Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora

Gaël Dias[1], Sylvie Guilloré[2] and José Gabriel Pereira Lopes[1]

[1]Universidade Nova de Lisboa, FCT/DI
Quinta da Torre, 2825-114, Monte da Caparica, Portugal
{ddg,gpl}@di.fct.unl.pt
[2]Université d'Orléans, Laboratoire d'Informatique Fondamentale d'Orléans
BP 6102 - 45061, Orléans Cédex 2, France
sylvie.guillore@lifo.univ-orleans.fr

Abstract

Multiword units are groups of words that occur together more often than expected by chance in sub-languages. *Président de la République*, *Coupe du monde* and *Traité de Maastricht* are multiword units. Unfortunately, most of the machine-readable dictionaries contain clearly insufficient information about multiword units[i]. Therefore, their automatic extraction from corpora is an important issue not only for natural language processing but also for applications on Information Retrieval, Information Extraction and Machine Translation. In this paper, we propose a new extraction system based on a new association measure, the Mutual Expectation, and a new acquisition process based on an algorithm of local maxima, the LocalMax algorithm.

1. Introduction

The acquisition of multiword units (MWUs) has long been a significant problem in natural language processing, being relegated to the borders of lexicographic treatment. For the past fifteen years, there has been a renewal in phraseology due to full access to large-scale text corpora in machine-readable formats that allowed testing assumptions made about syntactical regularities and flexibility constraints. From a statistical point of view, multiword units are groups of words that occur together more often than expected by chance. Compound nouns (*Président de la République*), compound verbs (*mettre en oeuvre*), adverbial locutions (*dès que possible*), prepositional locutions (*en matière de*), conjunctive locutions (*ainsi que*) and frozen forms (*Jacques Delors*) share the properties of MWUs. In this paper, we present and access a system exclusively based on statistics for massively extracting, from raw text, contiguous MWUs (i.e. uninterrupted sequences of words) and non-contiguous rigid MWUs (i.e. sequences of words interrupted by one or more gaps that are filled in by a small number of interchangeable words). In order to extract MWUs, a new association measure based on the concept of normalized expectation, the Mutual Expectation (ME) is conjugated with a new multiword unit acquisition process based on a algorithm of local maxima, the LocalMax algorithm [Silva et al.99]. The proposed approach copes with two major problems evidenced by all previous works in the literature: the definition of unsatisfactory association measures

[i] Two exceptions are the BBI Combinatory Dictionary of English [Benson86] and the DELAC and DELACS [Silberztein90].

and the ad hoc establishment of global association measure thresholds used to select MWUs among word groups. In order to evaluate the quality of the results obtained, a comparison with four other association measures proposed in the literature (the association ratio [Church90], the Dice coefficient [Smadja96], the Φ^2 [Gale91] and the Log-likelihood ratio [Dunning93]) is performed once those measures are normalized in order to accommodate the MWU length factor.

In the first two sections of this paper, we propose the core of the system by respectively introducing the Mutual Expectation measure and the LocalMax algorithm. In the third section, we discuss the comparative results obtained by applying to a French corpus of political debates, the LocalMax algorithm with the four association measures listed above.

2. The Mutual Expectation Measure

The transformation of the input text corpus into contingency tables, by counting contiguous and non-contiguous *n*-grams, allows the definition of mathematical models (or association measures) to describe the degree of cohesiveness that stands between words. However, the association measures presented so far in the literature (cf. [Church90], [Gale91], [Dunning93], [Smadja93] and [Smadja96]) are unsatisfactory as they only evaluate the degree of cohesiveness between two discrete random variables and do not generalize for the case of *n* variables. Moreover, many of these association measures rely too much on the marginal probabilities thus misevaluating the attraction between words. In this section, we introduce the Mutual Expectation measure based on the normalized expectation and the fair point of expectation methodology that allows the generalization of the association measures for the case of *n* words.

We define the normalized expectation (NE) existing between *n* words as the average expectation of one word occurring in a given position knowing the presence of the other *n*-1 words also constrained by their positions. The underlying concept is based on the conditional probability defined in (1) where $p(X = x, Y = y)$ 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*.

