Extracting Concepts from Dynamic Legislative Text Collections

Gaël Dias[12], Sara Madeira[1] & José Gabriel Pereira Lopes[2]

[1] Universidade da Beira Interior, Departamento de Informática

rua Marquês d'Ávila e Bolama, 6200-053 Covilhã Portugal

ddg@noe.ubi.pt, saramadeira@mail.telepac.pt

[2] Universidade Nova de Lisboa, Departamento de Informática

Quinta da Torre, 2825-114 Caparica Portugal - gpl@di.fct.unl.pt

Abstract

Selecting discriminating terms in order to represent the contents of texts is a critical problem for many applications in Information Retrieval. Most of the Information Retrieval systems index documents based on individual words that are not specific enough to evidence the contents of texts. As a consequence, there has been a growing interest in developing techniques for automatic term extraction. In this context, we propose a new architecture for retrieving relevant documents in a dynamic text collection. It combines the SINO search engine with the SENTA software designed for the automatic extraction of multiword lexemes. In this paper, we will particularly focus on the SENTA module that has recently been added to the global architecture.

Keywords: Multiword Lexical Unit Extraction, Information Retrieval, Web Interface.

1. Introduction

Selecting discriminating terms in order to represent the contents of texts is a critical problem for many applications in Information Retrieval. Ideally, the indexing terms should directly describe the concepts present in the documents. However, most of the Information Retrieval systems index documents based on individual words that are not specific enough to evidence the contents of texts. As a consequence, evolutionary retrieval systems use multiword terms previously extracted from text collections to represent the contents of texts (Evans and Lefferts 1993). Indeed, multiword terms embody meaningful sequences of words that are less ambiguous than single words and allow approximating more accurately the contents of texts.

However, most multiword terms are not listed in lexical databases. Indeed, the creation, the maintenance and the upgrade of terminological data banks often require a great deal of manual efforts that can not cope with the ever growing number of texts to analyse. Moreover, due to the constant dynamism of specialised languages, the set of multiword terms is opened and to be completed (Habert and Jacquemin 1997). As a consequence, there has been a growing interest in developing techniques for automatic term extraction. In the context of the PGR Project, funded by the Portuguese Ministry of Justice, we propose a new architecture for retrieving relevant documents in a dynamic legislative text collection (See Figure 1). It combines the SINO search engine (Quaresma *et al.* 1999) with the SENTA software designed for the automatic extraction of multiword lexemes (Dias *et al.* 1999). At this stage of the project, the set of multiword lexemes is manually checked and filtered out in order to insert useful indexing terms into the search engine thus producing high quality retrieval process.

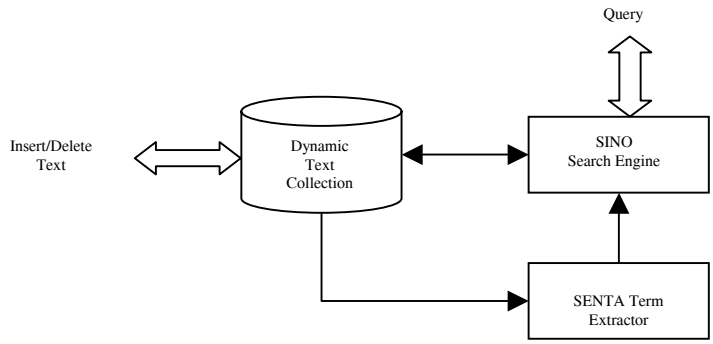

Figure 1: The Global Retrieval Architecture

In this paper, we will focus on the SENTA module that has recently been added to the global architecture of our system. SENTA (Software for the Extraction of N-ary Textual Associations) has been thought around two main principles (Dias *et al.* 2000). First, following **the rigidity principle**, we propose that the general information appearing in raw texts should be sufficient to extract meaningful multiword lexemes without applying domain-dependent or language-dependent heuristics. Second, following the **corpus integrity principle,** we propose that the input text corpus should not be modified at all (i.e., the text is neither lemmatised, nor pruned with lists of stop words). So, SENTA retrieves from naturally occurring text, contiguous and non-contiguous multiword lexemes on the basis of two complementary techniques: the Mutual Expectation measure and the LocalMaxs algorithm. One particularity of our architecture is to follow the changes in the text collection. Indeed, according to (Manning and Schütze 1999), lexical regularities appear and disappear as language evolves. Thus, a particular lexical relation that may not be

an expression at any given time *t*, may well form a multiword unit at time *t+1* and vice versa. So, whenever a new text is inserted or an old one deleted, SENTA is re-run over the collection. Thus, new expressions may be discovered and old ones may disappear.

