Multilingual Aspects of Multiword Lexical Units

Gaël Dias
José Gabriel Pereira Lopes
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Monte da Caparica, Portugal, 2725-114
{ddg,gpl}@di.fct.unl.pt

Sylvie Guilloré
Laboratoire d'Informatique Fondamentale
d'Orléans - U.F.R. Sciences
B.P. 6102 Cedex 02
Orléans, France, 45061
sylvie.guillore@lifo.univ-orleans.fr

Abstract

As most of the machine-readable dictionaries contain clearly insufficient information about multiword lexical units, there is a constant need to extend and tune specialized lexical databases to account for new expressions. In this paper, we present a system exclusively based on statistics that massively extracts from unrestricted text corpora contiguous and non-contiguous rigid multiword lexical units. For that purpose, a new association measure called the Mutual Expectation is conjugated with a new acquisition process based on an algorithm of local maxima. The system has been applied to a Portuguese, French, English and Italian parallel corpus and has evidenced that multiword lexical units embody a great deal of cross-language regularities.

Introduction

The acquisition of multiword lexical units (MWUs) has long been a significant problem in natural language processing. As most of the lexicons contain clearly insufficient information about multiword lexical units[1], there is a constant need to extend and tune specialized lexical databases to account for new expressions. As a consequence, the automatic extraction of multiword lexical units from corpora is an important issue not only for applications in Natural Language Processing but also for most of the systems of Information Retrieval, Information Extraction and Machine Translation. For the past fifteen years, the growing amount of text corpora available in machine-readable format has lead to new propositions for the treatment of fixed expressions emphasizing the evolution from formalisms towards lexicalization, that is the evolution from "general" rules towards rules specifying the usage of words on a case-by-case basis, such as in Gazdar (1985), Abeillé (1993) and Habert (1995).

From a statistical point of view, multiword lexical units are groups of words that occur together more often than expected by chance. Compound nouns like *European Social Fund*, compound verbs like *to take into account*, adverbial locutions like *as soon as possible*, prepositional locutions like *as defined in*, conjunctive locutions like *so that* and frozen forms like *inter alia* share the properties of multiword lexical units.

The research community has adopted three distinct policies in order to retrieve MWUs. Some approaches only extract contiguous multiword lexical units (i.e. uninterrupted sequences of words) requiring language-dependent information such as part of speech tags and basing their analysis on syntactical regularities like in Dagan (1994) and Bourigault (1996) or linguistic resources such as dictionaries like in Blank (1998). In order to scale up the acquisition process, other language-dependent approaches combine shallow morpho-syntactic information with statistics in order to evidence syntactical regularities and select the most probable candidate sequences of words like in Enguehard (1993), Justeson (1993), Daille (1995), Herviou (1996) and Feldman (1998). Finally, some purely statistical approaches propose language-independent techniques for

[1] Two exceptions are the BBI Combinatory Dictionary of English of Benson (1986) and the DELAC and DELACS of Silberztein (1990).

the extraction of contiguous and non-contiguous (i.e. fixed sequences of words interrupted by one or several gaps filled in by interchangeable words) multiword lexical units. They evidence regularities by means of association measure values that evaluate the mutual attraction existing between words in a sequence like in Church (1990), Smadja (1993), Chenxiang (1997) and Shimohata (1997).

In this paper, we propose a system based exclusively on a statistical methodology that retrieves from naturally occurring text, contiguous and non-contiguous multiword lexical units. In order to extract MWUs, a new association measure proposed by Dias-1 (1999) and based on the concept of normalized expectation, the Mutual Expectation is conjugated with a new multiword lexical unit acquisition process based on an algorithm of local maxima, the LocalMax algorithm introduced by Silva (1999). The proposed approach copes with two major problems evidenced by all previous works in the literature: the definition of unsatisfactory association measures and the ad hoc establishment of association measure thresholds used to select MWUs among word groups.

In the first section of this paper, we define the notion of multiword lexical units on a statistical basis. In the second section, we propose the transformation process of the input text corpus into contingency tables by counting contiguous and non-contiguous *n*-grams so that the data can suit to the purpose of statistical analysis. In the third and fourth sections, we respectively present the Mutual Expectation measure and the LocalMax algorithm for the election of MWUs. In the fifth section, we compare the Mutual Expectation with five other association measures and present the results obtained from the application of the mutual expectation with the LocalMax algorithm over a Portuguese, French, English and Italian parallel corpus.