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

Let's take the *n*-gram $[w_1\ p_{12}\ w_2\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$ where p_{1i}, for i=2,...,n, denotes the signed distance[ii] that separates word w_i from word w_1. In order to suit the definition of the NE, an *n*-gram can be considered as the composition of *n sub*-(*n-1*)-gram obtained from the *n*-gram by extracting one word at a time from it. This can be thought as giving rise to the occurrence of any of the following *n* events where the underline denotes the missing word from the *n*-gram:

$[\underline{\quad\quad} \ w_2\ p_{23}\ w_3\ ...\ p_{2i}\ w_i\ ...\ p_{2n}\ w_n]$,	word w_1 missing,
$[w_1\ \underline{\quad\quad}\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$,	word w_2 missing, ...
$[w_1\ p_{12}\ w_2\ p_{13}\ w_3 ... p_{1(i-1)}\ w_{(i-1)}\ \underline{\quad\quad}\ p_{1(i+1)}\ w_{(i+1)} ... p_{1n}\ w_n]$,	word w_i missing, ...
$[w_1\ p_{12}\ w_2\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1(n-1)}\ w_{(n-1)}\ \underline{\quad\quad}]$,	word w_n missing.

[ii] The sign "+" ("-") is used to represent words on the right (left) of w_1

In order to take into account all these events in just one measure, it is necessary to calculate an average conditional probability. This is realized by the fair point of expectation (FPE) which is defined in Equation (4), where $p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n])$, for i=3,...,n, is the probability of the occurrence of the (*n-1*)-gram $[w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]$ and $p\left(\left[w_1 \ldots \hat{p}_{1i}\ \hat{w}_i \ldots p_{1n}\ w_n\right]\right)$[iii] is the probability of the occurrence of one (*n-1*)-gram containing necessarily the first word w_1 of the *n*-gram.

$$FPE([w_1\, p_{12}\, w_2 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = \frac{1}{n}\left(p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]) + \sum_{i=2}^{n} p\left(\left[w_1 \ldots \hat{p}_{1i}\ \hat{w}_i \ldots p_{1n}\ w_n\right]\right)\right) \quad \textbf{(4)}$$

So, the NE of a generic *n*-gram is defined as being a "fair" conditional probability using the fair point of expectation and is defined in (5), where $p([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])$ is the probability of occurrence of the *n*-gram $[w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]$ and $FPE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])$ its normalized expectation.

$$NE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = \frac{p([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])}{FPE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])} \quad \textbf{(5)}$$

However, [Daille95] shows that one effective criterion for multiword unit identification is simple frequency. From this assumption, we deduce that between two *n*-grams with the same normalized expectation, the more frequent *n*-gram is more likely to be a multiword unit. So, the Mutual Expectation between *n* words is defined in (6) based on the normalized expectation, $NE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])$, and the simple frequency of the particular *n*-gram $[w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]$, $f([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])$.

$$ME([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = f([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) \times NE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) \quad \textbf{(6)}$$

3. The LocalMax Algorithm

Most of the approaches proposed for the extraction of multiword units are based on association measure thresholds (cf. [Church90], [Daille95] and [Smadja96]). This is defined by the underlying concept that there exits a limit association measure that allows to decide whether an *n*-gram is a multiword unit or not. But, these thresholds can only be justified experimentally and so are prone to error. Moreover, they may vary with the type, the size and the language of the document and vary obviously with the association measure. The LocalMax algorithm [Silva et al.99] proposes a more robust, flexible and fine tuned approach.

The LocalMax algorithm elects the multiword units from the set of all the valued *n*-grams based on two assumptions: all the association measures[iv] show that the more cohesive is a group of words, the higher its score will be, and multiword units are highly associated localized groups of words. As a consequence, an *n*-gram (let's name it *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)-gram* contained in the *n*-gram *N*, by Ω_{n-1} and the set of the association measure values of all the *(n+1)-gram* containing the *n*-gram *N*, by Ω_{n+1}. The LocalMax algorithm is defined as follows:

[iii] 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.

iv The conditional entropy measure is one of the exceptions.

