Extracting Multiword Terms from Document Collections

Joaquim Ferreira da Silva
Universidade Nova de Lisboa, FCT/DI
Quinta da Torre, 2725, Monte da Caparica
jfs@di.fct.unl.pt

Gabriel Pereira Lopes
Universidade Nova de Lisboa, FCT/DI
Quinta da Torre, 2725, Monte da Caparica
gpl@di.fct.unl.pt

Abstract

Multiword terms (MWTs) are relevant strings of words in text collections. Once they are automatically extracted, they may be used by an Information Retrieval system, suggesting its users possible conceptual interesting refinements of their information needs. As a matter of fact, these multiword terms point to relevant information, often corresponding to topics and subtopics in the text collection, and maybe quite useful specially for highly refining generic queries.

In this paper, we introduce the *LocalMaxs* algorithm, for automatically extracting multiword terms. This algorithm requires neither empirically suggested thresholds nor complex linguistic filters nor language specific morpho-syntactic rules. These features make this algorithm a suitable approach to extract MWTs from text collections written in any language. Moreover, by introducing the Fair Dispersion Point Normalization concept, we can deal with arbitrarily long MWTs and can compare the results obtained by using different word association measures for MWTs selection. We also introduce our own association measure, the SCP, to work with the *LocalMaxs* algorithm, and assess the results obtained by comparing it with related statistics-based measures (Specific Mutual Information, Dice, *Loglike* and ϕ^2 coefficients) used in experiments on a text collection. An Information Retrieval application using our approach is also presented.

1 Introduction

In Information Retrieval (IR) it is currently accepted that multiword terms enhance IR precision. There are doubts about its role, namely about whether they should work as real indexes or they should play a special role in the refinement phase of user information needs.

Our proposal leans towards the second approach. We claim that for each document collection that one wants to make accessible to general or restricted public, besides the normal indexing of the document collection made by a normal inverted file based indexing and retrieval engine, there should be a phase for automatic extraction of multiword terms from the collection. These multiword terms should constitute a separate document. So, if we want to allow the access to *n* collections of documents we should produce *n* documents and each one of these documents should contain the multiword terms of each text collection. Currently used indexing machinery should also index the collection of multiword terms. Acting like this, an information need required from an IR system brings up a number of documents from each collection. Apart from the traditional refinement possibilities using key-words or document descriptors, the system we present [10] may suggest the user the multiword terms containing the words used in the initial query, enabling an information need refinement over the whole collection of collections or over a specific collection. It is left to the user to decide if he/her wants to search over the set of the document collections or over just one collection.

While word based queries show up an enormous number of documents per collection, the use of multiword terms suggested by this IR system dramatically prunes the search space to just a few documents. In this paper we focus on the automatic extraction of multiword terms for any kind of document collection independently of the language used on those documents.

2 Motivation

Information Retrieval in large text or document collections can be rather tedious and unrewarding. Every one has the experience on having queried any existing information retrieval engines (Yahoo, Lycos,…) about a topic such as *Human Rights,* and having received information pointing to hundreds of documents where "Human" and "Rights" do appear, not necessarily the term "Human Rights". This is not encouraging specially for non-specialist users as they may not know how to pose a question, and once they do succeed, it is not practical to read every text containing "Human", "Rights" or, maybe, "Human Rights". Moreover, knowing that there are hundreds of occurrences of that specific expression, it would be much more useful, if the IR system could help the user informing him/her about the existing relevant subject matters (subtopics) on the topic *Human Rights* in the text collection, such as *European Convention on Human Rights*, *European Court of Human Rights*, *European Commission of Human Rights*, *etc*[1]. This surely would contribute as a guide for more efficient

[1] See Appendix A

searches. On the other hand, considering that MWTs are relevant expressions in the documents, they work as representing topics and subtopics by which the information can be grouped. Having this "natural" division of the information available, the user may select the subject / subtopic which he/her is interested in, instead of being informed that there are hundreds of occurrences of an expression, e.g. "Human Rights".

