Combining Evolutionary Computing and Similarity Measures to Extract Collocations from Unrestricted Texts

Gaël Dias and **Sérgio Nunes**
Centre of Mathematics
University of Beira Interior
Marquês d'Ávila e Bolama Street
6200-053, Portugal
ddg@noe.ubi.pt slnunes@yahoo.com

Abstract

In this paper, we focus on the suitability of natural selection for the extraction of multiword units from unrestricted texts. For that purpose, we combine numerous heuristics that are usually used on a stand-alone basis in a unique architecture called GALEMU.

1 Introduction

In this article, we present a system called GALEMU (Genetic ALgorithm for the Extraction of Multiword Units) designed to extract multiword units (MWUs) from unrestricted text corpora. GALEMU proposes an original architecture based on the combination of a genetic algorithm and a similarity measure. The basic idea of the application is simple. First, the text corpus is segmented into a set of positional N-grams from which significant individuals will have to be identified. Then, each N-gram is associated with a set of attribute values thus representing a specific chromosome of the overall population. Once the population has been defined, the maximisation of the fitness function provides the "best" genotype of the population. Finally, to identify relevant MWUs from the original population, a similarity measure evidences the relatedness between a specific N-gram in the population and the elected genotype.

2 Related Work

Some studies propose the application of binary association measures to evaluate the degree of cohesiveness between two words (Church & Hanks 90) (Gale & Church 91) (Smadja 93) (Shimohata *et al.* 97). As a consequence, bootstrapping techniques have to be applied to acquire associations with more than two words. Unfortunately, such techniques have shown their limitations (Habert & Jacquemin 93). In order to overcome the lack of generalization for N individual words, N-ary association measures have then been proposed (Frantzi & Ananiadou 96a) (Schneider & Renz 00). 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. However, one important remark must be pointed out. Both approaches rely only on a few association metrics. In most cases, just one association measure is used to evidence relevancy of a sequence of words. However, (Daille 96) and (Dias *et al.* 00a) have shown that many applications might benefit from the combination of different information. In order to overcome this drawback, we propose an original solution that relies on a Genetic Algorithm that combines various attributes for each N-gram so that relevancy may not rely on one exclusive association measure.

3 The Genetic Algorithm

The floating point representation genetic algorithms allow a one-gene-one-variable correspondence that eases the codification process. Consequently, each chromosome (i.e. any N-gram in the population) can easily be represented as a vector of real numbers, each number corresponding to a specific variable of the problem. In our context, we will define 7 variables that have been proposed in different studies as good heuristics for the identification of highly cohesive sequences of words.

3.1 Genes

Gene g_0: Association measures have been widely used in order to define the degree of cohesiveness of word N-grams. In particular, the more cohesive a sequence of words is, the more likely it is a MWU. In the specific context of positional N-grams, we will use the Mutual Expectation defined by (Dias *et al.* 00a).

Gene g_1: Aside from association measures, frequency is considered by many researchers (Daille 96) (Frantzi & Ananiadou 96b) as an effective criterion for MWU identification. So, highly frequent word N-grams are more likely to be MWUs

than unfrequent ones.

Gene g_2: However, (Frantzi & Ananiadou 96b) demonstrate that stand-alone frequency can lead to error in the acquisition process. Indeed, while the fact that an N-gram appearing in other longer N-grams is a negative factor for its relevancy, the word sequence increases in probability of importance as the number of these longer N-grams increases. We will consider this number as our third gene.

Gene g_3: Moreover, in the specific context of terminology, (Dias *et al.* 00b) evidence that complex terms are specific lexical relations that favour the occurrence of unfrequent single words in their core. As a consequence, for each N-gram, we will evaluate the arithmetic mean of the frequencies of all its constituents that we will call its marginal frequency.

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.

3.2 Fitness Function

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 a fitness function. From the previous assumptions, a simple fitness function can directly be suggested in equation 1 where X is a given chromosome.

$$f(X) = g_0 + g_1 + g_2 - g_3 \tag{1}$$

3.3 Handling Constraints

However, as stated in (Cooper & Steinberg 70), "*all optimization problems of the real word are, in fact, constrained problems*". For our specific task, three constraints can be evidenced that introduce three new genes.

Gene g_4 and Gene g_5: In order to select potential MWUs from a set of association measure valued N-grams, (Dias *et al.* 00a) have proposed an original methodology called the GenLocalMaxs that elects a positional N-gram if its association measure value is higher or equal than the association measure values of all its sub-groups of (N-1) words and if it is strictly higher than the association measure values of all its super-groups of (N+1) words. So, the fifth and sixth genes of each individual will respectively be the highest ME value of all the sub-groups of the considered genotype and the highest ME value of all its super-groups giving rise to the following constraints.

$$g_0 \geq g_4 \tag{2}$$

$$g_0 > g_5 \tag{3}$$

Gene g_6: Finally, (Frantzi & Ananiadou 96a) and (Justeson & Katz 93) 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 of the most frequent super-group of the considered individual introducing the following constraint.

$$g_6 < g_1 \tag{4}$$

So, if new individuals do not guarantee these constraints, they will be penalized in terms of fitness so that their probability to reproduce will be lowered by penalty functions that will not be described in this paper.