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

One important property of the LocalMax algorithm is that it can be tested with any association measure as they all share the first assumption explained above. In our system, an *n*-gram is a MWU if and only if its ME value is a local maximum.

4. Evaluation of the Results

In this section, we present the results obtained by applying the LocalMax algorithm and the Mutual Expectation to a French corpus of political debates taken from the European Parliament debates collection, that contains approximately 300000 words. The results are compared with the ones obtained by applying to the same corpus the LocalMax algorithm with the normalized association ratio[v], the normalized Dice coefficient[vi], the normalized Φ^2 [vii] and the normalized Log-likelihood ratio[viii].

Contiguous and non-contiguous rigid multiword units have been extracted. In the case of the extracted non-contiguous rigid multiword units, we analyze 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 multiword unit the less this unit is meaningful and the more it is likely to be an incorrect multiword unit. The first results (See Table 1) show that 72.43% of the extracted units are contiguous and only 27.57% are non-contiguous. From this result one can acknowledge that the non-contiguous rigid multiword units are less expressive in this sub-language than are the contiguous rigid multiword units. Nevertheless, their average frequency is very similar to the one of the extracted contiguous multiword units showing that they do not embody exceptions and that they reveal interesting phenomena of the sub-language. The extracted contiguous multiword units can be classified into four types. 61% are noun phrases and most of them are representative of the domain such as *Parlement Européen*, *Fonds social européen* or *États membres*. 18% embody verbal multiword units such as *faire le point, mettre en oeuvre*, *dresser la liste* or *il y a*. 17% are prepositional/adverbial locutions such as *en matière de, en raison de* or *en conformité avec*. Finally, 4% are prepositional structures such as *de la*, *sur le* or *dans les*. The extracted non-contiguous rigid multiword units can be classified into two main types. 78% are meaningful multiword units where the gap is a generalization of the multiword unit meaning. For instance, the concept of *transport de* ____ *dangereuses* will be specified by filling in the gap with possible occurrences such as *substances* or *matières*. 22% are multiword units that embody a syntactical relation. The coordination structure is evidenced by units such as *de* _____ *et de* and *n°* _____ *et n°*, and the negation structure by the *2*-gram *ne* _____ *pas*. The latter are much more frequent than the former as they characterize patterns of the language and not only of the sub-language. In a second stage, we measured the precision of the results based on two assumptions: multiword

v The normalized association ratio is the result of the application of the fair point of expectation methodology to the association ratio introduced by [Church90].

vi The normalized Dice coefficient is the result of the application of the fair point of expectation methodology to the Dice coefficient used by [Smadja96].

vii The normalized Φ^2 is the result of the application of the fair point of expectation methodology to the Φ^2 used by [Gale91].

viii The normalized Log-likelihood ratio is the result of the application of the fair point of expectation methodology to the Log-likelihood ratio used by [Dunning93].

units are valid units if they are grammatically appropriate units (i.e. compound nouns/names, compound verbs, compound prepositions/adverbs/conjunctions and frozen forms) or if they are meaningful units even though they are not grammatical[ix]. In these conditions, the system shows a precision of 86,59%. Unfortunately, we do not present the "classical" recall rate in this experiment due to the lack of a reference corpus where all multiword units are identified. However, we present the extraction rate, which is the percentage of well-extracted multiword units in relation with the size of the corpus. It was evaluated at 1,58% (See Table1) referring to the corpus with 300000 words.

