available for French : the Google 5-grams [5] (Michel et al., 2011). Its size reaches 1T words [6], and its coverage in terms of topic and register is supposedly wider than corpora of newspaper articles only. In a sense, the Google reference may be viewed as an attempt to a universal reference.

4 Evaluation

Most evaluations of MWE extraction systems are based on human judgments and restrict the validation process to the n-best candidates. Inevitably partial, this method is unable to estimate performance in terms of recall. To overcome these limitations, we use the evaluation method described by Evert and Krenn (2001). They propose an automatic method that consists in computing both recall and precision using various n-best samples. It involves the formation of a golden standard (i.e. a list of MWEs manually identified in a corpus) and a sorted list of MWEs extracted automatically by applying AM on the same corpus. The recall and precision rates are therefore calculated by comparing the n-best (where n increases from 0 till n in steps of x) to the golden standard list [7].

4.1 The test corpus

In this study, we use the corpus described in Laporte et al. (2006). It is a French corpus in which all MWEs have been manually annotated. It consists of two sub-corpora :

– the transcription, in a written style, of the October 3rd and 4th, 2006 meetings of the French National Assembly (FNA), and
– the complete text of Jules Verne's novel "Around the World in 80 Days", published in 1873 (JV).

These two sub-corpora respectively contain 98,969 and 69,877 words for a total of 3,951 and 1,103 MWEs [8]. We limit our evaluation to the FNA corpus in order to keep data consistent both in terms of register and time. We assume that these two variables have a direct impact on the use of MWEs, a hypothesis that seems to be confirmed by the rate of MWEs in both sub-corpora.

4.2 Extractor Parameters

Before evaluating the performance of each of the above mentioned references, we first assessed the influence of the various parameters involved in the extraction process and which affect the performance of the AMs. These parameters are the LocalMax, the smoothing technique, the lemmatization of the MWE constituents (LEMMA) [9] and the head-driven validation (HDV) [10]. To select the optimal parameters for our extractor, we established an additional reference (1000K words from *Le Soir*).

5. For the purposes of comparison, we also limited the size of the n-grams indexed in *Le Soir* to 5 words.

6. In order to model a contemporary language, we only kept the frequencies observed in texts written between 2000 and 2008.

7. We build these lists from MWE types to avoid introducing a bias in the evaluation process. Well-recognised high frequency MWEs might indeed gloss over poorly recognised low-frequency MWEs.

8. These occurrences correspond to 1,384 MWE types for the FNA corpus and 521 for the JV corpus.

9. The lemmatization of the MWE constituents is based on the assumption that the inflexion of the lemmas implies a dispersal of the frequency mass (the overall frequency of a lemma is split between its inflected forms) that may affect the behavior of the AMs.

10. The HDV aims to focus on the lexical heads of the MWE candidates. Therefore, function words (prepositions, conjunctions, etc.) are ignored and replaced by wildcards in the queries sent to the reference in order to keep the distance information. For instance, from the sequence *ministre de l'agriculture* (Minister for Agriculture), we derive the form *ministre * * agriculture*.

