Does Natural Selection Apply to Natural Language Processing? An Experiment for Multiword Unit Extraction

GAËL DIAS and SÉRGIO NUNES

Center of Mathematics
Beira Interior University, Covilhã, Portugal 6200-053

Abstract

In this paper, we focus on the suitability of natural selection for the extraction of Multiword Units (i.e. complex lexical units such as compound nouns, idiomatic expressions or phrase templates). For that purpose, a fitness function is defined whose maximization serves as a basis for the identification of pertinent word *N*-grams together with a similarity measure. In order to propose a suitable platform for evaluation, a software application called GALEMU (Genetic ALgorithm for the Extraction of Multiword Units) has been implemented. Finally, we will provide an experiment realized over an unannotated text corpus extracted from the database collection of the European Commission that evidences results with high precision rate.

Keywords
Multiword Unit Extraction, Genetic Algorithms, Similarity Measures.

1 Introduction

For the past ten years, there has been a renewal in phraseology mostly stimulated by full access to large-scale text corpora in machine-readable format. As a consequence, the evolution from formalisms towards lexicalization has lead to propose the hypothesis that the more a sequence of words is fixed, that is the less it accepts lexical and syntactical transformations, the more likely it should be a MWU. Compound nouns (*Human Rights*), compound names (*George W. Bush*), compound determinants (*a number of*), verbal locutions (*to give rise*), adverbial locutions (*as soon as possible*), prepositional locutions (*such as*) and conjunctive locutions (*on the other hand*) share the properties of MWUs.

Identifying MWUs in texts can be defined as a key step towards high-quality NLP applications. Indeed, it has been proved that Information Retrieval [8] and Machine Translation [9] applications would greatly benefit from advanced techniques capable of identifying important concepts in texts. In the context of Information Retrieval, selecting discriminating terms in order to represent the contents of texts is a critical problem. Ideally, the indexing terms should directly describe the concepts present in the documents. However, most of Information Retrieval systems index the documents based on individual words that are not specific enough to evidence discriminated information. In order to overcome this drawback, evolutionary retrieval systems [8] have been developed to represent the contents of texts using multiword terms previously extracted from text collection. In the context of Machine Translation, it is of common sense that MWUs are difficult to handle for non-native speakers. As a matter of fact, MWUs are indivisible lexical units in the sense that their meaning or function does not necessarily follow from the compositionality of the meaning of their component words. For example, the English compound noun *House of Commons* would not be translated on a word-by-word basis into French. Indeed, its correct French translation would be the MWU *Chambre des Communes* and not the concatenation of the translated words, i.e. *Maison des Communes*. As a consequence, most Machine Translation research groups have intensified their efforts in order to integrate modules that would be capable of identifying non-compositional complex lexical units in running texts [1] [9].

In this article, we present a tool designed to identify and extract MWUs from unrestricted text corpora. We named it GALEMU (Genetic ALgorithm for the Extraction of Multiword Units). GALEMU proposes an original architecture based on a floating point representation genetic algorithm and a set of different similarity measures. The basic idea of the application is simple. First, the text corpus to be analyzed is segmented into a set of positional *N*-grams (i.e. ordered vectors of *N* words) from which significant individuals will have to be identified. For that purpose, each positional *N*-gram is associated to a set of attribute values (e.g. frequency, degree of cohesion or size) thus representing a specific chromosome of the overall population. Then, a Genetic Algorithm will maximize a fitness function to provide the "best" genotype of the overall population. Finally, in order to extract relevant MWUs from the original population, a similarity measure will evidence the

Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference

relatedness between a specific positional *N*-gram in the population and the elected "best" individual. As a consequence, very close genotypes will be listed as pertinent word associations whereas unrelated chromosomes will be discarded.

In order to evaluate our methodology, some experiments have been realized over an unannotated corpus extracted from the European Commission data collection. As expected, compound nouns, names, determinants and verbal, adverbial, prepositional and conjunctive locutions have been extracted.