2. Data Preparation

The first step of our methodology performs the transformation of the input text into a set of n-grams. Indeed, a great deal of applied works in lexicography evidence that most of the lexical relations associate words separated by at most five other words and assess that multiword terms are specific lexical relations that share this property (Sinclair 1974). As a consequence, a multiword term can be defined in terms of structure as a specific word n-gram calculated in the immediate context of three words to the left hand side and three words to the right hand side of a pivot word. This situation is illustrated in Figure (2) for the pivot word *Lei* (*Law*) being given the input sentence (1). Indeed, *Lei de Imprensa* (*Press Law*) is a specific multiword term. By definition, a word n-gram is a vector of n words where each word is indexed by the signed distance that separates it from its associated pivot word. Consequently, an n-gram can be contiguous or non-contiguous whether the words involved in the n-gram represent or not a continuous sequence of words in the corpus.

(1) *O artigo 35 da Lei de Imprensa prevê esse procedimento em caso de burla agravada.*

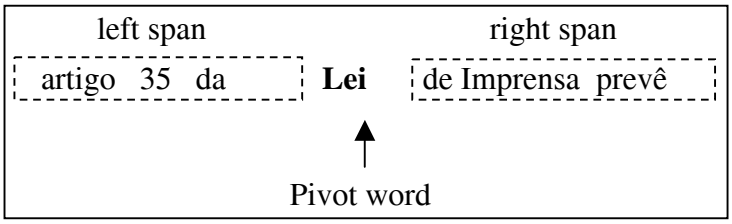

Figure 2: The Context Span

For instance, if we consider the sentence (1) as the current input text and "*Lei*" the pivot word, a contiguous and a non-contiguous word 3-gram are respectively illustrated in the following table.

w_1	$position_{12}$	w_2	$position_{13}$	w_3
Lei	+1	*de*	+2	*Imprensa*
Lei	-3	*artigo*	+3	*Prevê*

Table 1: Sample word 3-grams calculated from the pivot word *Lei.*

Generically, an n-gram is a vector of n textual units where each textual unit is indexed by the signed distance that separates it from its associated pivot textual unit. By convention, the pivot textual unit is always the first element of the vector and its signed distance is equivalent to zero. We represent an n-

gram by a vector [p_{11} u_1 p_{12} u_2 p_{13} u_3...p_{1i} u_i...p_{1n} u_n] where p_{11} is equal to zero and p_{1i} (for i=2 to n) denotes the signed distance that separates the textual unit u_i from the pivot textual unit u_1.

3. Normalized Expectation and Mutual Expectation

In order to evaluate the degree of cohesiveness existing between textual units, various mathematical models have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness between two textual units and do not generalise for the case of n individual textual units (Church and Hanks 1990, Gale 1991, Dunning 1993, Smadja 1993, Smadja 1996, Shimohata 1997). As a consequence, these mathematical models only allow the acquisition of binary associations and bootstrapping techniques have to be applied to acquire associations with more than two textual units. On the other hand, for the specific case of word associations, the proposed mathematical models tend to be over-sensitive to frequent words. In order to overcome both problems, we introduce a new association measure called the Mutual Expectation (ME) that evaluates the degree of rigidity that links together all the textual units contained in an n-gram ($\forall n, n \geq 2$) based on the concept of Normalised Expectation (NE) (Dias *et al*. 1999).

3.1 Normalised Expectation

The basic idea of the Normalised Expectation is to evaluate the cost, in terms of cohesiveness, of the loss of one textual unit in an n-gram. So, the more cohesive a group of textual units is, that is the less it accepts the loss of one of its components, the higher its Normalised Expectation will be. The underlying concept of the Normalised Expectation is based on the conditional probability defined in Equation (1).

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

Equation 1: Conditional Probability.

The definition of the conditional probability can be applied in order to measure the expectation of the occurrence of one textual unit in a given position knowing the occurrence of the other n-1 textual units also constrained by their positions. However, this definition does not accommodate the n-gram length factor. Naturally, an n-gram is associated to n possible conditional probabilities. The Normalised Expectation, based on a normalisation of the conditional probability, proposes an elegant solution to represent in a unique formula all the n conditional probabilities involved by an n-gram. For that purpose

we introduce the concept of the Fair Point of Expectation (FPE). In order to perform a sharp normalisation, the FPE is the arithmetic mean of the denominators of all the conditional probabilities. Theoretically, the Fair Point of Expectation is the arithmetic mean of the n joint probabilities of the (n-1)-grams contained in an n-gram and it is defined in Equation (2).

$$FPE([p_{11}u_1\, p_{12}\, u_2 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]) = \frac{1}{n}\left(p([p_{12}u_2 \ldots p_{2i}\, u_i \ldots p_{2n}\, u_n]) + \sum_{i=2}^{n} p\left(\left[p_{11}u_1 \ldots \hat{p}_{1i}\ \hat{u}_i \ldots p_{1n}\, u_n \right]\right)\right)$$

Equation 2: Fair Point of Expectation.