So, IR systems performance can be enhanced by using multiword terms (MWTs) automatically extracted from the text collection the IR system is working with. By MWTs we mean a string of words strongly connected (associated) which is relevant from an IR perspective, in that text collection. Thus, a MWT can be a compound noun such as *Ministro dos negócios estrangeiros (Minister of foreign affairs); Direitos do Homem (Human Rights)*; or a string of words pointing to an important subject or event in one or more documents (*Comissario culpa produtores –commissioner blame producers*; *relativas à autonomização da Faixa de Gaza –relative to autonomy of the Gaza Zone*). There are also MWTs that should be declared as stop-terms (*no caso de –a Portuguese locution meaning 'if'; assim que seja possível –as soon as possible*; etc.).

3 Extracting Relevant Expressions

In order to extract the relevant multiword terms from text collections, we have used three tools [12] that work together:

-The *LocalMaxs* algorithm
-The Symmetric Conditional Probability (SCP) statistical measure
-The Fair Dispersion Point Normalization

Before explaining the *LocalMaxs* algorithm, we must consider some important concepts. So, a *n*-gram is a string of words in the text collection[2]. For example: the word "president" is a *1*-gram; the string "President of" is a 2-gram; the string "President of the Republic" is a 4-gram. Furthermore, we say that a *n*-gram is a *hole-free* or *uninterrupted n*-gram if every physical position of the *n*-gram is occupied by just one possible word, i.e., within a *hole-free n*-gram there is no physical position corresponding to a "hole" that can be occupied by any word of a finite set of words.

LocalMaxs algorithm works based on the idea that each *n*-gram has a kind of "glue" sticking the words together within the *n*-gram. Different *n*-grams usually have different "glues". As a matter of fact one can intuitively accept that there is a strong "glue" within the *n*-gram *(Margaret, Thatcher)* i.e. between the words *Margaret* and *Thatcher*. On the other hand, one can not say that there is a strong "glue" for example within the *n*-gram *(or, uninterrupted)* or within the *n*-gram *(of, two)*. So, let us take for granted that we have a function *g(.)*[3] that measures the "glue" of each *n*-gram.

3.1 The *LocalMaxs* Algorithm

The *LocalMaxs* is an algorithm that works with any text collection as input and automatically produces multiword terms (MWTs) from that text collection. In the context of *LocalMaxs*, we define:

An *antecedent* (in size) of the *hole-free n*-gram $w_1,w_2\ldots w_n$, $ant((w_1\ldots w_n))$, is a *hole-free* sub-*n*-gram of the *n*-gram $w_1\ldots w_n$, having size *n-1*. i.e., the *(n-1)*-gram $w_1\ldots w_{n-1}$ or $w_2\ldots w_n$.

A *successor* (in size) of the *hole-free n*-gram $M = (w_1,w_2\ldots w_n)$, *succ(M)*, is a *hole-free (n+1)*-gram *N* such that *M* is an *ant(N)*. i.e., *succ(M)* contains the *n*-gram *M* and an additional word before (to the left) or after (to the right) *M*.

- Let *W* be a *hole-free n*-gram; we say that *W* is a MWT if[4]: (3.1)

$$g(W) \geq g(ant(W)) \wedge g(W) > g(succ(W)) \quad \forall_{ant(W),succ(W)} \quad (\text{if } W\text{'s size} \geq 3)$$

$$g(W) > g(succ(W)) \quad \forall_{succ(W)} \quad (\text{if } W\text{'s size} = 2)$$

Where *g(.)* is a function that measures the "glue" sticking the words together within the considered *n*-gram.

3.2 The Symmetrical Conditional Probability (SCP) Measure

Let us consider the bigram *(x,y)*. We say that the "glue" value of the bigram *(x,y)* measured by *SCP(.)* is:

$$SCP(x,y) = p(x \mid y).p(y \mid x) = \frac{p(x,y)}{p(y)} \cdot \frac{p(x,y)}{p(x)} = \frac{p(x,y)^2}{p(x) \cdot p(y)} \qquad (3.2)$$

[2] We use the notation $(w_1\ldots w_n)$ or $w_1\ldots w_n$ to refer the *n*-gram of length *n*.