1 Statistical Specification of MWUs

Most of the studies presented so far in the literature (cf. Justeson (1993), Dagan (1994), Daille (1995), Bourigault (1996) and Blank (1998)) concentrate on the specific area of terminology that deals essentially with the extraction of multiword lexical units of nominal type (i.e. terms). Terms can be considered as nominal compounds if they inherit well-known morphological and syntactical properties that have been stressed by studies on nominal compounding. However, multiword lexical units vary tremendously in the number of words involved, in the syntactic categories of the words, in the syntactic relations between the words and in how rigidly the individual words are used together. Consequently, their study has to take into account a more variegate set of linguistic phenomena than just the case of compound nouns. Compound verbs, adverbial locutions, prepositional locutions, conjunctive locutions and frozen forms also share the properties of multiword lexical units. Taking into account these observations, a linguistic specification of multiword lexical units seems to be a never-ending task. So, we propose a statistical specification of multiword lexical units based on the concept of collocations that has usually been misused in the literature about statistical extraction of lexical information. There has been a great deal of theoretical and applied works that have resulted in different characterizations of collocations as they depend on the aspects researchers are focusing at their studies (cf. Hausmann (1979), Cowie (1981) and Benson (1989)). An important comprehensive definition can be found in the work of Benson (1989) who defines a collocation as an "... arbitrary recurrent word combination". However, this definition is not sufficient for statistical analysis as it suggests frequency as the only relevant factor for the extraction of collocations. Smadja (1993) introduces the essential notion of plausibility for statistical analysis and defines a collocation as "... a recurrent combination of words that co-occur more often than expected by chance and that correspond to arbitrary word usage". But, this definition of collocation can not be used for statistical analysis, as statistical methods presented so far can not guarantee the arbitrariness of a group of words. Therefore, we focus on the difference between multiword lexical units and collocations as being the fact that a multiword lexical unit is a group of words that occur together more often than expected by chance.

Multiword lexical units have three properties. First, multiword lexical units do not embody exceptions as they are either frequently used or locally concentrated in a language or sub-language. For instance, *European Community* is widely and frequently used in the corpus of the European Parliament and the word combination *renewable energy sources* does not occur frequently in the corpus but its occurrences appear in only one paragraph thus referring to one particular important concept. Then, a multiword lexical unit is a group of words that occur together more often than expected by chance and as a consequence there exists a high level of cohesiveness between each word of the unit characterized by some kind of attraction between its components. For example, the two components of *Human Rights* are highly related as the presence of *Human* strongly suggests the occurrence of *Rights* and vice versa. Finally, the third property is of great importance when discussing about automatic extraction of lexical information. Indeed, multiword lexical units are domain-specific. Therefore, words that do not participate in a multiword lexical unit in one sub-language may be part of a unit in another sub-language. For instance, *data analysis* is clearly a multiword lexical unit when discussing about Statistics but surely is not when discussing about History.

We classify multiword lexical units into three types according to their structures. The first basic structure corresponds to contiguous multiword lexical units defined by uninterrupted sequences of words such as *single market* or *official languages*. The second structure corresponds to non-contiguous multiword lexical units that consist of fixed sequences of words interrupted by one or several gaps filled in by interchangeable words. For instance, *the _____ European Council* is a non-contiguous multiword lexical unit where the gap is likely to be filled in by names like *Lisbon* or *Luxembourg*. The last structure defines flexible multiword lexical units that correspond to free sequences of words. For example, *to be responsible for* is a flexible multiword lexical unit as it can be found in text in the form *to be successfully responsible for* or *to be for a long time responsible for*.

2 Data Preparation

According to Justeson (1993), the more a sequence of words is fixed, that is the less it accepts morphological and syntactical transformations, the more this sequence is likely to be a multiword lexical unit. Therefore the system does not modify the input text corpus by introducing any extra linguistic information.

Moreover, Smadja (1993) highlights that there is strong lexicographic evidence that most lexical relations associate words separated by at most five other words[2]. Therefore, multiword lexical units are specific contiguous or non-contiguous *n*-grams in a window of ten words long (i.e. five to the left of the considered word and five on its right hand side).

So, the first step of the system is to build all contiguous and non-contiguous *n*-grams from the non-modified input text. As an example, if sentence (1) is the current input and w_1 =*Maastricht* is the word under study, one non-contiguous and one contiguous *2*-gram containing w_1 are shown in Table 1.

(1) "*After difficult negociations, the Maastricht Treaty has been modified by all the State members.*"

Table 1: Two *2*-grams retrieved from sentence (1) containing *Maastricht*

w_1	$position_{12}$[3]	w_2
Maastricht	-3	*negociations*
Maastricht	+1	*Treaty*

By definition, MWUs are highly related groups of words, characterized by some kind of attraction between its components. Therefore, in order to investigate the suspected relationships between words, an *n*-dimensions contingency table is built for each *n*-gram providing a convenient display of the data for analysis. In

[2] Mason (1997) suggests that lexical relations involving one word vary in terms of word length. So, ideally there should exist a different span for every word under study.

