ATA – Automatic Term Acquisition

Joana Lúcio Paulo, Nuno J. Mamede

Instituto Superior Técnico – Universidade Técnica de Lisboa
L^{2}F – INESC-ID
R. Alves Redol, nº 9, Apartado 13069, 1000-029 Lisboa, Portugal
Joana.Paulo@RNL.IST.UTL.Pt, Nuno.Mamede@ACM.Org

Abstract: Terminological acquisition is an important issue when learning about Natural Language Processing (NLP) due to the constant terminological renewal caused by technological changes. Terms play a key role in several NLP activities such as machine translation, automatic indexing, text understanding, and information retrieval. This is especially true at this time when corpora in electronic format keep growing in number and variety. In this work we start by using morphological and syntactic information to locate candidate noun phrases, and then we use statistical information to improve result accuracy.

Introduction

The system we are about to present, ATA, is an Automatic Term Acquisition System that processes any given text and creates a list of nouns and noun phrases likely to be terminological units in that text.

The development of noun phrases extractors is a very delicate task constrained by robustness and accuracy. Robustness is an issue because it is subject to a strong restriction: that it can be used over a wide range of unrestricted texts gathered in large corpora. This means that the system has to be domain-independent. 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.

Structural and Functional Aspects

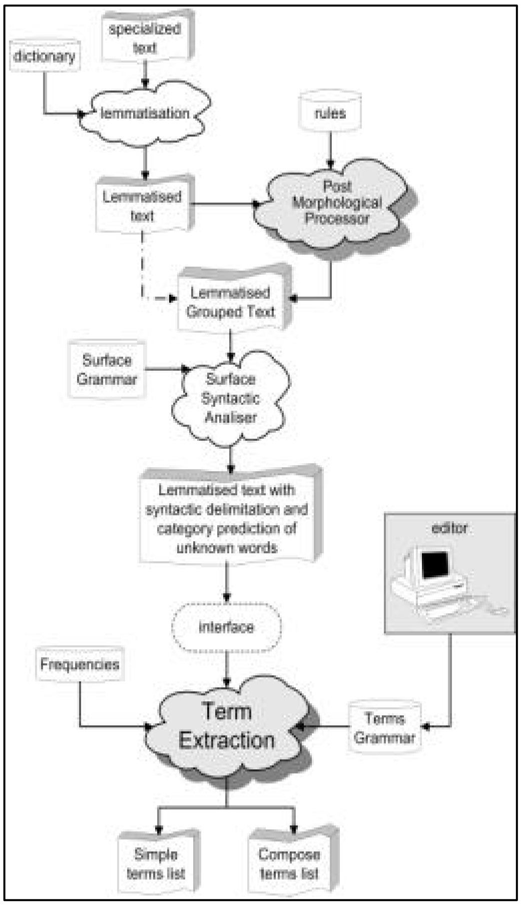
A design choice was made between performance and adaptability and between efficiency and flexibility: instead of hardwiring the linguistic information in the program, external files and associated scanning code are provided. This option allows us more flexibility and easier configuration while keeping the increased complexity at a minimum.

The general process is presented in Figure 1.

Taking the system as a whole, i.e., if we consider the larger system containing ATA, the input consists of a plain text, a dictionary, a set of rules defining word grouping, and a surface grammar, otherwise, ATA's input data consist of plain text previously analyzed both morphologically and syntactically.

ATA’s first lemmatises the text using an external dictionary. The resulting text is then passed to a post-morphological processor that detects and forms special groups according to recomposition and correspondence rules. The system groups the words in phrases (the phrase separators are described in another external file).

Before the main process begins, the text is sent to a syntactic analyser that, using a surface grammar groups the phrase constituents. The main reason for using a surface grammar is the efficiency and the robustness of the process.