In particular, 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 2 to n. Thus, the normalisation of the conditional probability is realised by the introduction of the FPE into the general definition of the conditional probability as defined in Equation (3).

$$NE([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]) = \frac{p([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n])}{FPE([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n])}.$$

Equation 3: Normalised Expectation.

3.2 Mutual Expectation

(Justeson and Katz 1993) and (Daille 1995) have shown in their studies that frequency is one of the most relevant statistics to identify multiword terms with specific syntactical patterns. The studies made by (Frantzi and Ananiadou 1996) in the context of the extraction of interrupted collocations also assess that the relative frequency is an important clue for the retrieval process. From this assumption, we deduce that between two word n-grams with the same Normalised Expectation, the most frequent word n-gram is more likely to be a relevant multiword unit. So, the Mutual Expectation between n words is defined in Equation (4) based on the Normalised Expectation and the relative frequency.

$$ME([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]) = p([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]) \times NE([p_{11}u_1 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n])$$

Equation 4: Mutual Expectation.

Comparing to the previously proposed mathematical models, the Mutual Expectation allows evaluating the degree of cohesiveness that links together all the textual units contained in an n-gram (i.e. $\forall n, n \geq 2$) as it accommodates the n-gram length factor.

4. Acquisition Process

Most of the approaches have based their selection process on the definition of global frequency thresholds and/or on the evaluation of global association measure thresholds (Church and Hanks 1990, Smadja 1993, Daille 1995, Shimohata 1997, Feldman 1998). This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether a word n-gram is a pertinent word association or not. However, these thresholds are prone to error as they depend on experimentation. Furthermore, they highlight evident constraints of flexibility, as they need to be re-tuned when the type, the size, the domain and the language of the document change (Habert *et al*. 1997). The LocalMaxs (Silva *et al*. 1999) proposes a more flexible and fine-tuned approach for the selection process as it concentrates on the identification of local maxima of association measure values. So, we may deduce that a word n-gram is a multiword term 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. Let *assoc* be an association measure, W an n-gram, Ω_{n-1} the set of all the (n-1)-grams contained in W, Ω_{n+1} the set of all the (n+1)-grams containing W and *sizeof* a function that returns the number of words of a word n-gram. The LocalMaxs is defined as follows:

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$,

W is a multiword term if $(sizeof(W)=2 \ \wedge \ assoc(W) > assoc(y))$

$\vee$

$(sizeof(W)\neq 2 \ \wedge \ assoc(W) \geq assoc(x) \ \wedge \ assoc(W) > assoc(y))$

5. The Web-Based Architecture of SENTA

The web-based implementation of SENTA has been realized at the Portuguese University of Beira Interior in collaboration with the New University of Lisbon. The application allows any authorized user to insert new texts (via browser) into the text collection and consult the set of the extracted multiword lexemes for further validation (See figure 3). When submitting the request to the Web Server, the text is pre-processed and stored in the Database. The three steps of SENTA are then run locally on the Database Server. Finally, the results are displayed (See Figure 4) in a table along with their frequency. The results show that relevant multiword terms are extracted: *normas legais* (*legal norms*), *Conselho Consultivo* (*Consultive Council*), *ex-administração ultramarina* (*ultramarine ex-administration*).