2 Related Work

In order to identify and extract MWUs, a great deal of distinct methodologies has been proposed. In this section, we will propose a rapid survey of the state-of-the-art of the statistical approach in order to enable the reader to outline the major differences proposed by our methodology compared to the ones illustrated in the literature. In this context, two typical methodologies can be evidenced. First, some studies propose the application of binary association measures to evaluate the degree of cohesiveness between two words [2] [7] [12] [20] [22]. As a consequence, bootstrapping techniques have to be applied to acquire associations with more than two words. A general algorithm of the methodology is given in Figure 1.

```
- Segment the input text into 2-grams
- Calculate the association measure
  value of each 2-gram
- Select as MWUs all 2-grams superior
  to a given association measure
  threshold
WHILE still MWUs to extract DO
  - Integrate as single words the
    selected 2-grams into the text
  - Segment the input text into 2-grams
  - Calculate the association measure
    value of each 2-gram
  - Select as MWUs all 2-grams
    superior to a given
    association measure threshold
END WHILE
```

Fig 1. The Bootstrapping Algorithm

Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams for the initiation of the iterative process [15]. In order to overcome the lack of generalization for *N* individual words, *N*-ary association measures have been proposed by the research community [11] [18] [19]. The basic idea is to evaluate the degree of cohesiveness of any sequence of words using a unique formula so that bootstrapping techniques can be avoided. The basic algorithm is then simplified as follows.

```
- Segment the input text into N-grams
- Calculate the association measure
  value of each N-gram
- Select as MWUs all N-grams superior
  to a given association measure
  threshold
```

Fig. 2: The *N*-ary Algorithm

However, two important remarks must be stressed out. On one hand, both methodologies rely on the definition of *ad hoc* association measure thresholds. This is defined by the underlying concept that there exists a limit value of the association measure that allows deciding whether an *N*-gram is a MWU or not. However, it is not clear whether the search space can be divided into two distinct subsets, one for the pertinent *N*-grams and one for the others, on the basis of one association measure value defined by the experimenter. We strongly believe that this coarse grain approach is not suitable for the specific task of MWU extraction. On the other hand, both approaches rely on a small number of computed information. In most cases, only one association measure is used to evidence relevancy of a sequence of words. However, [6] and [3] have shown that many applications might benefit from the combination of different information. In order to overcome both drawbacks, we propose an original idea that relies on two basic concepts: a Genetic Algorithm and a Similarity Measure. First, a floating point representation genetic algorithm processes a natural selection of the "best" positional *N*-gram (i.e. the typical MWU) among a population of attribute-valued sequences of words. In particular, various attributes are defined for each *N*-gram so that relevancy may not rely on one exclusive association measure. Then, a similarity measure evaluates the relatedness between each *N*-gram of the initial population and the typical selected MWU. As a consequence, very close *N*-grams are listed as pertinent word associations whereas unrelated ones are discarded on the basis of a confidence value. So, instead of defining global *ad hoc* thresholds, we propose an acquisition process based on similarity.

3 The Genetic Algorithm

A Genetic Algorithm (GA) is a stochastic algorithm whose search method models two basic natural phenomena: genetic inheritance and Darwinian strife for survival. In this context, a GA performs a multi-directional search over a sample space by maintaining a population of potential solutions and encourages information formation and exchange between individuals. As a consequence, the

population in consideration undergoes a simulated evolution so that, at each generation, the relatively "good" solutions reproduce and the relatively "bad" solutions die. In particular, as any evolution program, a GA must have the following five components:

- a genetic representation for potential solutions to the problem,
- a way to create an initial population of potential solutions,
- an evaluation function rating solutions in terms of fitness,
- genetic operators that alter the composition of children,
- values for various parameters that the GA uses (e.g. operator probabilities).

In this section, we will specifically focus on the genetic representation and the fitness function that we will use for our optimization problem. Indeed, unlikely the three other components whose techniques are generally well known and well established, problem representation and fitness play a key role for the success of GAs.

3.1 Floating Point Representation

The binary representation traditionally used in genetic algorithms has revealed some drawbacks when applied to multidimensional, high precision numerical problems. As a consequence, experiments have been realized for parameter optimization problems with real-coded genes together with specific genetic operators designed for them. On one hand, the conducted experiments indicate that floating point representation is faster, more consistent from run to run and provides better precision than the binary representation for large domains [17]. On the other hand, as intuitively closer to the problem space, the floating-point representation allows a one-gene-one-variable correspondence thus easing the codification process. Consequently, each chromosome can easily be represented as a vector of real numbers, each one corresponding to a specific variable of the problem. In the context of MWU extraction, we will define 7 variables that have been proposed in different studies as good heuristics for the identification of highly cohesive sequences of words.