[3] In Table 1, $position_{12}$ is the signed distance between w_1 and w_2. The sign "+" ("-") is used for words on the right (left) of w_1 .

order to build the contingency tables, we first need to define theoretically a Probability Space (Ω, $\mathcal{A}$, P[.]) where Ω is the Domain space, $\mathcal{A}$ the Event space and P[.] the Probability function:

- The event space $\mathcal{A}$ maps to each word w_i a binary discrete random variable X_{ip} that takes the value "1" if the word w_i appears in an *n*-gram at position[4] *p* and "0" if not.
- The Domain space Ω is the collection of all possible outcomes of a conceptual experiment over the instance space and is therefore defined as $\Omega=\{0, 1\}$.
- A good approximation for the Probability function P[.] is defined as the number of successes for a particular outcome divided by the number of instances.

The set of all *n*-grams built from the input text represents the instance space of the system and each *n*-gram provides a new independent Bernoulli trial for every variable X_{ip}. For example, if we take the random discrete variable X_{ip} which maps the word w_i =*Treaty* and the position p=+1, the outcome of the trial for the first *2*-gram of Table 1 is "0" and for the second *2*-gram is "1".

We may now build the contingency tables. For comprehension purposes, we only detail the case of the *2*-grams involving the definition of a two-dimension contingency table for each *2*-gram. We can define a *2*-gram as being a generic triplet [w_1 p_{12} w_2] where w_1 and w_2 are two words and p_{12} denotes the signed distance that separates both words. As defined above, w_1 and w_2 are respectively mapped to two discrete random variables X_{1p} and X_{2k}[5] whose cohesiveness has to be tested in order to measure their attraction. A contingency table is defined as in Table 2 for each triplet [w_1 p_{12} w_2] of the instance space.

Table 2: A contingency table for *2*-grams

	X_{2k}	$\neg X_{2k}$	**Total**
X_{1p}	$f(w_1, p_{12}, w_2)$	$f(w_1, p_{12}, \neg w_2)$	$f(w_1)$
$\neg X_{1p}$	$f(\neg w_1, p_{12}, w_2)$	$f(\neg w_1, p_{12}, \neg w_2)$	$f(\neg w_1)$
Total	$f(w_2)$	$f(\neg w_2)$	N

where N is the number of words present in the input text, $f(w_1, p_{12}, w_2)$ is the frequency of w_1, w_2 occurring together at position p_{12}, $f(w_1, p_{12}, \neg w_2)$ is the frequency of w_1 occurring with words other than w_2 at position p_{12}, $f(\neg w_1, p_{12}, w_2)$ is the frequency of w_2 occurring with words other than w_1 at position p_{21}, $f(\neg w_1, p_{12}, \neg w_2)$ is the frequency of w_1, w_2 never occurring at position p_{12} , $f(w_1)$ and $f(w_2)$ are the respective marginal frequencies of w_1 and w_2 , $f(\neg w_1)$ and $f(\neg w_2)$ are respectively equal to N- $f(w_1)$ and N- $f(w_2)$.

3 The Mutual Expectation measure

By definition, multiword lexical units are groups of words that occur together more often than expected by chance. From this assumption, we define a new mathematical model to describe the degree of cohesiveness that stands between the words contained in an *n*-gram. The association measures presented so far in the literature (cf. Church (1990), Gale (1991), Smadja (1993), Dunning (1993), Smadja (1996) and Silva (1999)) are not satisfactory as they only evaluate the degree of cohesiveness between two sub-groups of an *n*-gram. Futhermore, as they rely too much on the marginal probabilities of the word occurrences, they miscalculate the cohesiveness values. In order to overcome both problems, we present the Mutual Expectation measure (ME) based on the Normalized Expectation (NE) and introduced by Dias-1 (1999).

3.1 Normalized Expectation

We define the normalized expectation existing between *n* words as the average expectation of the occurrence of one word in a given position knowing the occurrence of the other *n*-1 words also constrained by their positions. For example, the average expectation of the *3*-gram [*Council +1 of +2 Ministers*] must take into account the expectation of occurring *Ministers* after *Council*

[4] The position *p* is the position of the word w_i in relation with the first word of the *n*-gram.

[5] Positions *p* and *k* must satisfy the constraint imposed by p_{12} that the two words occur together at the signed distance p_{12}.

of, but also the expectation of the preposition *of* linking together *Council* and *Ministers* and finally the expectation of occurring *Council* before *of Ministers*. This situation is graphically illustrated in Table 3 where one possible expectation corresponds to one respective row.

The basic idea of the normalized expectation is to evaluate the cost, in terms of cohesiveness, of the possible loss of one word in an *n*-gram. The more cohesive a word group is, that is the less it accepts the loss of one of its components, the higher its normalized expectation will be.

