Terms Spotting with Linguistics and Statistics

Joana L. Paulo (L^2F - INESC-ID / IST)
joana.paulo@l2f.inesc-id.pt
Nuno J. Mamede (L^2F - INESC-ID / IST)
nuno.mamede@l2f.inesc-id.pt

Rua Alves Redol 9, 1000-029 Lisboa, Portugal

Abstract

Terminologies are useful in all areas that use specialized languages. The development of terminologies is a hard work, when manually done. It can be assisted with tools to ease and improve the achievement of such a work. In this article, we present ATA, an automatic terms extractor that uses both linguistic and statistical information. This tools was tested with Portuguese language but can be easily extended to other languages.

1 Introduction

In the last few years, computational linguists, applied linguists, translators, interpreters, scientific journalists and computer engineers have been interested in automatically extracting terminology from texts. Different goals have led these professional groups to design software tools to directly extract terminology from texts, basically, all kind of Natural Language Processing (NLP) applications that work with specialized domains and that consequently need special vocabulary.

ATA, is an Automatic Term Acquisition System that processes technical texts and produces a list of noun phrases likely to be terminological units in that field. It was tested using Portuguese language but is easily extendable to any other language.

This article opens with some background knowledge. Then structural and functional aspects of ATA are presented: the architecture, the input and output data and the main process, that responsible for the term's extraction. After this, we describe the evaluation process enumerating some ATA's results.

2 Background

In this context, a *term* is a linguistic representation of a concept by means of a simple noun or a noun phrase [11]. We consider two term types: simple and compound. Other phraseological structures characterizing some knowledge domains are not in the scope of ATA.

Simple terms consists of a single lexical unit, a graphical word. The complexity associated with the detection of this kind of terms arises from their unremarkable appearance. This means that there is no way for one to be distinguished from another, unless

the system has a morphological structure analyzer which can sort term-candidates by the occurrence of specific affixes or roots which is not the case of ATA.

Compound terms consists of more than one lexical unit (graphical form). Thus, they are less prone to ambiguity than simple terms. Nevertheless, they require a previous syntactical study to verify whether a set of words actually defines a term's syntactical structure.

According to several works [15, 12, 5, 11] all lexical units have an associated frequency corresponding to the number of times they appear in a corpus. Using this information we can decide whether a word can eventually be a term: items that are nouns and that appear more than a given number of times can be considered as candidates to be simple terms; words with other categories must be kept in order to complete the processing of compound terms.

Most systems designed for this kind of task take a plain text and extract from it a list of candidate terms. To make the terminologists' task easier, this list is provided with its context and assorted additional information (such as relative frequency for that word and for its root).

The most used techniques for this task are:

Statistical based systems: detect lexical units whose frequency is higher than a given corpus-based threshold definition. The problem with this approach is that it fails to detect low-frequency terms.

Systems that use linguistic knowledge: detect recurrent patterns from complex terminological units such as noun-adjective and noun-preposition-noun. Patterns to be detected are assumed to have been designed by linguists.

Hybrid systems: start detecting some basic linguistic structures, such as noun or prepositional phrases, and then, after the candidate terms have been identified, the relevant statistical information is used to decide whether they correspond to a term. This will be our methodology.

3 ATA's Structural and Functional Aspects

In this section we describe ATA's structure and functionality. First we take into account the application's structure, that is, the architecture of the system. Then the input and the output data are described. We kept the description of the main process for the next section.

The used architecture (fig. 1) is similar to the proposed for simple terms in [7]. Taking the system as a whole, it first lemmatizes the text using an external dictionary (SMorph). The resulting text is then passed to a post-morphological processor (PAsMo) that detects and forms special groups according to recomposition and correspondence rules. The system groups the words in phrases (the phrase separators have been previously described in another external file). Before the main process begins, the text is sent to a syntactic analyzer (SuSAna) that, using a surface grammar groups the phrase constituents. Then the main process extracts the words candidate to terms (GeTerms). After this, a statistical-based process (Anota) enriches the candidate list with the relevant statistical information. The last process (Decision) evaluates the candidate list and filters those that can be terms according to the syntactic and statistical restrictions.

