Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment

Gaël Dias*, Sérgio Nunes†

*Centre of Mathematics
Universidade da Beira Interior, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal
ddg@di.ubi.pt

†Centre of Informatics
Universidade da Beira Interior, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal
sergio@ubi.pt

Abstract

In this paper, we present an application of Genetic Algorithms to extract Multiword Units (i.e. complex lexical units such as compound nouns, idiomatic expressions or phrase templates). For that purpose, a fitness function will be defined whose maximization will serve as a basis for the identification of pertinent word N-grams (i.e ordered vectors of N words) based on different similarity measures. Finally, we will provide an experiment realized over an English Linux Manual that evidences promising results.

1. Introduction

The acquisition of Multiword Units (MWUs) has long been a significant problem in Natural Language Processing (NLP). Indeed, most of the work in knowledge acquisition has aimed at extracting explicit information from texts (i.e. knowledge about the world) and has generally neglected the extraction of implicit information (i.e. knowledge about the language). However, 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[1] 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[2].

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 innovative 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 (Dias et al., 2000a). For that purpose, each positional N-gram is associated to a set of attribute values (e.g. frequency, degree of cohesion, size etc.) that will represent a specific chromosome of the overall population. Once the population is defined, the maximization of the fitness function will provide the "best" genotype that hopefully will be a global maximum. Finally, in order to extract relevant MWUs from the original population, a set of different similarity measures will evidence the 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.

So, in order to evaluate our methodology, tests have been realized over a Linux Manual written in English. As expected, compound nouns/names/determinants and verbal/adverbial/prepositional/conjunctive locutions have been extracted. However, we will access that different results can be obtained using different scenari of extraction i.e. different similarity measures.

2. The Gentetic 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-directionnal search over a sample space by maintaining a population of potential solutions and also 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

[1] i.e. The evolution from"general" rules towards rules specifying the usage of words on a case-by-case basis.

[2] This classification is proposed by (Gross, 1996).

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.

2.1. Floating Point Representation

The binary representation traditionally used in genetic algorithms has revealed some drawbacks when applied to multidimensional, high precision numerical problems (Michalewicz, 1996). 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 (Goldberg, 1989). 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 positionnal N-grams, (Dias et al., 2000b) have defined a new N-ary association measure called the Mutual Expectation that does not under-evaluate the degree of cohesion of sequences of words containing frequent single words. Our first gene will model the Mutual Expectation of any given N-gram.

Gene x_1: Aside from association measures, frequency is considered by many researchers (Daille, 1995) (Justeson and Katz, 1993) (Frantzi and Ananiadou, 1996b) as an effective criterion for Multiword Unit identification. So, highly frequent word N-grams are more likely to be MWUs than unfrequent ones. Consequently, we propose that the second gene of any N-gram individual should be its relative frequency.

Gene x_2: However, (Frantzi and Ananiadou, 1996b) 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, (Dias et al., 1999) evidence that complex terms (i.e. terminologically relevant MWUs) are specific lexical relations that favor the occurrence of unfrequent single words in their core. So, the more an N-gram contains frequent single words in its inside, 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 optimization problem.

2.2. Fitness Function

To distinguish between different solutions, we use an objective (evaluation) 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 in equation 1 where X is a given chromosome.

$$g(X) = x_0 + x_1 + x_2 - x_3 \quad (1)$$

However, real-word optimization problems are generally constrained. The specific task of extracting MWUs does not avoid this general rule. As a consequence, aside from the definition of the fitness function, a specific set of domain constraints and inequalities will have to be defined.

2.3. Handling Constraints

In the context of floating point representation genetic algorithms, many researches have been carried out in order to define suitable optimization processes. However, as stated in (Cooper and Steinberg, 1970), "*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, (Silva et al., 1999) 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 in the inequations 2 and 3.

$$x_0 \geq x_4 \tag{2}$$

$$x_0 > x_5 \tag{3}$$

Gene x_6: Finally, (Frantzi and Ananiadou, 1996a) and (Justeson and Katz, 1993) 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.

$$x_6 < x_1 \tag{4}$$

Additionally, new constraints can be evidenced thus introducing knowledge about the problem. For instance, it is clear that the marginal frequency of an N-gram must be superior or at least equal to its relative frequency.

$$x_3 \geq x_1 \tag{5}$$

In the same way, the number of different super-groups of a given N-gram can not be superior to its relative frequency, thus giving rise to the following constraint.

$$x_2 \leq x_1 \tag{6}$$

3. Similarity Measures

The application of the genetic algorithm 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.