[3] We will write *g(W)* for the *g(.)* value of the generic *n*-gram *W*, and $g((w_1\ldots w_n))$ for the *g(.)* value of the *n*-gram $w_1\ldots w_n$ once we want to keep *g(.)* as a one-argument function. We will instanciate this generic function by using various *n*-gram word association functions, namely *SCP(.)*, that will be a one-argument function too. So, we can write for example *SCP(W)*, $SCP((w_1,w_2,w_3))$, $SCP((w_1\ldots w_n))$, etc.

[4] Since a MWT must be a relevant *n*-gram, *LocalMaxs* does not produce MWTs with just one occurrence in the text collection. This criterion was applied just because it reduces drastically the processing time mainly in large text collections and augments precision.

Where *p(x,y)*, *p(x)* and *p(y)* are the probabilities of occurrence of the bigram (*x,y)* and unigrams *x* and *y* in the text collection; *p(x|y)* stands for the conditional probability of occurrence of *x* in the first (left) position of any bigram given that *y* appears in the second (right) position of the same bigram. Similarly *p(y|x)* stands for the probability of occurrence of *y* in the second (right) position of any bigram given that *x* appears in the first (left) position of the same bigram.

3.3 The Fair Dispersion Point Normalization

Considering the denominator of the equation 3.2 we can think about any *n*-gram as a "pseudo-bigram" having a left part (*x*) and a right part (*y*). The Fair Dispersion Point Normalization or simply Fair Dispersion "transforms" any *n*-gram of any size into a "pseudo-bigram" and reflects the "glue" between each two adjacent words of the whole original *n*-gram. Thus, applying the Fair Dispersion Point Normalization to the *SCP(.)* measure in order to measure the "glue" of the *n*-gram $w_1 \ldots w_n$, the denominator of the equation 3.2 is changed to:

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(w_1 \ldots w_i) \cdot p(w_{i+1} \ldots w_n) \quad (3.3)$$

So, we will have

$$SCP\_f((w_1 \ldots w_n)) = \frac{p(w_1 \ldots w_n)^2}{Avp} \quad (3.4)$$

In *SCP_f(.)* we have added *"_f"* for "fair" (from Fair Dispersion) to *SCP(.)*.

As it is shown in [12], the Fair Dispersion Point Normalization concept can be applied to other statistical measures in order to obtain a "fair" measure of the association / "glue" of any *n*-gram with *n > 2*.

3.4 Choosing a Suitable Statistical Measure for Information Retrieval Purposes

In order to extract MWTs, we needed to measure the "glue" of every *n*-gram and we had to choose a statistical measure to work with the *LocalMaxs* algorithm concerning IR purposes. Thus, several measures were tested using the Fair Dispersion normalization[5] namely: the Specific Mutual Information (SI) [3], [4] and [1], the SCP [12], the Dice coefficient [7], the *Loglike* coefficient [8], and ϕ^2 *coefficient* [9]. Table 1 contains scores for these statistical measures working with the Fair Dispersion Point Normalization and the *LocalMaxs* algorithm. We have used an average size text collection (919,253 words)[6].

3.4.1 The Evaluation Criterion

The *LocalMaxs* algorithm extracts *n*-grams, which are potential MWTs. In order to decide if an extracted *n*-gram is a MWT or not, we considered as MWTs the relevant expressions: proper names, such as *Butros Butros-Gali, Republica Centro Africana*[7]*, etc.*; compound names, ex: *câmara municipal de Lisboa (Lisbon town hall), convenção dos Direitos Humanos (Human Rights convention)*, entrada em vigor *(coming into force)* etc. and other relevant *n*-grams from the IR perspective, e.g. *afectadas pela guerra civil (affected by the civil war), detido por tráfico de droga (stopped for dealing in narcotic drugs)*, etc.

There are other interesting MWTs, extracted by the *LocalMaxs* algorithm working with some of the statistical measures that we are testing here. Although those MWTs are interesting units in the lexical and/or morphological point of view, they can not all be considered MWTs, as they have no meaning for IR purposes, e.g. some frozen phrases such as *em todo o caso (in any case), segundo consta (according with what is said), mais ou menos (more or less), etc.*

These considerations were taken into account for evaluating the suitability of each statistics-based measure that we tested.