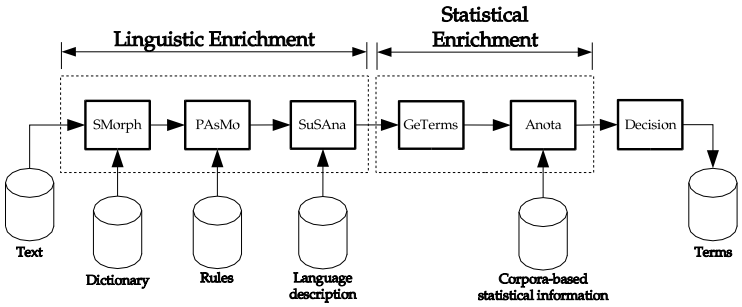

Fig. 1. Main Architecture.

3.1 SMorph: tokenization, lemmatization, morphological analysis

Enriches each word with its morphological characterization. This step needs the dictionary file, which contains not only simple units (corresponding to graphical forms), but also complex units, such as prepositional and adverbial locutions. For this we use SMorph [3] that allows the construction of large dictionaries required for the linguistic analysis of texts. The user declares the dictionary by specifying five types of rules, which are converted into a compact binary file containing a finite state automata. The dictionary is used for generating all inflected forms of a lemma and for segmenting and analyzing a text.

3.2 PAsMo: post-lemmatization

Rewrites the text according to rules based on the morphological features of the words and breaking it into sentences (according to the chosen punctuation). This is done using recomposition and correspondence rules and a list of symbols to be used as breaks in the text. For this we use PAsMo [1]. that can rewrite dates, compound nouns, consecutive unknown words, numbers, and so on. PAsMo may also be used to translate tags, thus facilitating the interface with the syntactic analyzer.

3.3 SuSAna: syntactic analysis

Is the stage where we analyze the text grouping the words of each phrase into syntactic chunks. This process is done using surface grammars. For this we use SuSAna [9, 4]. Since SuSAna allows to select the syntactic chunks we want to extract, the first selection is made at this point. SuSAna proved to be a good choice as the extension of the existent rules was easy and simple.

3.4 GeTerms: candidate terms selection

At this point all the linguistic enrichment is done and we are able to select the candidate terms according to the linguist restrictions and to sum up the number of occurrencies of each one.

3.5 Anota: statistical enrichment

In order to compare the frequencies observed in the specialized text and the reference ones we have to join this information. This isolates the decision process with proved to be a good decision when we wanted to test the system's performance. If one of the criteria when finding terms is the frequency of a word when comparing with general corpora, we need to extract frequencies for each lemma. For that the same architecture will be used enriched with a tool for resolve ambiguity that remains at the end of the process chain. The corpora we are going to use is CETEM-Público [14].

3.6 Decision: terms selection

Finally, when all the information was collected we can decide for each candidate whether it is a term. This is done according to two parameters: the minimal number of occurrencies; a comparison factor between the occurrence in the specialized text and in the reference.

4 Evaluation

The development of noun phrases extractors is a very delicate task constrained by robustness and accuracy.

Robustness is subject to a strong restriction: it can be used over a wide range of unrestricted texts gathered in large corpora. This means that it has to be domain-independent, that is, it cannot use any a priori semantic or conceptual information. From the point of view of the surface syntactic analysis, the extraction is more difficult, since the system is domain-independent because each domain can have specific restricted surface structures ([8] restricted the extraction to medical terms which have few possible nominal structures).

Accuracy is also an issue because the noun phrases extracted by the system are the candidate terms that will be proposed to the user building a domain's terminology.