Gene x_0: As evidenced in the previous section, association measures have been widely used in order to define the degree of cohesiveness of word *N*-grams. So, the more cohesive a word sequence is (i.e. the higher its association measure value is), the more likely it is a MWU. Thus, association measures are good heuristics for the identification of relevant word associations. In the specific context of positional *N*-grams, [6] have defined a new *N*-ary association measure called the Mutual Expectation (ME) that does not under-evaluate the degree of cohesion of sequences of words containing frequent single words. The intuitive idea of the ME is to evaluate the probability of appearance of each word in a *N*-gram conditioned by the appearance of the other *N-1*. So, our first gene will model the association measure of any given *N*-gram.

Gene x_1: Aside from association measures, frequency is considered by many researchers [3] [10] as an effective criterion for multiword unit identification. So, highly frequent word *N*-grams are more likely to be MWUs than infrequent ones. Consequently, we propose that the second gene of any *N*-gram individual should be its relative frequency (i.e. the quotient between its frequency and the total number of *N*-grams built from the input text for a given *N*).

Gene x_2: However, [10] demonstrate that stand-alone frequency can lead to error in the acquisition process. Let's consider the following word sequence: *soft contact lenses*. Suppose that its frequency is high enough so that it should be considered a candidate MWU. *A fortiori*, both 2-grams *soft contact* and *contact lenses* should also be regarded as potential MWUs. However, only the latter is a pertinent word association. This is due to the fact that while *contact lenses* can occur in the text by itself, *soft contact* will always appear within *soft contact lenses*. So, while the fact that an *N*-gram appearing in other longer *N*-grams (i.e. super-groups) is a negative factor for its relevancy, the word sequence increases in probability of importance (i.e. it increases in independence from its context) as the number of these longer *N*-grams increases. We will consider this number as our third variable-gene.

Gene x_3: Moreover, in the specific context of terminology, [5] evidence that complex terms (i.e. terminologically relevant MWUs) are specific lexical relations that favor the occurrence of infrequent single words in their core. Indeed, following the Zipf's law [24], very frequent words tend to be meaningless. So, the more an *N*-gram contains frequent single words, the less relevant it should be. As a consequence, for each positional *N*-gram, we evaluate the arithmetic mean of the frequencies of all its constituents in order to measure its relevancy. We will call it the marginal frequency. In this context, a high marginal frequency would induce irrelevancy. This measure will be our fourth gene.

Before going on with the definition of our 3 remaining genes, we will introduce the fitness function that our genetic algorithm will have to maximize. As a matter of fact, we will see that the remaining genes will only introduce constraints in our search problem.

3.2 Fitness Function

As mentioned earlier, a GA performs a simulated evolution over a population where relatively "good" solutions reproduce and relatively "bad" ones die from generation to generation. To distinguish between different solutions, we use an objective function, which plays the role of the environment. This function is called the fitness function. In the context of our research, we need to select pertinent individuals in terms of word associations among the set of attribute-valued positional *N*-grams. From the previous assumptions, a simple fitness function can directly be suggested. Indeed, a potential MWU is a particular *N*-gram with a high association measure, a high frequency, a high number of longer strings in which it appears and a small marginal frequency. A straightforward fitness function is thus defined as follows where *X* is a given chromosome.

$$g(X) = x_0 + x_1 + x_2 - x_3$$

However, extracting MWUs is a constrained task. As a consequence, aside from the definition of the fitness function, a specific set of inequalities will have to be defined.

3.3 Handling Constraints

As stated in [4], "A little observation and reflection will reveal that all optimization problems of the real word are, in fact, constrained problems". As a consequence, this assumption will lead to the introduction of three new genes that will be used to penalize infeasible solutions.