Then the main process extracts the words that are term candidates. After that, a statistical-based process evaluates the candidate lists. At this stage, a file containing the frequency of each word computed over a general-content corpus will give us information that allows us to detect simple terms by comparing word frequencies of the entities in the text being analyzed and the ones in the corpus.

The output is finally produced and sent to the system's user. ATA's output is divided into two sets both of which may be empty. The first set contains simple terms identified in the text. The second set contains compound terms detected in the text. A specialist should confirm the correction of both sets. Given the characteristics of simple terms, their processing starts by collecting all input words appearing with frequency higher than the threshold defined by the corpus. From this collection, words classified as nouns may be considered as simple term candidates. Those classified as unknown but considered as nouns due to the syntactic process are also considered as candidates. The candidate list is going to be processed before output production. A noun is considered as a simple term if it occurs with a higher frequency than corpus-based threshold.

Compound terms obey some syntactical and grammar restrictions that make the detection process easier. The structure of the word groups that are to be considered as compound term candidates are described in an external file. A group of words is a compound term if it respects a previously defined structure. Like simple terms, compound terms also have frequency constraints.

In addition, the system must distinguish between high-frequency words occurring within compound terms and isolated occurrences of the same words. Format information must also be considered by the system. The text format depends on the editor. It is necessary to create an external format description. The idea is to give more or less importance to a word according to its format.

Evaluation

The two most frequently used metrics in the assessment of this type of system are recall and precision. 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 analysed text that are not detected by the system. Errors in the assignment of morphological categories or syntactic analysis are also shared by these systems.

We will evaluate our system using these methods, hopping to achieve results similar to those of similar systems. Also as a test, we are going to use the system to create automatic subject indexes of specialized books and compare them with the original ones. Perfecting these procedures requires the adoption of experimental processes, with numerous tests carried on large-scale corpora.

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 a system for such a task as automatic term acquisition. The ATA system will probably 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.

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. Alongside term detection, we find the task of automatic document indexing. This is a field where natural language techniques are applied to find word chunks that index a given document and that are terms.

Although some of the processes are already implemented, at this time the system is not yet completely functional. That is, the final process, the one responsible for term extraction, has not been implemented yet. Currently, processing stops after syntactic analysis. For now, the architecture and the main algorithm are defined so the implementation process can begin as soon as possible.

Changing language, currently, means having to restart the whole system. This entails considering the new language's dictionary, analysing the morphological behaviour of that language to write new rules, analysing a general corpus and computing word frequency and getting a surface grammar. That work can be done for simple terms in order to prove that the system is language-independent.

References

[Bourigault 1992] Bourigault, D. (1992) "Surface grammatical analysis for the extraction of terminological noun phrases", in *Proceedings of the 15th International Conference on Computational Linguistics, COLING'92*, p. 977-981, Nantes

[Clarkson & Rosenfeld 1997] P.R. Clarkson and R. Rosenfeld (1997); "Statistical Language Modeling Using the CMU-Cambridge Toolkit"; *Proceedings ESCA Eurospeech*

[Faiza 1999] Abbaci Faiza, Développement du Module Post-SMorph. Mémoire de DEA de linguistique et informatique, GRIL, Université Blaise Pascal, Clermont-Ferrand, 1999

[Jacquemin & Bourigault 2000] Jacquemin, C., et Bourigault, D. (2000), "Term Extraction and Automatic Indexing", R. Mitkov, editor, *Handbook of Computational Linguistics*, Oxford University Press, Oxford

[Justeson 1995] Justeson, J. S., Katz S. M. (1995) "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, p. 9-27.

[Neto 1996] Marquez Neto, A. P. (1996) "Terminologia e Corpus Linguístico", *Revista Internacional de Língua Portuguesa - RILP nº 15*, p. 100-108

[Silva 1999] Ferreira da Silva, J., Pereira Lopes, G. (1999) "A local maxima method and a fair dispersion normalization for extracting multi-words units from corpora", *International Conference on Mathematics of Language, Orlando, July 99*