The two most frequently used metrics in the evaluation of this type of system are recall and precision [13]. Recall is defined as the relationship between the sum of retrieved terms and the sum of existing terms in the document that is being explored. Precision accounts for the relationship between those extracted terms that are actually terms and the aggregate of candidate terms that are found. These metrics can be seen as the capacity to extract all terms from a document (recall) and the capacity to discriminate between those units detected by the system which are terms and those which are not (precision).

Systems based on linguistic knowledge tend to use noise and silence as a measure of efficiency. Noise attempts to assess the rate between discarded candidates and accepted ones; silence attempts to assess those terms contained in an analyzed text that are not detected by the system. Errors in the assignment of morphological categories or syntactic analysis are also shared by these systems.

The system was evaluated using the described methods (see 4), hopping to achieve results similar to those of systems for Portuguese [6, 16] and for other foreign languages as English [12] and French [7].

The tests were done starting with an nautical annotated corpus. A linguistic identified the terms in the text. Then we used to system to get all the proposed terms. The described measures were calculated comparing the results from both tasks. The results were as follows.

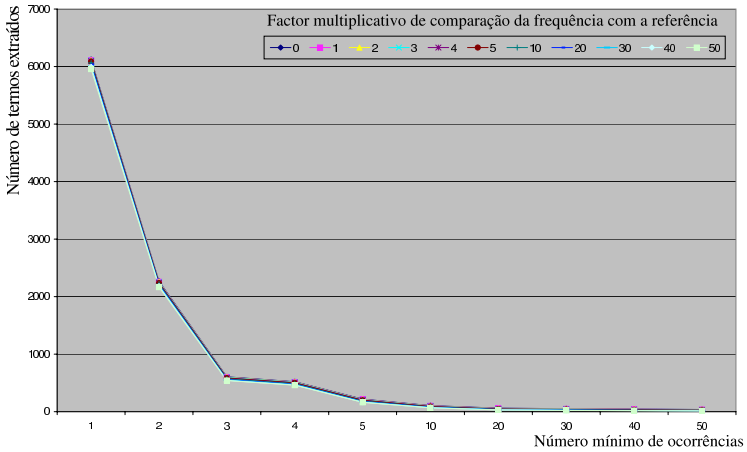

Fig. 2. Detected Terms.

Figure 2 shows the number of terms detected when, for the same input, we change the extraction parameters. From left to right we have the minimum of occurrencies we ask to a candidate appear in order to consider it as a term. The different lines represent the variation with the multiplicative factor between the specialized frequency and the reference one. As expected, as bigger the parameters, lower the number of detected terms. A detailed analysis of this variations can be found in [2].

The price to pay for an automatic acquisition without any intermediary human validation is twofold: the procedure can let relevant information pass through undetected; it can acquire false information. This is the reason why perfecting these procedures requires the adoption of experimental processes, with numerous tests carried on large-scale corpora, that ensure the global empirical validity of the procedures.

Figure 3 shows the variation of precision and recall when the minimum of occurrencies requested increases. Observing it we find that the best value for this parameter is 3. With this configuration we have the best equilibrium between precision and recall, that is, the noise isn't to much and still we can detect same terms.

5 Conclusions and Future Directions

In this article a new system for automatic term acquisition has been described. We are especially interested in studying the capability and the implications of building an automatic term acquisition system.

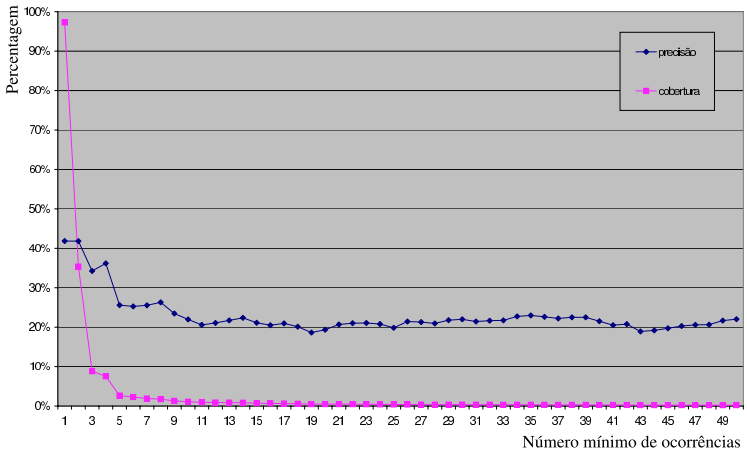

Fig. 3. Precision vs. Recall.