Gene x_4 and Gene x_5: In order to select potential MWUs from a set of association measure valued *N*-grams, [21] have proposed an original methodology that does not rely on global thresholds. The basic idea is simple. A positional *N*-gram is a MWU if its association measure value is higher or equal than the association measure values of all its sub-groups of (N-1) words (i.e. all the *N-1*-grams contained in it) and if it is strictly higher than the association measure values of all its super-groups of (N+1) words (i.e. all the *N+1*-grams containing it). So, for our optimization problem, the fifth and sixth genes of each individual will respectively be the highest Mutual Expectation value of all the sub-groups of the considered genotype and the highest Mutual Expectation value of all its super-groups. As a consequence, two constraints will directly be formulated.

$$x_0 \geq x_4$$
$$x_0 > x_5$$

Gene x_6: Finally, [11] propose that longer *N*-grams should be preferred to smaller ones. In particular, if the frequency of a given *N*-gram is equal to the frequency of a longer *N*-gram that contains it, the former should not be considered as a relevant word association. As a consequence, our seventh gene-variable will evidence the frequency value of the most frequent super-group of the considered individual. For that purpose, the following constraint will be formulated.

$$x_6 < x_1$$

3.4 Penalty Function

The definition of constraints implies the introduction of penalty functions whose goal is to penalize infeasible solutions. Indeed, if new individuals do not guarantee the constraints, they must be penalized in terms of fitness so that their probability to reproduce is lowered. As a consequence, the fitness function *g(X)* will be transformed in a more generic one called *eval(X)* defined as follows where $penal_k(X)$ is the penalty function attributed to the chromosome *X* when the constraint *k* is not verified.

$$eval(X) = \begin{cases} g(X), X \text{ feasible} \\ g(X) - \sum_k penal_k(X), \text{ otherwise} \end{cases}$$

In particular, as all the variable values have been normalized, the penalizations will not depend on scale problems. As a consequence, in order to penalize the fitness function, we will use the same methodology for each constraint. The penalty function will thus be based on the distance of the current value of the variable compared to its limit value. For example, for the constraint on gene x_4, the penalty function would be defined as follows.

$$penal_1(X) = \begin{cases} \frac{x_4 - x_0}{x_0}, \text{ if } x_4 > x_0 \\ 0, \text{ otherwise} \end{cases}$$

3.5 The Algorithm

Once the fitness function have been defined together with its constraints, the goal of the genetic algorithm is to find the values of each gene that together maximize the fitness function. The overall algorithm is illustrated in Figure 3.

```
- Build the initial population
  from the text corpus
- Random sample of the initial
  population
WHILE still generations to run DO
     - Evaluate population (fitness)
     - Stochastic selection of individuals
     - Crossover
     - Mutation
     - Apply Elitism
END
```

Fig 3. The Genetic Algorithm

4 Similarity Measure

The application of the GA over the initial population is likely to provide the "best" genotype that is supposed to evidence the "typical'" MWU. However, work still need to be done in order to identify pertinent word associations. For that purpose, we will use a similarity measure whose goal will be to evaluate the relatedness between each *N*-gram built from the initial population and the "typical" selected MWU. As a consequence, very close *N*-grams will be listed as pertinent word associations whereas unrelated ones will be discarded on the basis of a confidence value.

When variables are measured quantitatively, it is natural to evaluate similarity as a measure of distance. The basic idea is simple: the more distant two pairs of units are, the less similar they are. For that purpose, five different measures have been implemented in GALEMU based on [16]. We will access one of them that has been used for this particular experiment. Suppose *Xi=(Xi1, Xi2, ..., Xip)* a row vector of observations on *p* variables associated with a label *i*. The distance between two units *i* and *j* is defined as *Dij=f(Xi, Xj)* where *f* is some function of the observed values. The following function is called the Bray-Curtis Distance.

$$Dij = \frac{\sum_{k=1}^{p} \left| X_{ik} - X_{jk} \right|}{X_{ik} + X_{jk}}$$

In the context of our work, the application of these similarity (or distance) measures is straightforward. Indeed, *Xj* may be regarded as the elected chromosome and *Xi* as a particular individual of the initial population that will have to be compared with the "typical" MWU.