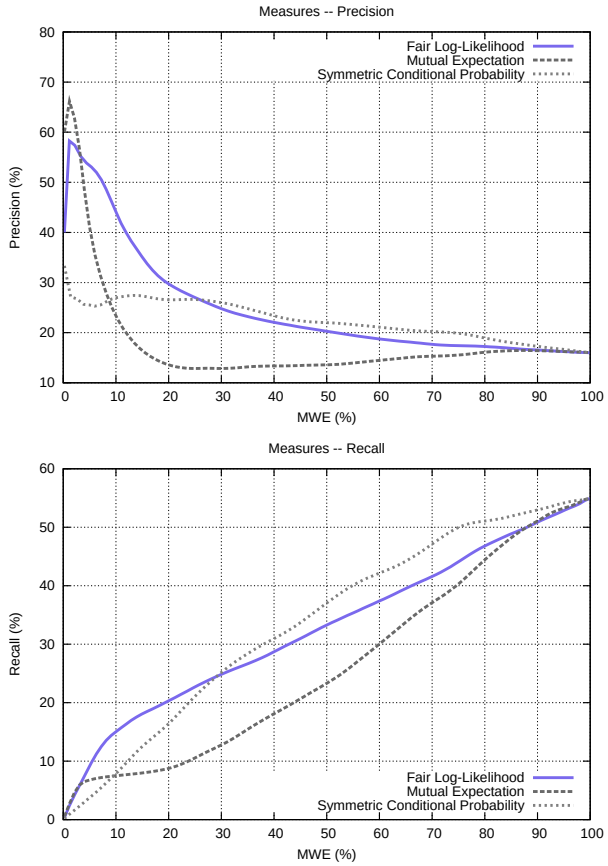

FIGURE 1: Evaluation of AMs

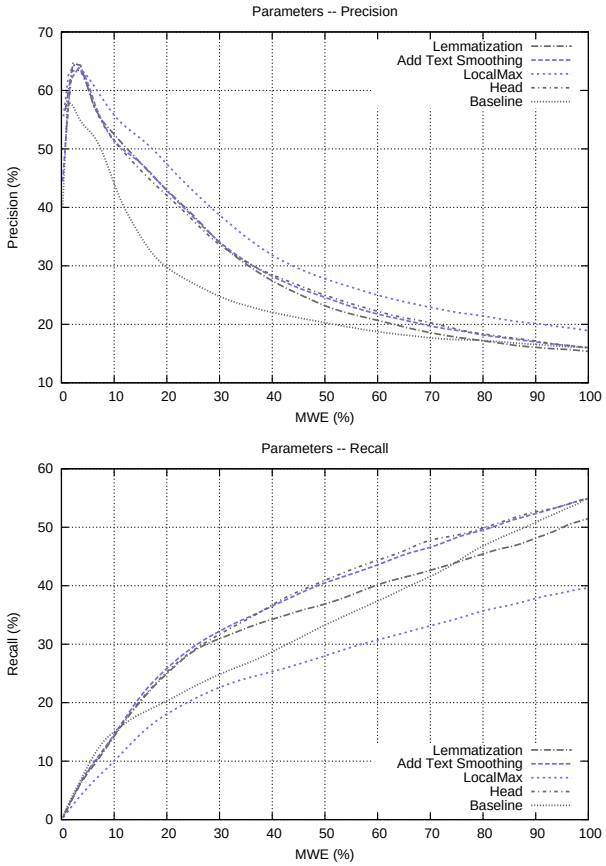

FIGURE 2: Evaluation of the parameters

The first step of this selection procedure was to define a baseline. For this purpose, we compared the precision and recall rates of our three AMs (see Figure 1) and kept only the best, namely the *log-likelihood ratio*, for the rest of our experiments. While the ME provides better precision for the top five percent of the extracted units, the *log-likelihood ratio* appears more reliable in that it maintains its efficiency over time (for recall as well as precision). The SCP, for its part, displays more stable results but does not reach sufficient precision.

On the basis of this baseline, we then separately compared the contribution of each of the four parameters. Results are reported in Figure 2 and detailed in the following subsections.

4.2.1 The LocalMax

Figure 2 shows that the LocalMax significantly improves the precision of the extraction. It emerges as the most relevant parameter at this level. However, unlike other parameters, its application directly affects the recall that falls below our baseline. This may not be a problem for certain applications. In our case, we aim to index and classify documents. Therefore, while we can accommodate a lower precision, we cannot entirely neglect the recall. We thus abandoned this parameter which, moreover, indubitably increases the processing time in that it requires the use of approximate matching (see Section 3.1).

4.2.2 The *Add-text smoothing*

Smoothing is another aspect worthy of consideration. No matter how large the reference used is, it will never constitute more than a subset of the language. Therefore, it is necessary to find a solution to estimate the frequency of unobserved n-grams. For the baseline, we used a simple "add-one“ (or Laplace) smoothing (Manning and Schütze, 1999) which presents a severe flaw when the size of the n-grams to smooth increases : the normalization pro-

cess discounts too much probability mass from observed events.

We therefore compare this simple method with another one we consider more "natural" : the "add-text" smoothing that adds the text to process to the reference. We view this method as more natural to the extent that it simulates a standard MWE extraction process. In this case, the reference complements the frequency universe of the input corpus as if it formed a homogeneous whole. Figure 2 demonstrates a clear superiority of the second smoothing procedure over the first one which was therefore discarded.

4.2.3 Lemmatization and HDV

The lemmatization and HDV follow a similar curve with regard to precision, although HDV is better for recall. Nonetheless, this difference only appears when precision falls below 35%. This does not seem sufficient to reject the lemmatization process whose computation time is significantly lower than for the HDV. We therefore limit the use of this last parameter to the reference built from Google whose n-grams cannot be lemmatized due to lack of context. [11]

4.3 Evaluation of the references

The estimation of the parameters allowed us to establish a specific evaluation framework. Two sets of parameters were defined depending on whether they apply to Google (ATS + HDV) or to the references built from *Le Soir* (ATS + LEMMA). From a practical standpoint, we limited the MWE extraction to nominal units of size inferior to five in order to meet the characteristics of our test corpus (the annotations of which are limited to nominal sequences), on the one hand, and to allow comparability of results on the other hand (the n-grams from Google do not exceed the order 5).