The ATA system proved to be a useful helper in solving the problem of the semi-automatic building of terminological indexes and will be used on different kinds of specialized documents. However the results can be better if we work on the linguistic resources. These aren't accurated at the present time and being at the initial stage any linguistic error has is effects on the all process. Some work is already being done hear.

We think that knowledge acquisition for knowledge-based systems is also a suitable experimentation ground for such a terminology extraction system, provided that an appropriate tool exists to represent and record information.

Changing language, currently, means retrieving linguistic resources for each tool. Nevertheless, the architecture eases the task as the only thing that as to be changed are the inputs. Such an extension would need the new language's dictionary, analyzing the morphological behavior of that language to write new rules, analyzing a general corpus and computing word frequency and getting a surface grammar. That work can be done in order to prove that the system is language-independent.

Format information could be considered by the system. The text format depends on the editor. It is necessary to create an external format description. If the text format has been lost at some point, the initial text has to be considered. The idea is to give more or less importance to a word according to its format [10]. For instance, if a word appears in a title it must be considered more important that a word that is in the middle of a paragraph. Thus, it is possible to create a hierarchical classification that considers input text format. This classification should consider: Titles of documents, sections, and subsections; Bold, italic, and underline; Type and size of letter; Caps usage; Headers and footers; Footnotes; Quotations; Other styles.

References

1. PAsMo - pós-análise morfológica. Technical report, 2001.
2. ATA - aquisição de termos automática. Master's thesis, 2004.

3. Salah Aït-Mokhtar. *L'analyse Présyntaxique en une seule étape*. PhD thesis, Université Blaise Pascal, 1998.
4. Fernando Batista. Análise sintáctica de superfície e consistência de regras. Master's thesis, Instituto Superior Técnico, UTL, 2002. (work in progress).
5. D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'92*, 1992. p. 977-981.
6. J. Ferreira da Silva and G. Pereira Lopes. A local maxima method and a fair dispersion normalization for extracting multi-words units from corpora. *International Conference on Mathematics of Language, Orlando*, July 1999.
7. 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.
8. Rosa Estopà. *Les unitats terminológiques polilexemàtiques en els lèxics especialitzats: dret i medicina*. PhD thesis, Institut Universitari de Lingüística Aplicada, Barcelona, UPF, 1999.
9. Caroline Hagège. *Analyse syntaxique automatique du portugais*. Thèse de doctorat, Université Blaise Pascal, GRIL, Clermont-Ferrand, 2000.
10. C. Jacquemin. Quelques exemples d'application du traitement automatique d es langues en accès à l'information. *5emes Journées Internationales d'Analyse de Données Textuelles (JADT)*, 1, 2000.
11. C. Jacquemin and D. Bourigault. Term extraction and automatic indexin. *R. Mitkov, editor, Handbook of Computational Linguistics*, 2000.
12. J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, p. 9-27*, 1995.
13. C. D. Manning and H. Shutze. *Foundations of Statistical Natural Language Processing*. MIT Press, London, 1999.
14. MCT and Público. Cetempúblico - corpus de extractos de textos electrónicos, 2000.
15. A. P. Marquez Neto. Terminologia e corpus linguístico. *Revista Internacional de Língua Portuguesa - RILP n. 15, p. 100-108*, 1996.
16. J. Silva, G. Dias, S. Guilloré, and G. Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *9th Portuguese Conference on Artificial Intelligence*, 1695:113–132, September 1999.