5 Experiments and Results

In order to evaluate our methodology, some experiments have been performed over a text corpus of approximately 200000 words extracted from the collection of debates of the European Commission. In particular, many experiments have been conducted using different genetic operators, different parameter values (e.g. number of generations, operator probabilities) and different similarity measures. However, it is not possible to summarize them all in this section. As a consequence, we will specifically focus, on the results obtained by applying different similarity measures thresholds to the acquisition process. In that context, we will propose a performance evaluation based on precision rate (# of correct MWUs / # of extracted MWUs) and spread (# of correctly extracted MWUs) depending on the considered thresholds. We will access that precision and spread combine together so that whenever spread increases, precision decreases.

It is usually difficult to determine whether a word association is a MWU or not. Many authors assess this problem [20] [22]. However, in order to evaluate our results, a clear definition of the phenomena embodied by the MWUs is required. In this context, one of the best linguistic studies of MWUs is proposed by [14]. In particular, G. Gross proposes a classification of the MWUs in 7 categories: compound nouns, names, determinants and verbal, adverbial, prepositional, conjunctive locutions. So, each *N*-gram that may be classified according to these guidelines will be considered as a correct MWU. In order to perform a homogeneous evaluation, we first define a typical set of parameters that is shown in Table 1.

Parameter	***Value***
Number of Generations	6000
Size of the Sample Population	100
Mutation Probability	0.01
Crossover Probability	0.6
Mutation Type	Non-uniform
Crossover Type	Uniform
Similarity Measure	Bray-Curtis
Error rates	{0.01, 0.003, 0.0007}

Table 1: Parameters of the Experiment

In this first part of our evaluation, we will analyze the results of the extraction on a qualitative basis. For that purpose, we will try to classify the elected *N*-

grams following [14]'s classification. First of all, compound nouns represent the greatest part of the extracted MWUs. Indeed, most MWUs are complex terms. Examples are shown in Table 2.

ME	*Rel. Freq.*	*MWU*
0.000001	0.00001	Bank loans
0.000001	0.00001	fresh fruit
0.000002	0.00001	Conservative Party
0.000007	0.00001	Greek trade unionists
0.000007	0.00001	Vance-Owen peace plan
0.000007	0.00001	two-wheel motor vehicles

Table 2: Extracted Compound Nouns

Due to its focus, the text also contains a wide variety of compound names that we illustrate in Table 3.

ME	*Rel. Freq.*	*MWU*
0.000002	0.000010	Allan Donelly
0.000001	0.000010	Bosnia Herzegovina
0.000003	0.000010	Financial Times
0.000010	0.000010	Roth- Behrendt
0.000011	0.000015	Mr Bofill Abeilhe
0.000007	0.000015	Professors Tangerman

Table 3: Extracted Compound Names

Another recurrent linguistic phenomenon in modern English is embodied by verbal locutions. Some examples are given in Table 4.

ME	*Rel. Freq.*	*MWU*
0.000001	0.000010	Giving rise
0.000001	0.000010	moving ahead
0.000005	0.000010	make an effort
0.000004	0.000030	to refer to
0.000017	0.000030	take place at
0.000009	0.000025	to draw attention

Table 4: Extracted Verbal Locutions

However, multiword units are not restricted to compound nouns, names and verbal locutions. Indeed, many other linguistic phenomena are characterized by MWUs, namely, compound determinants and adverbial, prepositional, conjunctive locutions as shown in Table 5. Nevertheless, it is clear that they count for a smaller portion of the MWUs than the other three categories.

ME	*Rel. Freq.*	*MWU*
0.000013	0.000040	so far as
0.000015	0.000060	as long as
0.000002	0.000010	in spite of
0.000001	0.000015	a lack of
0.000001	0.000010	All in all
0.000001	0.000010	in order for

Table 5: Extracted Locutions and Compound

Based on these qualitative results, we can easily propose, in this second part, a performance evaluation built on precision and spread. Thus, all the MWUs that will be classified following [14]'s list of categories will be counted as correct units. As a consequence, all the extracted *N*-grams that will not satisfy these requirements will be considered "incorrect" for the acquisition process. In these conditions of experimentation, we tested three different similarity measures thresholds that gave rise, as expected, to three different results as shown in Table 6.