Table 1: Comparative results between five association measures

	Mutual Expectation	**Normalized Association Ratio**	**Normalized Dice Coefficient**	**Normalized Φ^2**	**Normalized Log-likelihood**
% of CMWU[x]	72.43	48.14	62.49	58.99	60.47
% of NCMWU[xi]	27.57	51.86	37.51	41.01	39.53
Average freq. CMWU	7.11	2.24	9.11	9.79	5.39
Average freq. NCMWU	7.12	2.14	8.78	3.64	4.23
Average length of MWU	3.32	3.33	2.06	2.85	2.36
% Precision	86.59	54.34	47.01	68.34	41.81
% Extraction rate	1.58	0,84	1.56	0.95	3.05

The results obtained by applying the LocalMax algorithm to the other association measures, compared with the ones obtained for the ME, show that the Mutual Expectation measure evidences significant improvements in terms of precision and correctness of the elected MWUs. The normalized association ratio makes rare word groups look more similar than they really are[xii] and as a consequence the average frequency of the extracted multiword units falls to 2.24 for the case of the CMWUs raising a weak extraction rate. Besides, almost no *2*-grams are extracted thus over-evaluating the average length of the units to 3.33 words. The normalized Dice coefficient gives good results in terms of coverage but shows one of the worst precision rate of all. Indeed, the Dice coefficient tends to elect exclusively *2*-grams showing an average length around 2.06 words. Moreover, the average-frequency rows of Table 1 highlight the fact that the normalized Dice coefficient tends to elect preferably frequent MWUs. The normalized Φ^2 shows one of the best precision rate of all, but its extraction rate is weak comparing to most of the other measures. Satisfyingly, it tends to elect a more variegate set of MWUs than most of the other measures as it is evidenced by an average length of 2.85. Finally, the normalized Log-likelihood ratio reveals the worst precision rate of all measures in contrast with its extraction rate that evidences the best result. But, similarly to the normalized Dice coefficient, it tends to elect exclusively *2*-grams showing an average length around 2.36 words. Besides, the precision of the elected *n*-grams, for *n* higher than 2, is weak causing the low precision rate result. But, the measures presented by the four other authors all raise the typical problem of high frequency words as they highly depend on the marginal probabilities. Indeed, they underestimate the degree of cohesiveness between words when the marginal probability of one variable (i.e. one word) is high.

ix This choice can easily be argued as a precision measure should be calculated in relation with a particular task. For instance, one may calculate the precision of the extracted multiword units for machine translation purposes, for information retrieval purposes or for lexicographic purposes.
x CMWU stands for Contiguous MultiWord Unit.
xi NCMWU stands for Non-Contiguous MultiWord Unit.
xii This confirms the results obtained by [Daille98].

5. Conclusion

We proposed in this paper a language independent statistically-based system to automatically extract contiguous and non-contiguous rigid multiword units from unrestricted text corpora. The method introduces a new association measure, the Mutual Expectation, and a new multiword unit acquisition process, the LocalMax algorithm [Silva et al.99]. The experiments realized on a 300000-words corpus of the legal domain evaluated the precision of the system at 86,59%. We compared the Mutual Expectation measure with four other normalized association measures, the association ratio [Church90], the Dice coefficient [Smadja96], the Φ^2 [Gale91] and the Log-likelihood ratio [Dunning93] by running the system in the four cases. The comparative results showed that the Mutual Expectation gives higher precision, overcomes the problem of highly frequent words raised by the four other measures and satisfyingly tends to elect longer multiword units. Finally, the system ensures total portability. It is applicable to various languages as it uses plain text corpora and requires only the general information appearing in it [Dias99].

References

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

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

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

Daille B. et al. (1998), An evaluation of Statistical scores for word association*, Tbilisi Symposium on Logic, Language and Computation* (CSLI)

Dias G. et al. (1999), Mutual Expectation and LocalMax Algorithm for Multiword Lexical Unit Extraction, *Paper submitted at EPIA'99*

Dunning T. (1993), Accurate Methods for the Statistics of Surprise and Coincidence, *ACL*, Vol. 19-1

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

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

Silva, J.F. and Lopes, J.G.P. (1999), A Local Maxima Method and a Fair Dispersion Normalization for Extracting multiword units, *In Proceedings of the 6th Meeting on Mathematics of Language (MOL6),* Orlando, Florida July 23-25, 1999

Smadja F. (1993), Retrieving Collocations From Text: XTRACT, *Computational Linguistics*, Vol. 19 (1)

Smadja F. (1996), Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Association for Computational Linguistics*, Vol. 22-1