4 Similarity Measures

In order to identify relevant word N-grams, we now need to know how similar a particular N-gram is to the typical MWU. For that purpose, we will use a similarity measure called the Bray and Curtis distance. So, suppose that $X_i = (X_{i_0}, X_{i_1}, ..., X_{i_p})$ is a row vector of observations on $p+1$ variables associated with a label i, the distance between two units i (the "best" genotype) and j (any chromosome) is defined as follows.

$$D_{ij} = \frac{\sum_{k=0}^{p} |X_{i_k} - X_{j_k}|}{\sum_{k=0}^{p} X_{i_k} + X_{j_k}} \quad Bray/Curtis \tag{5}$$

5 Experiments and Results

In order to evaluate our methodology, some experiments have been performed over a multilingual corpus of approximatively 200,000 words,

written in English, French and Portuguese. As evaluation is concerned, it is interesting to notice that most studies of statistical extractors have traditionally been realized for English and have neglected to test the ability of the architectures to deal with different languages. As a consequence, in this paper, we will specifically focus on the performances of GALEMU when applied to English, French and Portuguese.

As evaluation indicators, we will propose the precision rates obtained for the three languages based on (Gross 96)'s classification of MWUs. For that purpose, we will respectively evidence the results of the acquisition process for the first 100, 200, 300 and 400 most similar N-grams in table 1.

Ranked	English	French	Portuguese
100	81	62	79
200	80	63	72
300	79	68	70
400	73	65	69

Table 1: Precision Rates in %

The figures are clear. The performance of GALEMU strongly depends on the language into consideration. Indeed, over the same corpus, the results can vary from 81% to 62% between French and English i.e. a 19% difference. Although the difference tends to shrink as the number of considered N-grams increases, the gap between French and English still remains high with 8% evidenced. The same situation also stands for Portuguese but at a lesser degree. Two major causes can be pointed at for that figures. First, GALEMU tends to favour the extraction of 2-grams. This situation clearly benefits the results for English. Indeed, as French and Portuguese make great use of prepositions to construct complex terms, MWUs are usually long sequences of words i.e. from 3 to 6 words. The other important reason for these results has to do with the fact that French and Portuguese are languages where flexion plays an important role. As a consequence, the same linguistic phenomenon can be evidenced in the corpus in different graphical forms thus making difficult the observation of regularities.

As a conclusion, it is clear that the information contained in unannotated corpora is fundamental to identify important concepts of the language. However, depending on the language into consideration, this information can reveal insufficiencies and linguistic treatment may be needed.

6 Conclusion

In this paper, we have defined a global architecture based on a genetic algorithm and a similarity measure that combines several heuristics that are usually used on a stand-alone basis in a unique architecture called GALEMU. A multilingual evaluation has stressed out that different results can be obtained for different languages thus evidencing the limitation of raw-text-based architectures.

References

(Church & Hanks 90) K. Church and P. Hanks. Word association norms mutual information and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

(Cooper & Steinberg 70) L. Cooper and D. Steinberg. *Introduction to Methods of Optimization*. W.B. Saunders, London, 1970.

(Daille 96) B. Daille. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act combining symbolic and statistical approaches to language*, pages 49–66, 1996.

(Dias *et al.* 00a) G. Dias, S. Guilloré, J-C. Bassano, and G. Lopes. Extraction automatique d'units lexicales complexes: Un enjeu fondamental pour la recherche documentaire. *Traitement Automatique des Langues*, 41(2):447–473, 2000.

(Dias *et al.* 00b) G. Dias, S. Guilloré, and G. Lopes. Benefiting from multi-domain corpora for extracting terminologically relevant multiword lexical units. *9th EURALEX International Congress*, pages 339–350, 2000.

(Frantzi & Ananiadou 96a) K. Frantzi and S. Ananiadou. Extracting nested collocations. *International Conference on Computational Linguistics (COLING)*, pages 41–46, 1996.

(Frantzi & Ananiadou 96b) K. Frantzi and S. Ananiadou. A hybrid approach to term recognition. *NLP and Industrial Applications*, pages 93–98, 1996.

(Gale & Church 91) W. Gale and K. Church. Concordances for parallel texts. *7th Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, pages 40–62, 1991.

(Gross 96) G. Gross. *Les expressions figées en franais*. Ophrys, Paris, 1996.

(Habert & Jacquemin 93) B. Habert and C. Jacquemin. Noms composés, termes, dénominations complexes: Problématiques linguistiques et traitements automatiques. *Traitement Automatique des Langues*, 34(2):5–41, 1993.

(Justeson & Katz 93) J. Justeson and S. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. Technical report, IBM, 1993.

(Schneider & Renz 00) R. Schneider and I. Renz. The relevance of frequency lists for error correction and robust lemmatization. *5emes Journées Internationales d'Analyse de Données Textuelles (JADT)*, 2000.

(Shimohata *et al.* 97) S. Shimohata, T. Sugio, and J. Nagata. Retrieving collocations by co-occurrences and word order constraints. *35th annual meeting of the Association for Computational Linguistics*, pages 476–481, 1997.

(Smadja 93) F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.