Table 3: Example of expectations to take into account in order to evaluate the NE

Expectation to occur the word	**Knowing the gapped 3-gram**
Council	[_____ *+1 of +2 Ministers*]
of	[*Council +1* _____ *+2 Ministers*]
Ministers	[*Council +1 of +2* _____]

The underlying concept of the normalized expectation is based on the conditional probability defined in Equation (1). The conditional probability measures the expectation of the occurrence of the event $X=x$ knowing that the event $Y=y$ stands. $p(X=x,Y=y)$ is the joint discrete density function between the two random variables X, Y and $p(Y=y)$ is the marginal discrete density function of the variable Y.

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad (1)$$

As defined in the previous section, each word of the text corpus is mapped to a discrete random variable in the Probability Space (Ω, $\mathcal{A}$, P[.]). Consequently, the definition of the conditional probability can be applied in order to measure the expectation of the occurrence of one word in a given position knowing the occurrence of the other *n*-1 words also constrained by their positions. However, this definition does not accomodate the *n*-gram length factor. For example, Table 3 clearly points at three possible conditional probabilities for a *3*-gram. Naturally, an *n*-gram is associated to *n* possible conditional probabilities. It is clear that the conditional probability definition needs to be normalized in order to take into account all the conditional probabilities involved by an *n*-gram.

One way to solve the normalization problem is to measure the Fair Point of Expectation (FPE). In order to perform the normalization process, it is convenient to evaluate the gravity center of the denominators of all the possible conditional probabilities thus defining an average event called the fair point of expectation. Basically, the FPE is the arithmetic mean the *n* joint probabilities[6] of the (*n*-1)-grams contained in an *n*-gram. In other words, the FPE is defined as the average point of expectation embodying all the particular points of expectation, thus reducing the *n* particular points of expectation to just one average point. The FPE for an *n*-gram is defined in Equation (2).

$$FPE([w_1\, p_{12}\, w_2 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = \frac{1}{n}\left(p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]) + \sum_{i=2}^{n} p\left(\left[w_1 \ldots \hat{p}_{1i}\ \hat{w}_i \ldots p_{1n}\ w_n \right]\right)\right) \qquad (2)$$

$p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n])$, for i=3,...,n, is the probability of the occurrence of the (*n-1*)-grams $[w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]$ which is the result of the extraction of w_1 from the whole *n*-gram and $p\left(\left[w_1 \ldots \hat{p}_{1i}\ \hat{w}_i \ldots p_{1n}\ w_n \right]\right)$ is the probability of the occurrence of one (*n-1*)-gram containing necessarily the first word w_1. The "^" corresponds to a convention frequently used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from 1 to n.

Hence, the normalization of the conditional probability is realized by the introduction of the fair point of expectation into the general definition of the conditional probability. The symmetric resulting measure is called the normalized expectation and is proposed as a "fair" conditional probability. It is defined in Equation (3).

[6] In the case of *n*=2, the FPE is the arithmetic mean of the marginal probabilities.

$$NE([w_1 ... p_{1i}\ w_i ... p_{1n}\ w_n]) = \frac{p([w_1 ... p_{1i}\ w_i ... p_{1n}\ w_n])}{FPE([w_1 ... p_{1i}\ w_i ... p_{1n}\ w_n])} \quad (3)$$

$p([w_1...p_{1i}w_i...p_{1n}w_n])$ is the probability of the *n*-gram $[w_1...p_{1i}w_i...\ p_{1n}\ w_n]$ occurring among all the other *n*-grams and $FPE([w_1...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n])$ is the fair point of expectation defined in Equation (2).

3.2 Mutual Expectation

Daille (1995) shows that one effective criterion for multiword lexical unit identification is simple frequency. From this assumption, we pose that between two *n*-grams with the same normalized expectation, that is with the same value measuring the possible loss of one word in an *n*-gram, the most frequent *n*-gram is more likely to be a multiword unit. So, the Mutual Expectation between *n* words is defined in Equation (4) based on the normalized expectation and the simple frequency.

$$ME([w_1...p_{1i}\ w_i\ ...\ p_{1n}\ w_n]) = f([w_1...p_{1i}\ w_i\ ...\ p_{1n}\ w_n]) \times NE([w_1...p_{1i}\ w_i\ ...\ p_{1n}\ w_n]) \quad (4)$$

$f([w_1\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n])$ and $NE([w_1\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n])$ are respectively the absolute frequency of the particular *n*-gram $[w_1\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$ and its normalized expectation.