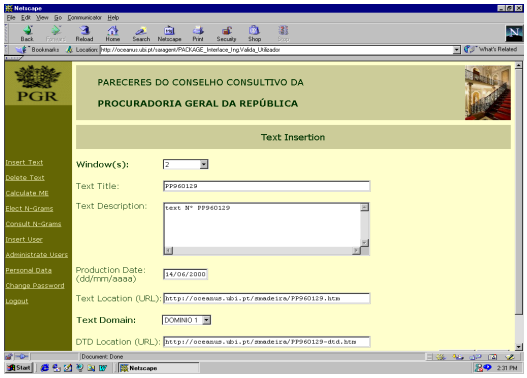

Figure 3: Text Insertion

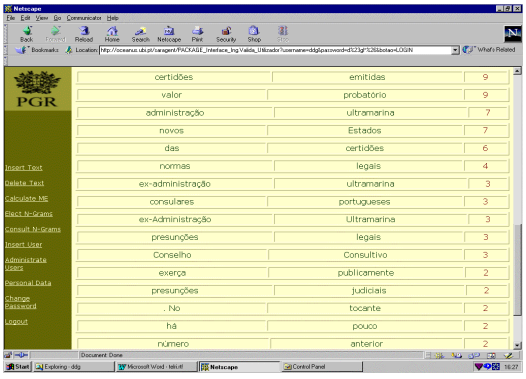

Figure 4: Consult Page

From this interface, an expert in Law terminology can then easily select the relevant multiword terms to be integrated as indexing terms in the search engine SINO. This stage is still done manually but we are working on a fully automated version that would avoid human intervention and post-editing. So, the user is guided by SINO in his search for information by accessing a list of complex terms that embody fundamental concepts of the document collection. For example, if one is interested in getting information about crime, the system suggests a list of complex terms related to the query. Thus, the user is able to refine his search by selecting one of the terms in the list and access the most relevant documents. As illustrated in Figure (5), the user may choose one of the following phrases related to crime: *crime militar* (*military crime*) or *crime internacional* (*international crime*).

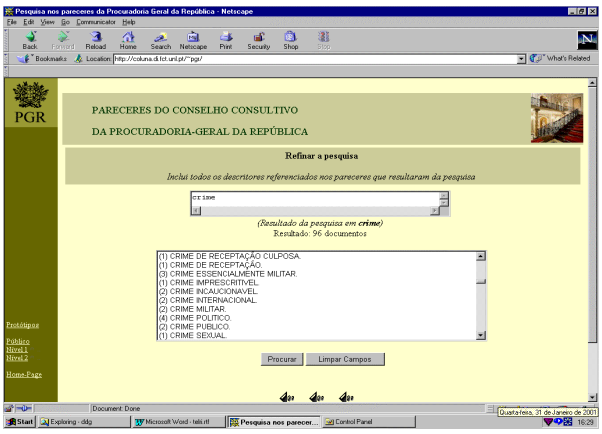

Figure 5: SINO search engine

6. Conclusion

In this paper, we have proposed a web-based integrated solution for enhanced Information Retrieval which combines the search engine SINO with the term extractor SENTA. This work is the result of the

collaboration between two Portuguese Universities for the purpose of the ”PGR-Acesso Selectivo aos pareceres da Procuradoria Geral da República” project that is being funded by the Portuguese Ministry of Justice. Our fundamental goal is the automatic extraction of multiword lexemes (concepts) to improve information retrieval by introducing new indexing terms (a fundamental issue in Information Retrieval). We are actually planning to improve our Consult Interface by introducing a set of tools (concordancer, hypertext links and other association measures) to ease the decision making of terminologists. The application can be accessed by the following URL:

http://oceanus.ubi.pt/saragent/package_interface.form_password.

References

Church, K.W. & P. Hanks. 1990. "Word Association Norms Mutual Information and Lexicography". *Computational Linguistics*, 16 (1) :23--9.

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. and S. Guilloré and J-C. Bassano and J.G.P. Lopes. 2000. "Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?". *Recherche d'Informations Assistée par Ordinateur (RIAO'2000)*, Paris, France.

Dias G. and S. Guilloré and J.G.P. Lopes. 1999. "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora". *Traitement Automatique des Langues Naturelles*. Institut d'Etudes Scientifiques, Cargèse, France.

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

Evans, D. & R. Lefferts. 1993. "Design and Evaluation of the CLARIT-TREC-2 System". *TREC93:* 137150.

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

Frantzi K.T. and S. Ananiadou . 1996. "Retrieving Collocations by Co-occurrences and Word Order Constraint". *16th International Conference on Computational Linguistics (COLING'96)*. 41-46. Copenhagen, Denmark.

Gale, W. and K. Church. 1991. "Concordances for Parallel Texts". *Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*. Oxford.

Habert, B. and C. Jacquemin. 1993. "Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques". *Traitement Automatique des Langues* 34(2). Association pour le Traitement Automatique des langues, France.

Habert B. and A. Nazarenko and A. Salem. 1997. *Les linguistiques du Corpus*. Paris, Armand Colin.

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

Quaresma P. and I. Rodrigues and J.G.P. Lopes. 1998. "PGR Project: The Portuguese Attorney General Decisions on the Web". *The Law in the Information Society*. Instituto per la documentazione giuridica del CNR, C. Ciampi, E. Marinai (ed.), Florence, Italy.

Manning C. D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Silva, J. and G. Dias and S. Guilloré and J.G.P. Lopes. 1999. "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units". *9th Portuguese Conference in Artificial Intelligence.* Springer-Verlag.

Sinclair J.. 1974. "English Lexical Collocations: A study in computational linguistics". Singapore, reprinted as chapter 2 of Foley, J. A. (ed). (1996), J. M. Sinclair on *Lexis and Lexicography*, Uni Press.

Shimohata S. 1997. "Retrieving Collocations by Co-occurrences and Word Order Constraints". ACL-EACL'97.476-481.

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

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