Initially, we considered the extraction of MWEs in the whole evaluation corpus. Results displayed in Figure 3 provide an advantage over the use of a reference with respect to the extraction carried out on the test corpus only. In addition, we see a clear improvement in performance with respect to that obtainable with a dictionary of MWEs. [12]

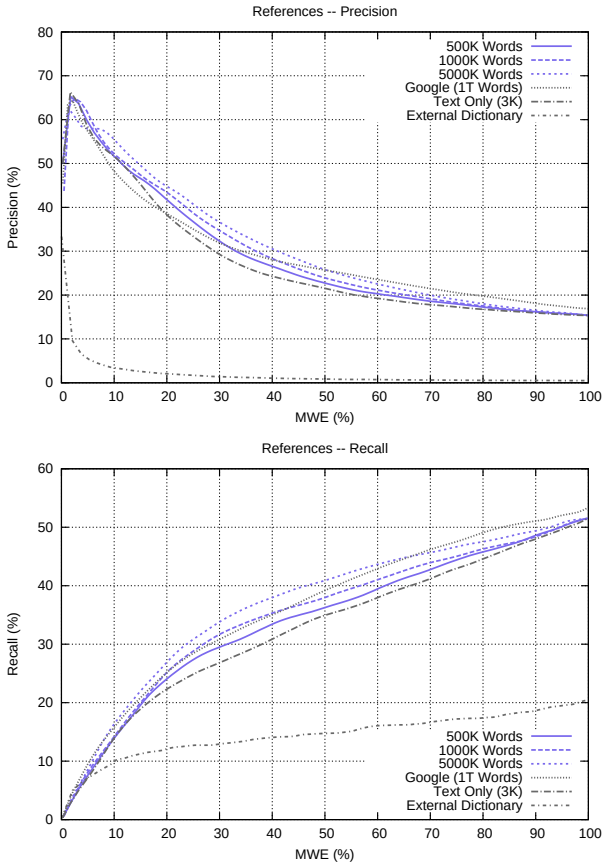

FIGURE 3: Evaluation on the 100K Corpus

In a second step, we wanted to test the efficiency of our references in the more adverse context of a short text. We randomly selected 3K words of our test corpus to simulate a short text while maintaining a sufficient number of MWEs (i.e. 151 nominal MWEs). Results shown in Figure 4 further confirm our first experience and validate our concept of a reference in a real application context.

Beyond validating the use of a frequency base, these results also confirm the general idea that the size of the corpus used for the reference matters. The differences between the references of 500K, 1000K and 5000K words showed a continuous improvement both in precision and recall. The results obtained with the Google reference are more surprising, since they do not meet that growing trend. However, given the number of errors that those n-grams contain (mainly due to the OCR-ization and tokeni-

11. References constructed on the basis of the newspaper *Le Soir* have been reindexed from a lemmatized text.

12. The MWE dictionary used in this experiment was initially derived from the corpus of 5000K words used to build the corresponding reference.

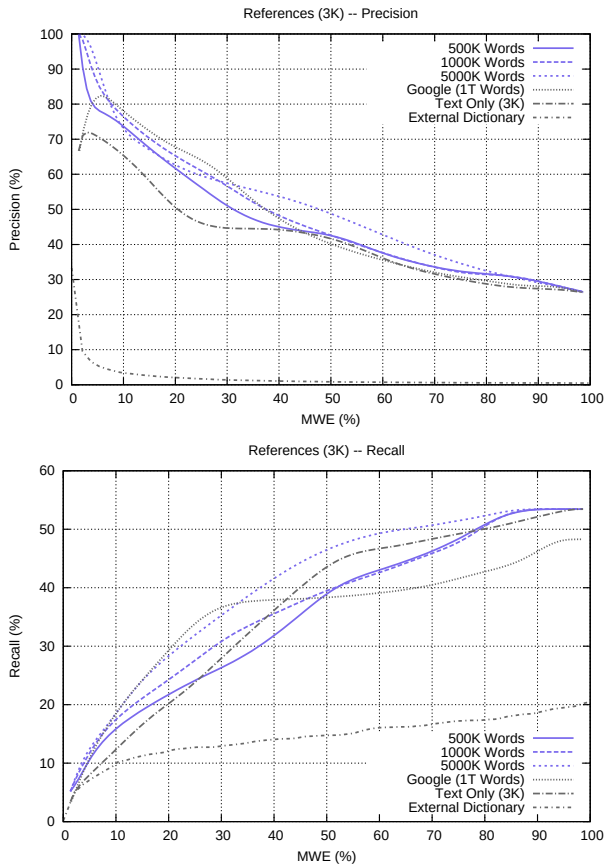

FIGURE 4: Evaluation on the 3K Corpus

zation processes), the result remains satisfactory. It even confirms to some extent the importance of size in the sense that preprocessing errors are being mitigated by the global mass of the frequencies.