4 The LocalMax Algorithm

Being the association measure value associated to each *n*-gram, the only feature available to the system in order to extract multiword lexical unit candidates, most of the approaches proposed in the literature have based their selection process on association measure thresholds like in Church (1990), Daille (1995), Smadja (1996) and Shimohata (1997). This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether an *n*-gram is a multiword lexical unit or not. However, these thresholds can only be justified experimentally and so are prone to error. Moreover, the association measures tend to favor certain properties of the multiword lexical units and as a consequence, the coarse grain threshold methodology may reject unjustifiably potential expressions in the set of all valued *n*-grams. Finally, the thresholds may vary with the type, the size and the language of the document and vary obviously with the association measure. The LocalMax algorithm proposed by Silva (1999), based on local maxima association measure values, proposes a more robust, flexible and fine-tuned approach for the election of multiword lexical units.

The LocalMax algorithm elects the multiword lexical units from the set of all the cohesiveness-valued *n*-grams based on two assumptions. First, the association measures show that the more cohesive a group of words is, the higher its score[7] will be. Second, multiword lexical units are highly associated localized groups of words. From these two assumptions, we may deduce that an *n*-gram is a multiword lexical unit if the degree of cohesiveness between its *n* words is higher or equal than the degree of cohesiveness of any sub-group of *(n-1)* words contained in the *n*-gram and if it is strictly higher than the degree of cohesiveness of any super-group of *(n+1)* words containing all the words of the *n*-gram. As a consequence, an *n*-gram, let's say *N*, is a multiword lexical unit if its association measure value, *val(N)*, is a local maximum. Let's define the set of the association measure values of all the *(n-1)*-grams contained in the *n*-gram *N*, by Ω_{n-1} and the set of the association measure values of all the *(n+1)*-grams containing the *n*-gram *N*, by Ω_{n+1}. The LocalMax algorithm is defined as follows:

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$
    if N=2 then
        if $val(N) > val(y)$ then
            $val(N)$ is a local maximum
    else
        if $val(x) \leq val(N)$ and $val(N) > val(y)$ then
            $val(N)$ is a local maximum

The LocalMax algorithm avoids the ad hoc definition of any global association measure threshold and focuses on the identification of local variations of the association measure values. This methodology overcomes the problems of reliability and portability of the

[7] The conditional entropy measure is one of the exceptions.

previously proposed approaches. Indeed, any association measure that shares the first assumption (i.e. the more cohesive a group of words is, the higher its score will be) can be tested on this algorithm.

5 Evaluation of the Results

In a preliminary study, we compare the results obtained by applying the LocalMax algorithm with five association measures, including the Mutual Expectation, to a Portuguese, English, French and Italian parallel corpus of political debates taken from the European Parliament debates collection with approximately 300000-words for each language. Then, we detail the results obtained with the four languages for the particular case of the Mutual Expectation.

5.1 Comparison between Five Association Measures

We applied the LocalMax algorithm with the normalized Association Ratio (N_AR)[8], the normalized Dice Coefficient (N_DC)[9], the normalized Φ^2 (N_PHI)[10], the normalized Log-likelihood Ratio (N_LOG)[11] and the mutual expectation (ME) on the Portuguese, English, French and Italian parallel corpus and compared the results.

There is no consensus among the research community about how to evaluate the output of multiword lexical unit extraction systems. Indeed, the quality of the output strongly depends on the task being tackled, as a lexicographer or a translator may not evaluate the same results in the same manner. A precision measure should surely be calculated in relation with a particular task. However, in order to define some "general" rule to measure the precision of the system, we propose the following two assumptions. Multiword units are valid units if they are grammatically appropriate units (by grammatically appropriate units we refer to compound nouns/names, compound verbs, prepositional/adverbial/conjunctive locutions) or if they are meaningful units even though they are not grammatical. Besides, the evaluation of extraction systems is usually performed with the well-known recall rate. However, we do not present the "classical" recall rate in this experiment due to the lack of a reference corpus where all MWUs are identified. Instead, we present the extraction rate, a measure of coverage, defined as the percentage of well-extracted MWUs in relation with the size of the corpus (by well-extracted we mean that the extracted MWUs are precise according to the definition of precision). The comparative results of precision rate and extraction rate, for the five association measures in the four languages, are respectively illustrated in Fig.1 and Fig.2.