---

5 As a matter of fact, any *n*-gram can be divided in a left and a right part choosing any point between two adjacent words within the *n*-gram. In this way, one can measure the "glue" using some usual statistical measure without the Fair Dispersion, but the results are relatively poor. The enhancements obtained in Precision and Recall when the Fair Dispersion is introduced in several statistical measures are shown in [12].

6 This text collection corresponds to some news from *Lusa* (the Portuguese News agency) in January 1994.

7 Note the spelling error in 'Republica' that should have been written as 'República'. However real *text collection* is like that and we can not escape from it as there are texts that may reproduce parts of other texts where the graphical form of words does not correspond to currently accepted way of writing.

3.4.2 The results

Table 1 – The scores for statistics-based measures

Statistics-based measure: g(.)=	Precision (average)	Extracted MWTs (count)
SCP_f(.)	81.00%	24476
SI_f(.)	75.00%	20906
ϕ^2*_f(.)*	76.00%	24711
Dice_f(.)	58.00%	32381
LogLike_f(.)	53.00%	40602

The Precision column means the average percentage of correct MWTs obtained under the criterion of the section 3.4.1. It is not possible to calculate the exact number of MWTs in the text collection, so that we may measure how close to that number is the number of extracted MWTs obtained by each statistics-based measure (Recall). As a matter of fact we are not facing the problem of counting very well defined objects like nouns or verbs of a text collection, but counting relevant expressions for IR purposes. So, the column Extracted MWTs, which gives the number of MWTs extracted by each considered measure, works as an indirect measure of Recall. (Remember that we have discarded every "MWT" that occurred just once).

Although there are very large MWTs, for example the 9-gram *protestos contra as violações dos direitos humanos em Timor Leste (protests against violations of human rights in East Timor)*, for reasons of processing time, in this test we have limited the MWTs produced by the *LocalMaxs* algorithm from 2-grams to 7-grams.

3.4.3 Discussion of the results

As we can see from table 1, *SCP_f* measure gets the best Precision and a comparatively good value for Extracted MWTs. By using the *LocalMaxs* algorithm with any of the statistics-based measures *SCP_f, SI_f* or ϕ^2*_f,* Precision was increased from approximately 50% to the values presented above (see [12]). However *SI_f* has a relative low score for Extracted MWTs. *Dice_f* and specially the *Loglike_f* measure showed to be not very selective. They extract many expressions (high frequency for Extracted MWTs), but many of them are not relevant, they just have high frequency, ex: *dar ao (to give to the), dos outros (of the others),* etc. Moreover, as it is discussed in [13], the Dice and the *Loglike* based measures do extract a lot of uninteresting units and fail to extract other units that are interesting and are selected by the other three word association measures. Thus, we have chosen the *SCP_f* measure to work with *LocalMaxs* algorithm in order to extract relevant expressions, for IR purposes (MWTs), from text collections.

4 Retrieving Information from Large Document Collections

Recently we have been working on an IR project (PGR[8]), using a 2,722,476-word text collection. We have applied our approach to this text collection[9]. In this section we assess the results.

4.1 The results

Using the *LocalMaxs* algorithm and the *SCP_f* measure, we have attained 84% Precision and 94,039 MWTs from this 2,722,476-word text collection. In this experience the *LocalMaxs* algorithm was prepared to produce MWTs from 2-grams to 8-grams. Table 2 shows the number of extracted MWTs versus the number of occurrences in the text collection for some typical searches.

Table 2 – The number of extracted MWTs versus the number of occurrences in the text collection for typical searches

Topic	Extracted MWTs (count)	Occurrences in the text collection
Direito Civil (Civil Code)	5	40
Direito Penal (Penal Code)	4	38
Direitos Humanos (Human Rights)	15	90
Direito Internacional (International Code)	9	32
Droga (narcotic drugs)	17	96

4.2 Discussion

Table 2 shows that there are 5 distinct MWTs for the topic "Direito Civil", but there are 40 occurrences of that expression ("Direito Civil") in the text collection. As we can see by appendix A, these 5 MWTs are: *Direito Civil (Civil Code), Noções Fundamentais do Direito Civil (Fundamental Notions of the Civil Code),*

[8] PGR means *Procuradoria Geral da República* (Portuguese General Attorney).