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, different measures have been defined. We will access four of them that have been implemented in GALEMU. So, suppose that $X_i = (X_{i_1}, X_{i_2}, ..., X_{i_p})$ is a row vector of observations on p variables associated with a label i. The distance between two units i and j is defined as $D_{ij} = f(X_i, X_j)$ where f is some function of the observed values. The following functions have been proposed.

$$D^1_{ij} = \frac{1}{p}\sum_{k=1}^{p}(X_{i_k} - X_{j_k})^2 \quad Euclidean \tag{7}$$

$$D^2_{ij} = \frac{1}{p}\sum_{k=1}^{p}\frac{(X_{i_k} - X_{j_k})^2}{(X_{i_k} + X_{j_k})^2} \quad Divergence \tag{8}$$

$$D^3_{ij} = \frac{\sum_{k=1}^{p}|X_{i_k} - X_{j_k}|}{\sum_{k=1}^{p}(X_{i_k} + X_{j_k})} \quad Bray/Curtis \tag{9}$$

$$D^4_{ij} = \frac{\sum_{k=1}^{p}|X_{i_k} - X_{j_k}|}{\sum_{k=1}^{p}max(X_{i_k}, X_{j_k})} \quad Soergel \tag{10}$$

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

4. Experiment and Results

In order to evaluate our methodology, tests have been performed over an English Linux Manual of approximatively 54000 words[3] that has been extracted from the IJS-ELAN text collection encoded with the *Text Encoding Initiative* format (Erjavec, 1999). In this paper, we will specifically focus on the results obtained by applying different similarity measures over the same initial population. 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). Although it is usually difficult to determine whether a word association is a MWU or not, we will try to stick to (Gross, 1996)'s classification. In order to perform an homogeneous evaluation, we defined a typical set of parameters that is shown in Table 1.

Parameter	Value
Number of Generations	6000
Size of the Sample Population	50
Mutation Probability	0.01
Crossover Probability	0.6
Mutation Type	Non-uniform
Crossover Type	Uniform
Distance Error	0.15

Table 1: Parameters of the Experiment

As expected, a great deal of linguistic phenomena have been identified. Thus, compound nouns, names, determinants and verbal, adverbial, prepositional, conjunctive

[3]This corpus may be considered as a small one compared to the standards in the field. However, purely statistical methodologies increase in quality as the size of the corpus increases. As a consequence, we will show that our technique is suitable to small corpora and should easily demonstrate better behavior with bigger texts.

locutions have been identified for all scenari. We propose a sample of the extracted MWUs in Table 2.

Extracted MWUs	
Home Page	Patrick Volkerding
operating system	Red Hat
in some way	permission denied
something goes wrong	free of charge
as well as	X Window System
to compile a kernel	" file not found "

Table 2: Sample Extracted MWUs

However, different results of precision and spread are revealed as shown in Table 3.

Distance	Precision (%)	Spread (# of units)
D^1	64	73
D^2	62	29
D^3	71	131
D^4	70	53

Table 3: Comparative Results

A quick look at the results show that two distinct sets of measures can be identified. On one side, the *Euclidean* and the *Divergence* distances evidence results below 65 % precision whereas the *Bray and Curtis* and *Soergel* measures demonstrate precision results around 70 %. These results are not surprising at all. Indeed, for both sets, the formulae are quite similar. However, it is interesting to notice that 9 % precision can be gained whether one uses the *Euclidean* measure or the *Bray and Curtis* distance to evaluate the results. Similarly, the values for spread show a great discrepancy between measures in the same set. Indeed, the Soergel distance only extracts 40 % of the expressions elected with the *Bray and Curtis* measure and the same can be assessed between the *Euclidean* and the *Divergence* distances. However, a detailed analysis of the formulae (namely the denominators) show that these results were to be expected. Nevertheless, they evidence a classical evaluation problem. As a matter of fact, the differences evidenced in the acquisition process are due to the definition of the same threshold for all the four measures. This should clearly be avoided. But then, what would be the basis for an impartial evaluation? We will leave this question in open.

5. Conclusions

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 different similarity measures. In particular, we have provided an experiment realized over an English Linux Manual that evidences promising results. However, different issues may be obtained depending on the similarity measure in use. The system will be soon on-line at http://galemu.di.ubi.pt.

6. References

Cooper, L. and D. Steinberg, 1970. *Introduction to Methods of Optimization*. London: W.B. Saunders.

Daille, B., 1995. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act combining symbolic and statistical approaches to language*.

Dias, G., S. Guilloré, J-C. Bassano, and J.G.P. Lopes, 2000a. Extraction automatique d'units lexicales complexes: Un enjeu fondamental pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2).

Dias, G., S. Guilloré, and J.G.P. Lopes, 2000b. 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.*

Dias, G., S. Vintar, S. Guilloré, and J.G.P. Lopes, 1999. Identifying and integrating terminologically relevant multiword units in the ijs-elan slovene-english parallel corpus. *10th Computer Linguistics in the Netherlands*.

Erjavec, T., 1999. A tei encoding of aligned corpora as translation memories. *EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99).*

Frantzi, K. and S. Ananiadou, 1996a. Extracting nested collocations. *International Conference on Computational Linguistics*.

Frantzi, K. and S. Ananiadou, 1996b. A hybrid approach to term recognition. *NLP & Industrial Applications*.

Goldberg, D.E, 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading: Addison-Wesley.

Gross, G., 1996. *Les expressions figées en franais*. Paris: Ophrys.

Justeson, J. and S. Katz, 1993. Technical terminology: Some linguistic properties and an algorithm for identification in text. Technical report, IBM.

Michalewicz, Z., 1996. *Genetic Algorithms + Data Structures = Evolution Programs*. Berlin Heidelberg New York: Springer.

Silva, J., G. Dias, S. Guilloré, and G. Lopes, 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *9th Portuguese Conference in Artificial Intelligence*:21–24.