Independently from the language under analysis, the Mutual Expectation shows significant improvement in terms of precision and reveals a high extraction rate in relation with all the other measures. The most important drawback that we can express against all the measures presented by the four other authors is that they raise the typical problem of high frequency words as they highly depend on the marginal probabilities. Indeed, they underestimate the degree of cohesiveness when the marginal probability of one word is high. For instance, the N_AR, the N_DC, the N_PHI and the N_LOG elect the multiword lexical unit *Code _____ Practice* although the probability that the preposition *of* fills in the gap is one. In fact, the following *3*-gram [*Code +1 of +2 Practice*] gets unjustifiably a lower value of cohesiveness than the *2*-gram [*Code +2 Practice*]. Indeed, the high frequency of the preposition *of* underestimates the cohesiveness value of the *3*-gram. On the opposite, the ME elects the MWU *Code of Practice,* as it does not depend on marginal probabilities except for the case of *2*-grams. So, all the non-contiguous multiword lexical units extracted with the mutual expectation measure define correct units as the gaps correspond to the occurrences of at least two different tokens. The problem shown by the other measures is

[8] The N_AR is the application of the fair point of expectation methodology to the association ratio introduced by Church (1990).

[9] The N_DC is the application of the fair point of expectation methodology to the dice coefficient introduced by Smadja (1996).

[10] The N_PHI is the application of the fair point of expectation methodology to the Pearson's coefficient introduced by Gale (1991).

[11] The N_LOG is the application of the fair point of expectation methodology to the Log-likelihood ratio introduced by Dunning (1993).

illustrated by low precision rates. The results shown in Fig.1 and Fig.2 confirm the general tendency of retrieval systems that the higher the extraction rate is, the lower the precision rate will be, with the only exception of the mutual expectation measure. Indeed, the ME shows the best precision rate and the second best extraction rate. The results also illustrate regularities in the ranking of the association measures independently of the language under study. Indeed, in terms of precision, we could order, across languages, the association measures from the more precise to the less precise as follows: ME, N_PHI, N_AR, N_DC and N_LOG (See Fig.1).

Fig. 1. Comparative precision rate

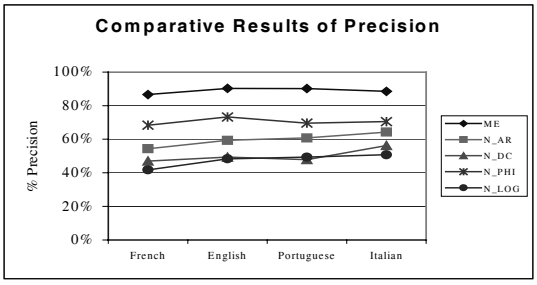
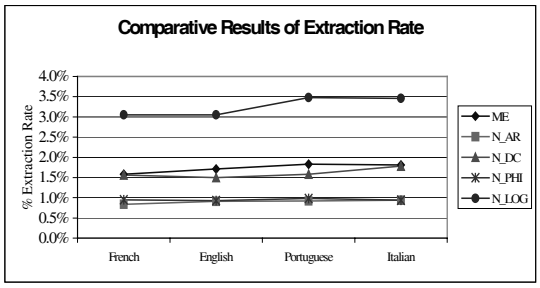


Fig. 2. Comparative extraction rate



Similarly, the following order could be set for the extraction rate: N_LOG, ME, N_DC, N_PHI and N_AR (See Fig.3). Our experiment clearly evidences that no association measure tends to favor any language in particular.

Moreover, all association measures tend to follow similarly each language characteristics. For the case of the extraction rate (See Fig.2), all curves show an identical shape illustrating that all models incline to extract more multiword lexical units for Italian than for any other language and retrieve fewer expressions for French than for any other language. For the case of the precision rate (See Fig.2), the curves of the N_AR, the N_DC and the N_LOG vary similarly tending to be more precise for Italian than for any other language and oppositely less precise for French than for any other language. The ME and the N_PHI have slightly different behaviors although they show similar variations along with languages. Indeed, both measures show the worst precision rate for French as the other measures do.

All these observations made about similar cross-language behaviors between association measures, in terms of precision and extraction rate strengthen the idea that the concept of multiword lexical unit can be "universally" pictured by means of association measure value regularities.

5.2 Comparison between Four Languages

In this section, we detail the results obtained for the particular case of the Mutual Expectation for the four languages. Contiguous multiword lexical units (CMWUs) and non-contiguous rigid multiword lexical units (NCMWUs) have been extracted. In the case of the extracted NCMWUs, we analyzed the results obtained for units containing exactly one gap leaving for further study the analysis of all the units containing two or more gaps. Indeed, the relevance of such units is difficult to judge and a case by case analysis is needed. However, the reader may retain the basic idea that the more gaps there exists in a MWU the less this unit is meaningful and the more it is likely to be an incorrect multiword lexical unit.