[9] The application is accessible in http://coluna.di.fct.unl.pt/~pgrd

Teoria Geral de Direito Civil (General theory of Civil Code), Teoria Geral do Direito Civil[10] *(General theory of the Civil Code)* and *Simulação em Direito Civil (Simulation in Civil Code).* As we can see, these *MWTs* correspond to subtopics of the topic "Direito Civil" concerning this concrete text collection. As a matter of fact, subtopics are dependent of the text collection we are using. If we had a larger text collection having just documents about "Direito Civil", we would be able to extract many more subtopics than we are extracting now, which seems correct from an IR perspective. However, our approach extracts some MWTs that are not perfect subtopics (they are just part of it), e.g. *Europeu dos Direitos do Homem (European of Human Rights)* –see appendix A for the topic "Direitos do Homem" – these kind of MWTs are still useful, since they clearly point to something important, as the user can easily guess: this MWT should point to *Tribunal Europeu dos Direitos do Homem (European Court of Human Rights)*, which is not extracted. The reason why this is not extracted is because the "glue" value of the *6*-gram *Tribunal Europeu dos Direitos do Homem* is lower than the "glue" value of the *5*-gram *Europeu dos Direitos do Homem*, as the word *Tribunal (Court)* is very frequent is this text collection. As a matter of fact, the *SCP_f* measure, as well as the *SI_f* and the ϕ^2*_f*, penalizes the "glue" values of the *n*-grams starting or ending in stop words. Though *Tribunal* is not a stop word, it is very frequent in this specific text collection. See also *LocalMaxs* algorithm definition (3.1) for a better understanding about the *n*-grams which are not extracted as MWTs.

As we said, appendix A contains samples of extracted MWTs corresponding to topics and subtopics from this text collection.

5 Conclusions and Future Work

Using the *LocalMaxs* algorithm, the SCP measure and the Fair Dispersion Point Normalization -*SCP_f,* it is possible to extract relevant multiword terms. From an Information Retrieval perspective, these multiword terms point to relevant information, often corresponding to topics and subtopics in the text collection.

The approach used in this work can extract relevant expressions in several languages [11], since *LocalMaxs* approach does not depend on morpho-syntactic information. This makes this approach a very useful tool for cross language Information Retrieval.

In related literature, the extraction of relevant expressions has been based mostly on statistical measurement of the *2*-grams and *3*-grams frequency and on the application of more or less complex linguistics filters and specific morpho-syntactic operations [2] and [5]. The work we present in this paper, confirms that the *LocalMaxs* algorithm is a more robust approach, reducing the commitments associated to that complexity and take a greater advantage of the statistics. The results presented in tables 1 and 2 confirm also the usefulness of the Fair Dispersion Point Normalization. Through the introduction of this concept, the whole *n*-gram's "glue" is reflected in a *normalised* "pseudo-bigram" whatever the *n*-gram's length is. This concept was tested for different statistical measures and gave rise to great improvements to every previously used word association measures [12].

The *LocalMaxs* algorithm detects the *n*-grams that must be elected under the criterion explained in section 3.1, being a serious alternative to threshold based approaches.

In order to choose a suitable statistical measure to calculate the "glue" that sticks the words together in every possible contiguous *n*-gram, several statistics-based measures were tested: Specific mutual information (SI), Symmetrical Conditional Probability (SCP), *Loglike*, ϕ^2, and Dice coefficients. These measures were tested using the Fair Dispersion Normalization. Despite the good results obtained with *SI_f* and ϕ^2*_f* word association measures, the *SCP_f* measure presented the best Precision (81%) for a medium size text collection and extracted a relatively high number of MWTs. That's why we have chosen this measure to work with the *LocalMaxs* algorithm for a larger text collection, considering the Information Retrieval purpose of the project we have been working (PGR). Due to the size augmentation of the working *corpus*, in this larger text collection an 84% Precision was obtained and over 94,000 MWTs were extracted.