Threshold	*Precision (%)*	*Spread (# of units)*
0.0007	61.74	326
0.003	37.31	1031
0.01	32.23	1238

Table 6: Performance Evaluation

As expected, the larger the error rate is (i.e. the further the potential MWU is from the typical MWU), the lower its precision rate and the larger its spread are. In order to compare our results with existing methodologies, we access the figures evidenced by two important studies made by [22] and [20]. First, [22] presents an experiment of XTRACT over a 10 million-word corpus of *The Jerusalem Post*. For that purpose, F. Smadja proposes three criteria to classify potential MWUs: YY (a very good quality MWU), Y (a MWU with less quality) and N (surely not a MWU). In these conditions, 40% precision rate is reached for the unannotated version of the corpus. However, we must point at the fact that the definition of the quality of a MWU is somewhat vague. After all, what is a good quality MWU? In these conditions, a comparison between our results and the ones illustrated by [22] is hazardous. In the same context, [20] proposes an experiment over a 1 million-word written-English corpus for which he looks for four different types of potential MWUs: complete sentences (CS), grammatical units (GU), meaningful units (MU) and functional fragments (F). In these conditions, S. Shimohata evidences a precision rate of 67%. Once again, the most important point of the evaluation resides in the fact that the explored phenomena are not well defined. Indeed, for example, the definition of meaningful units proposed by S. Shimohata is somewhat unclear. In fact, he considers as a meaningful unit, all the potential MWUs that may convey some sense even they are not grammatically correct! According to us, such a definition should be avoided as it leads to unjustified high performance. As evidenced above, the comparison between extractors is a difficult task as the material to be explored lacks of well-founded definitions. As a consequence, the results of precision and spread should only be used as estimators of the quality of the acquisition process in very restricted

conditions. It is clear that our results can not directly be compared to the ones evidenced by [22] and [20]. However, from our experience in the field, we strongly believe that the results obtained by our architecture are at least as good as the ones obtained from both studies considering precision and spread.

We must point at the fact that a complete evaluation should also take into account many parameters such as portability to new languages, scalability in terms of text size, adaptability to new text domains, external reviewing etc. However, the scope of the paper does not allow all this. An important criterion (among others) that should appear in a complete evaluation of extractors is the speed of the acquisition process. Indeed, the preoccupation to build performing tools is one of the most important challenges of this century for new tasks in Natural Language Processing. In this context, GALEMU (See Figure 5) performed outstandingly well comparing to other systems.

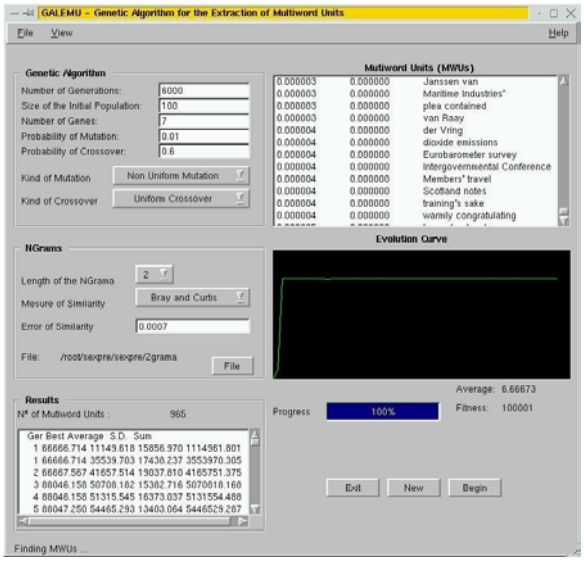

Fig. 5: GALEMU Application

All our experiments have been tested on a 166 Mhz Intel Pentium PC with 128 Mbytes of RAM and a 20 Gb Hard Disk. As an indication, the GA took 14m57s to identify the best genotypes for each value of *N* i.e. N=2..6 and 1m54s to select the potential MWUs.

7 Conclusions and Future Work

In this paper, we have presented an application of Genetic Algorithms for the specific task of Multiword Unit extraction. For that purpose, a fitness function, together with a set of constraints, has been defined whose maximization has served as a basis for the identification of pertinent word *N*-grams based on a similarity measure. In order to propose a suitable platform for evaluation, a software application called GALEMU (Genetic ALgorithm for the Extraction of Multiword Units) has been implemented. GALEMU distinguishes itself from previous architectures by its acquisition process based on similarity. Furthermore, GALEMU uses different heuristics to identify relevant word associations thus taking advantage of various useful information evidenced by the constant evolution of researches. In the near future, we intend to verify whether GALEMU would benefit from the introduction of new variables that have proved to be good heuristics for our task or whether some feature selection should be applied to prevent from redundancy of attributes. In particular, we intend to introduce new association measures such as the Dice coefficient [23], the association ratio [2], the Φ^2 test [12] and the Log-likelihood ratio [7], as each one tends to favor different characteristics of MWUs. We are also clearly convinced that improvements may be introduced by the weighting of the fitness function as well as the similarity measures. Indeed, it is clear that some informations are more important than others for the process of selection.