5 Conclusion and perspectives

In this paper, we presented an MWE extraction system based on the use of frequency references. We have shown that its use enables MWE extraction on short texts with performances that are at least comparable to those achieved by standard solutions and far superior to solutions based on the use of MWE dictionaries.

Moreover, as this system has been integrated within an indexing engine, various issues were raised, some of which constitute avenues for future research. First, since our indexer aims at the identification of entities and terms specific to a given specialty area, the question of data representativeness is of particular importance. It is not clear to what extent a given reference can be applied to various types of texts. We only noticed that the Google reference, whose features were less similar to the test corpus, nevertheless yielded satisfactory results in comparison with our other references that better fitted the test corpus features.

MWE	500 K	1000 K	5000 K	Google
même groupe	0.73	1.44	3.85	1,746.03
nouveaux instruments	3.81	3.3	49.83	2,793.65
lettres de noblesse	33.99	52.43	232.51	27,202.17

TABLE 2: Examples of MWEs candidates whose *log-likelihood ratio* is not significant on a small corpus and becomes extremely significant on a large corpus. They are compared to the score of an actual MWE.

In addition, our results show that the threshold issue remains relevant. Although the LocalMax seems to allow better discrimination of the MWE candidates, it is not selective enough to keep only the actual MWEs. On the other hand, as the size of the references increases, some results of the AMs based on the *log-likelihood ratio* reach high values that can no longer be interpreted by a chi-square significance test (see Table 2).

We believe that our references offer an interesting perspective to face this problem. The stability of their frequencies makes it possible to define a threshold corresponding to a specific percentage of precision and recall (set according to the needs of a given application). Therefore, as long as the size of the analyzed texts remains limited – which can be controlled –, the efficiency of this threshold should remain constant. Further experimentations on this aspect are however required to determine to what extent this assumption stands true as the size of the analyzed texts grows.

References

K.W. Church and W.A. Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.

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

J. da Silva and G.P. Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*.

B. Daille. 1995. Combined approach for terminology extraction : lexical statistics and linguistic filtering. Technical report, Lancaster University.

G. Dias, S. Guilloré, and J.G.P. Lopes. 1999. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *Proceedings of the 6th Conference on the Traitement Automatique des Langues Naturelles (TALN1999)*, pages 333–339.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1) :61–74.

S. Evert and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 188–195.

A. Kilgarriff. 2005. Language is never ever ever random. *Corpus linguistics and linguistic theory*, 1(2) :263–276.

E. Laporte, T. Nakamura, and S. Voyatzi. 2006. A french corpus annotated for multiword expressions with adverbial function. In *Proceedings of the Language Resources and Evaluation Conference (LREC) : Linguistic Annotation Workshop*, pages 48–51.

C.D. Manning and H. Schütze, editors. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014) :176–182.

D.R. Morrison. 1968. PATRICIA—practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, 15(4) :514–534.

L. Nerima, V. Seretan, and E. Wehrli. 2006. Le problème des collocations en TAL. *Nouveaux cahiers de linguistique française*, 27 :95–115.

J. Nivre and J. Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of LREC-04 Workshop on Methodologies & Evaluation of Multiword Units in Real-world Applications*, pages 37–46.

S. Paumier. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée.

D. Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1530–1536.

P. Pecina and P. Schlesinger. 2006. Combining association measures for collocation extraction. In *Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, pages 651–658.

Z. Ren, Y. L, J. Cao, Q. Liu, and Y. Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications*, pages 47–54.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2001. Multiword expressions : A pain in the neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

S. Sekine. 2008. A linguistic knowledge discovery tool : Very large ngram database search with arbitrary wildcards. In *COLING : Companion volume : Demonstrations*, pages 181–184.

V. Seretan, L. Nerima, and E. Wehrli. 2003. Extraction of Multi-Word Collocations Using Syntactic Bigram Composition. In *Proceedings of the 4th International Conference on Recent Advances in NLP (RANLP2003)*, pages 424–431.

J. da Silva, G. Dias, S. Guilloré, and J. Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Progress in Artificial Intelligence*, pages 849–849.

F. Smadja. 1993. Retrieving collocations from text : Xtract. *Computational Linguistics*, 19 :143–177.

O. Vechtomova. 2005. The role of multi-word units in interactive information retrieval. In D.E. Losada and J.M. Fernández-Luna, editors, *ECIR 2005, LNCS 3408*, pages 403–420. Springer-Verlag, Berlin.

P. Watrin. 2007. Collocations et traitement automatique des langues. In *Actes du 26e Colloque international sur le lexique et la grammaire*, pages 1530–1536.