The availability of the text collection's MWTs enabled a more natural and comfortable interaction with the novice users, not knowing which descriptors have been assigned to each opinion of the General Attorney. Moreover, the text collection's MWTs extracted by this approach work as a guide for generic searches, which is very useful, as we will show in appendix A.

The same approach has also been used in the General Chronicle of Spain written in Medieval Portuguese, and the results once again showed how this methodology can be an important support for lexicographers [13]. The *LocalMaxs* algorithm has also been useful for working with another word association measure, the Mutual Expectation measure, on discontinuous or non-*hole-free n*-grams [11] and [6].

[10] The expressions *Noções Fundamentais **do** Direito Civil* and *Noções Fundamentais **de** Direito Civil* means the same, but they appear in these forms in the *text collection.*

We stress that the organization of automatically extracted terms from text collections into a hierarchy of concepts, is useful for visualization. Additionally, work must still be done in order to automatically classify documents and aid judges to fill in the information on key-words, law areas, referred laws and opinions, etc.

Appendix A

This appendix contains samples of MWTs extracted from the *PGR* text collection with 2,722,476 words. The MWTs were extracted by using the *LocalMaxs* algorithm, and the *SCP_f* statistics-based measure. We signal by "*" those terms which by themselves are not terminology terms.

Topic: Direito Civil (Civil Code)

Direito Civil (Civil Code)
Noções Fundamentais do Direito Civil (Fundamental Notions of the Civil Code)
Teoria Geral de Direito Civil (General theory of Civil Code)
Teoria Geral do Direito Civil (General theory of the Civil Code)
Simulação em Direito Civil (Simulation in Civil Code)

Topic: Direito Penal (Penal Code)

Assistência Jurídica Mútua nas Matérias de Direito Penal (Mutul Juridical assistance in Penal Code matters)
Matérias de Direito Penal (Penal Code matters)
Direito Penal Fiscal (Fiscal Penal Code)

Topic: Direitos do Homem (Human Rights)

Direitos do Homem e das Liberdades Fundamentais (Human Rights and of Fundamental liberties)
** Europeu dos Direitos do Homem não contém (European of Human Rights does not contain)*
Convenção de Salvaguarda dos Direitos do Homem (Safeguard Convention of Human Rights)
Declaração Universal dos Direitos do Homem (Universal Declaration of Human Rights)
** Direitos do Homem não contém disciplina (human Rights does not contain discipline)*
Extinção da Comissão Europeia dos Direitos do Homem (Extinction of European Commission of Human Rights)
Protecção dos Direitos do Homem e das Liberdades (Protection of Human Rights and Liberties)
** 6 da Convenção Europeia dos Direitos do Homem (6 of the European Convention of Human Rights)*
Convenção para a Protecção dos Direitos do Homem (Convention for the Protection of Human Rights)
** Europeia dos Direitos do Homem (European –* convention | commission*– of Human Rights)*
** Europeu dos Direitos do Homem (European of Human Rights)*
Salvaguarda dos Direitos do Homem (Safeguard of Human Rights)
Direitos do Homem (Human Rights)

Topic: Direito Internacional (International Code)

Conferencia da Haia de Direito Internacional Privado[11]
Conferencia da Haia do Direito Internacional Privado[13]
Conferencia de Haia de Direito Internacional Privado (Private International Code Conference of Haia)
elaborado pela Comissão de Direito Internacional (developed by the Commission of International Code)
** Haia de Direito Internacional Privado (Haia of Private International Code)*
** Haia do Direito Internacional Privado (Haia of Private International Code)*
Lições de Direito Internacional Privado (Lessons of Private International Code)
Comissão de Direito Internacional (Commission of International Code)[12]
Comissão do Direito Internacional (Commission of International Code)[14]

Topic: droga (narcotic drugs)

** sentido de combater o tráfico de droga (-in- order to fight the dealing in narcotic drugs)*
** drogas no mar não suscita juízos (narcotic drugs in the sea does not promote judgements)*
repressão do tráfico ilícito de drogas (repression of the illegal dealing in narcotic drugs)
tráfico de droga pelos países produtores (dealing in narcotic drugs by the producing countries)
tráfico ilícito de drogas no mar (illegal dealing in narcotic drugs in the sea)
combater o tráfico de droga pelos países produtores (to fight the dealing in narcotic drugs by the producing countries)

---

[11] These two MWT corresponds to the subtopic *conferência de Haia de Direito Internacional Privado (Private International Code Conference of Haia).* They correspond to correctly spelled existing in the text collection. That correspond to indecisions prior to terminological freezing of expressions.