5.2.1 Qualitative Results

The extracted CMWUs can be classified into four types independently of the language (See Table 4): noun phrases (NP) such as *Council of Ministers*, verbal lexical units (VP) such as *to comply with*, prepositional/conjunctive/adverbial locutions (LOC) such as *as soon as possible* and prepositional/relative/coordination structures (STR) such as *in the* or *by the*. Similarly, the extracted non-contiguous MWUs can also be classified into four different types independently of the language: noun phrases such as *recent*

_____ *incidents* where the gap may be filled in by *violent* or *appalling*, verbal phrases such as *to* _____ *CO2 emissions* where possible instances to fulfill the gap are *reduce* or *limit*, syntactical structures (STR) such as the NP structure *the* _____ *of* and templates (TEMP) which represent long idiomatic domain dependent phrases such as *Written Question nº* _____ *by* where the gap may be fulfilled by any number.

The analysis of Table 4 assesses and enlarges to the case of French, Portuguese and Italian, Justeson's remark (1993) that more than 70% of technical terms are multiword lexical units. Indeed, for all four languages, more than 70% of the extracted multiword lexical units are noun phrases. Moreover, all languages show astonishingly similar distributions among categories. In the case of French, English and Italian the multiword lexical units are preferably noun phrases and verbal units. For the case of Portuguese, the results show that the acquisition process elects more locutions than verbal units. However, we believe that the specific domain of legislation may originate the latter result and we stress that the comparative results obtained confirm the study presented by Abeillé (1993) who postulate that multiword lexical units embody general grammatical rules.

Table 4. Classification of extracted multiword lexical units

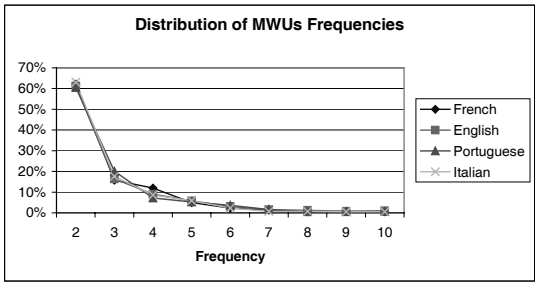
	CMWU			
	%NP	%VP	%LOC	%STR
French	73.43	16.25	7.18	3.14
English	73.26	18.60	6.28	1.86
Portuguese	74.48	9.49	11.27	4.76
Italian	76.21	13.85	6.93	3.01

	NCMWU			
	%NP	%VP	%STR	%TMP
French	59.35	23.22	10.97	6.46
English	58.41	21.78	7.93	11.88
Portuguese	65.32	16.13	16.13	2.42
Italian	64.88	21.37	9.16	4.59

5.2.2 Quantitative Results

Another important cross-language result that we obtained during our experiments is the fact that most of the extracted multiword lexical units occur only twice in the corpus[12]. Fig.3 reveals a representation of the distribution of the MWUs frequencies for all the four languages and assesses Dunning's remark (1993) that texts are composed largely of rare events. Indeed, independently of the language, more than 60% of the extracted MWUs occur only twice in the corpus[13]. This particular result is due to the application of the LocalMax algorithm. Indeed, a localized analysis of association measure values avoids the problem of many statistical approaches that only elect frequent MWUs as they base their study on the foundational assumption that the events being analyzed must be relatively common.

Fig. 3. The Distribution of MWUs Frequencies



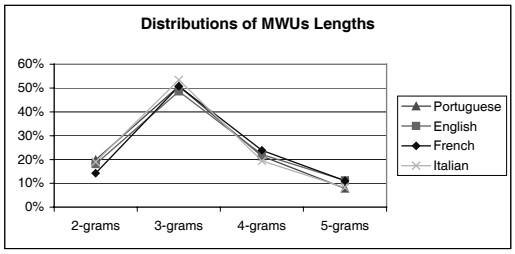
Another interesting result concerns the length of the multiword lexical units. Most of the studies on MWUs extraction rely on the definition of syntactical patterns thus coercing in some way the number of words of the extracted multiword lexical units. However, to our knowledge, no study has ever put forward a decisive result that would allow one to determine in advance the length of a multiword lexical unit. Fig.4 illustrates the distributions of the elected MWUs in terms of word length. The results clearly reveal that most of the multiword lexical units contain between two and four words independently from the language used. This result combined to the fact that most of the extracted MWUs are contiguous is of great interest for the window-based approaches that

[12] All the other expectation measures experimented showed the same result except for the case of the N_DC that preferably elects MWUs that occur three times in the corpus.

[13] The system does not elect one-frequency MWUs although Dias-2 (1999) points at a partial solution.

lack of foundational proves in order to set the size of the window used.