References

1. R. Basili, What can be learned from Raw Texts? An Integrated Tool for the Acquisition of Case Rules, Taxonomic Relations and Disambiguation Criteria, journal of Machine Translation, 8, pp 147-173, 1993.
2. K. Church and P. Hanks, Word Association Norms Mutual Information and Lexicography, Computational Linguistics, 16-1, pp 23-29, 1990.
3. B. Daille, Study and Implementation of Combined Techniques for Automatic Extraction of Terminology, The balancing act combining symbolic and statistical approaches to language, 1995.
4. L. Cooper and D. Steinberg, Introduction to Methods of Optimization, W.B. Saunders, London, 1970.
5. G. Dias, S. Vintar, S. Guilloré and J.P.G. Lopes, Identifying and Integrating Terminologically Relevant Multiword Units in the IJS-ELAN Slovene-English Parallel Corpus, 10th Computational Linguistics in the Netherlands, November 1999.
6. G. Dias, S. Guilloré and J.G.P. Lopes, Normalisation of Association Measures for Multiword Lexical Unit Extraction, International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications, Tunisia, 2000.

7. T. Dunning, Accurate Methods for the Statistics of Surprise and Coincidence, Association for Computational Linguistics, 19-1,1993.
8. D. Evans, Concept Management in text via Natural Language Processing: the CLARIT approach, Working Notes of the 1990 AAAI Symposium on Text-Based Intelligent Systems, Stanford University, March, pp 93-95, 1990.
9. P. Fung, A Pattern Matching Method for finding Noun and Proper Noun Translations, 33rd Annual Meeting of the Association for Computational Linguistics, pp 236-243, Boston, MA, USA, 1995.
10. K. Frantzi and S. Ananiadou, A Hybrid Approach to Term Recognition, NLP & Industrial Applications, Canada, pp 93-98, 1996.
11. K. Frantzi and S. Ananiadou, Extracting Nested Collocations, International Conference on Computational Linguistics, Copenhagen, pp 41-46, 1996.
12. W. Gale, Concordances for Parallel Texts, Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora, 1991.
13. D.E Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Reading, 1989.
14. G. Gross, Les expressions figées en français, Ophrys, Paris, 1996.
15. B. Habert and C. Jacquemin, Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques, Traitement Automatique des Langues, Association pour le Traitement des Langues,France, 34:2, pp 5-41, 1993.
16. S. Katz, N. Johnson and C. Read, Measures of Similarity, Dissimilarity, and Distance, Encyclopedia of Statistical Sciences, John Wiley & Sons, New York, 5, 1982.
17. Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer, Berlin Heidelberg New York, 1996.
18. A. Salem, La Pratique des segments répétés, Klincksieck, Paris, 1987.
19. R. Schneider and I. Renz, The Relevance of Frequency Lists for Error Correction and Robust Lemmatization, Journées d'Analyse de Données Textuelles, EPFL, Lausanne, 2000.
20. S. Shimohata, Retrieving Collocations by Co-occurrences and Word Order Constraints, ACL-EACL, 1997.
21. J. Silva, G. Dias, S. Guilloré and J.G.P. Lopes, Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, 9th Portuguese Conference in Artificial Intelligence, September, 21-34, 1999.
22. F. Smadja, Retrieving Collocations From Text: XTRACT, Computational Linguistics, 19-1, pp 117-143, 1993.
23. F. Smadja, K.R. McKeown and V. Hatzivassiloglou, Translating Collocations for Bilingual Lexicons: A Statistical Approach, Computational Linguistics, 22-1, 1996.
24. G. K. Zipf, The Psychobiology of Language, an Introduction to Dynamic Philology, Houghton-Mifflin, Boston, 1935.