[12] These two MWT correspond to the subtopic *Comissão de Direito Internacional (Commission of International Code).* They correspond to correctly spelled existing in the text collection. That correspond to indecisions prior to terminological freezing of expressions.

** drogas no mar não suscita juízos de desconformidade (narcotic drugs in the sea does not promote discordance judgements)*

repressão do tráfico ilícito de drogas no mar (repression of illegal dealing in narcotic drugs in the sea)

combater o tráfico de droga (to fight the dealing in narcotic drugs)

recuperação dos toxicodependentes da droga (recuperation of the narcotic drugs addicts)

** droga pelos países produtores (narcotic drugs by the producing countries)*

tráfico ilícito de droga (illegal dealing in narcotic drugs)[13]

tráfico ilícito de drogas (illegal dealing in narcotic drugs)[15]

drogas no mar (narcotic drugs in the sea)

toxicodependentes da droga (narcotic drugs addicts)

tráfico de droga (dealing in narcotic drugs)

tráfico de drogas (dealing in narcotic drugs)

References

1. Bahl, L., & Brown, P., Sousa, P., Mercer, R.: Maximum Mutual Information of Hidden Markov Model Parameters for Speech Recognition. In Proceedings, International Conference on Acoustics, Speech, and Signal Processing Society, Institute of Electronics and Communication Engineers of Japan, and Acoustical Society of Japan (1986)
2. Bourigault, D., Jacquemin, C.: Term Extraction and Term Clustering: an Integrated Platform for Computer Aided Terminology. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, p. 15-22, Bergen, Norway June (1999)
3. Church, K. et al.: Word Association Norms Mutual Information and Lexicography, Computational Linguistics, Vol. 16 (1). (1990) 23-29
4. Church, K., Gale, W., Hanks, P., Hindle, D.: Using Statistical Linguistics in Lexical Analysis. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, edited by Uri Zernik. Lawrence Erlbaum, Hilldale, New Jersey (1991) 115-165
5. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act Combining Symbolic and Statistical Approaches to Language, MIT Press (1995)
6. Dias, G., Gilloré, S., Lopes, G.: Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora. In Proceedings of the TALN'99, p. 333-338. Corse, july 12-17 (1999)
7. Dice, L.: Measures of the Amount of Ecologic Association between Species. Journal of Ecology, 26: 297-302 (1945)
8. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, Association for Computational Linguistics, Vol. 19-1. (1993)
9. Gale, W.: Concordances for Parallel Texts, Proceedings of Seventh Annual Conference of the UW Centre for the New OED and Text Research, Using Corpora, Oxford (1991)
10. Quaresma, P., Rodrigues, I., Lopes, G., Almeida, T., Garcia, E., Lima, A.: "PGR project: the Portuguese Attorney General Decisions on the Web". Proceedings of the law in the Information Society. Florence, December (1998)
11. Silva, J., Dias, G., Guilloré S., Lopes, G.: Using *LocalMaxs* Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. Lectures Notes in Artificial Intelligence, Springer-Verlag, Vol. 1695, p. 113-132 (1999)
12. Silva, J., Lopes, G.: A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In Proceedings of the 6th Meeting on the Mathematics of Language, p. 369-381. Orlando, July 23-25 (1999)
13. Silva, J., Lopes, G., Xavier, M., Vicente, G.: Relevant Expressions in Large Corpora. In Proceedings of the Atelier-TALN99, Corpus et TAL, p. 86-94. Corse, july 12-17 (1999)

[13] These two MWTs correspond to the subtopic *tráfico ilícito de droga (illegal dealing in narcotic drugs).* They correspond to correctly spelled existing in the text collection. That corresponds to indecisions prior to terminological freezing of expressions.