Fig. 4. The Distribution of MWUs Lengths



Conclusion

We proposed in this paper a language independent statistically-based system to automatically extract contiguous and non-contiguous rigid multiword lexical units from unrestricted text corpora. The method introduces a new association measure, the Mutual Expectation and a new multiword lexical unit acquisition process the LocalMax algorithm. We compared the mutual expectation with four other association measures, the normalized Association Ratio, the normalized Dice Coefficient, the normalized Φ^2 and the normalized Log-likelihood Ratio. The comparative results showed that the Mutual Expectation gives a higher precision than all other four measures. We also tried out our system on a Portuguese, French, English and Italian parallel corpus and the results highlighted the fact that the concept of multiword lexical unit embodies a great deal of cross-language regularities beyond grammatical and flexibility constraints, namely occurrence and length distribution consistencies. We hardly believe that the success of applications in the areas of Machine Translation, Information Retrieval, Cross-Language Information Retrieval and Information Extraction will rely on the preprocessing of corpora in order to benefit from their intrinsic information. Indeed, for the specific case of Machine Translation, we manually evidenced that 33.34% of the Portuguese MWUs were translated into an English MWU, 37.01% into an Italian MWU and 45.79% into a French MWU. The extraction of implicit knowledge (knowledge about the language) such as multiword lexical units will enable more precise text processing and as a consequence will lead to an adequate normalization of texts in order to extract more cross-language explicit information (knowledge about the world).

References

Abeillé A. (1993) *Les nouvelles syntaxes: Grammaires d'unification et analyse du Français*, Armand Colin, Paris, France.

Benson M. (1986) *The BBI Combinatory Dictionary of English: a Guide to Word Combinations*, John Benjamins.

Benson M. (1989) *The Structure of the Collocational Dictionary*, International Journal of Lexicography.

Blank I. (1998) *Computer-Aided Analysis of Multilingual Patent Documentation*, In "Proceedings of the First LREC", pp. 765--771.

Bourigault D. (1996) *Lexter, a Natural Language Processing Tool for Terminology Extraction*, In "Proceedings of the 7th EURALEX International Congress".

Chengxiang Z. (1997) *Exploiting Context to Identify Lexical Atoms: a Statistical View of Linguistic Context*, cmp-lg/9701001, 2 Jan 1997.

Church K. (1990) *Word Association Norms Mutual Information and Lexicography*, Computational Linguistics, 16/1, pp. 23--29.

Cowie A. (1981) *The Treatment of Collocations and Idioms in Learners' Dictionaries*, Applied Linguistics, Vol. 11.

Dagan I. (1994) *Termight: Identifying and Translating Technical Terminology*, In "Proceedings of the 4th Conference on Applied Natural Language Processing", ACL Proceedings.

Daille B. (1995) *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*, The balancing act combining symbolic and statistical approaches to language, MIT Press.

Dias G. (1999), nº1, Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora, In "Proceedings of TALN'99", Cargèse, France.

Dias G. (1999), nº2, *A Comparative Study of Mathematical Models for Multiword Lexical Unit Extraction*, submitted to Revue Technique et Science Informatiques.

Dunning T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence*, Association for Computational Linguistics, 19/1.

Enguehard C. (1993) *Acquisition de Terminologie à partir de Gros Corpus*, In "Proceedings of Informatique & Langue Naturelle ILN'93", pp. 373--384

Feldman R. (1998) *Text Mining at the Term Level*, In "Proceedings of PKDD'98", Lecture Notes in AI 1510, Springer Verlag.

Gale, W. (1991) *Concordances for Parallel Texts*, In "Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora" Oxford, England.

Gazdar G. (1985) *Generalized Phrase structure Grammar*, Harvard University Press, Cambridge, MA.

Habert B. (1997) *Les linguistiques du Corpus*, Armand Colin, Paris, France.

Hausmann F. (1979) *Un dictionnaire des collocations est-il possible?*, Travaux de linguistique et de littérature, Vol. 17., pp. 187—195.

Herviou M.L. (1996) *Construction de terminologies: une chaîne de traitement supportée par un atelier intégrant outils linguistiques et statistiques*, Technical Report 96NO00018, EDF-DER.

Justeson J. (1993) *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text*, IBM Research Report, RC 18906 (82591) 5/18/93.

Mason O. (1997) The Weight of Words: an Investigation of Lexical Gravity, In "Proceedings of PALC'97".

Silberztein M. (1990) *Le Dictionnaire Electronique des Mots Composés*, Langue Française, Vol. 87, pp. 79--83.

Silva J. (1999) *A local Maxima Method and a Fair Dispersion Normalization for Extracting multiword units*, In "Proceedings of MOL'6", Orlando, USA.

Shimohata S. (1997) *Retrieving Collocations by Co-occurrences and Word Order Constraints*, In "Proceedings of ACL-EACL'97", pp. 476—481.

Smadja F. (1993) *Retrieving Collocations From Text: XTRACT*, Computational Linguistics, 19/1, pp. 143—177.

Smadja F. (1993) *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Association for Computational Linguistics, 22/1.