Easy Automatic Terms Acquisition with ATA and Galinha

Joana L. Paulo, David M. de Matos, Nuno J. Mamede

L2F – Spoken Language System Laboratory
INESC-ID Lisboa / IST, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
{joana.paulo, david.matos, nuno.mamede}@l2f.inesc-id.pt

Abstract

ATA (Paulo, 2002) is a system for Automatic Term Acquisition that takes a text from a specific field and analyses it in order to decide which of the detected nouns and noun phrases ought to be considered terminological units. ATA uses a well known architecture (Daille, 1996), taking advantage of the system's modularity which lets us modify each module independently, thus improving the whole system. Currently, ATA is being evaluated over a Portuguese nautical corpus: in the final version of the article, evaluation results will be discussed. Galinha (Galaxy Interface Handler) (Matos, 2002) is a system that integrates multiple linguistic resources and tools. Galinha enables easy module integration and testing of prototypical configurations, thereby reducing the effort and backtracking usual in the construction of modular applications. Joining ATA and Galinha allowed us to provide a web graphical interface to make it easier to automatically acquire terms while accessing to the intermediate results of each module.

1. ATA

ATA is divided into three main modules (see figure 1): linguistic enrichment and selection of those units that may be terms due to their syntactical categories; enrichment of candidates with corpora-based statistical information; and decision about whether they are terms and should be proposed to the user.

In the linguistic analysis sequence, `SMorph` (Aït-Mokhtar, 1998) lemmatizes and annotates morphologically the text using a dictionary. Then, `PAsMo` (Paulo, 2001) rewrites the text according to recomposition and correspondence rules. `PAsMo` also groups the words in phrases. The syntactic analyser `SuSAna` (Batista, 2002) groups phrase constituents. A filtering tool, `GeTerms`, selects those structures that, given their syntactical features, can be terms. That is, for Portuguese, all noun phrases founded on the text.

After that, in the statistical sequence, `Anota` enriches the selected expressions with their statistical information.

Finally, the `Decision` module evaluates the candidate lists, producing the final results to be presented to the user. This is done, by comparing the occurrence of the candidate term in the specialized text and its occurrence on a newspaper corpora analyzed by the same chain process.

The output is a list of words that can be terms. This list may be divided into two sets, both of which may be empty: the first set contains simple term candidates, identified in the text; the second set contains compound term candidates.

Even though the two types of terms to be detected (simple and compound) have different characteristics, we handle them in the same way, by delegating on the grammar the responsibility for customized processing. In an hybrid system such as this, high-frequency terms will be detected statistically, while low-frequency terms will be detected through the grammar of terms. Afterwards, it will be necessary to review the candidate terms. This step is always necessary since not even human annotators eventually find an agreement about the terms in a text.

For evaluation purposes, we analyzed a 114 thousand-word corpus and asked the system for its terms. Then we compare the given list with a list of terms manually detected by linguists. For now we are still running tests and we are trying to experimentally find the best parameters according to which we will say that a noun phrase or noun is a term: the minimal number of occurrences that a term should have and the multiplicative factor when comparing the occurrence on corpora to the occurrence on the specialized text.

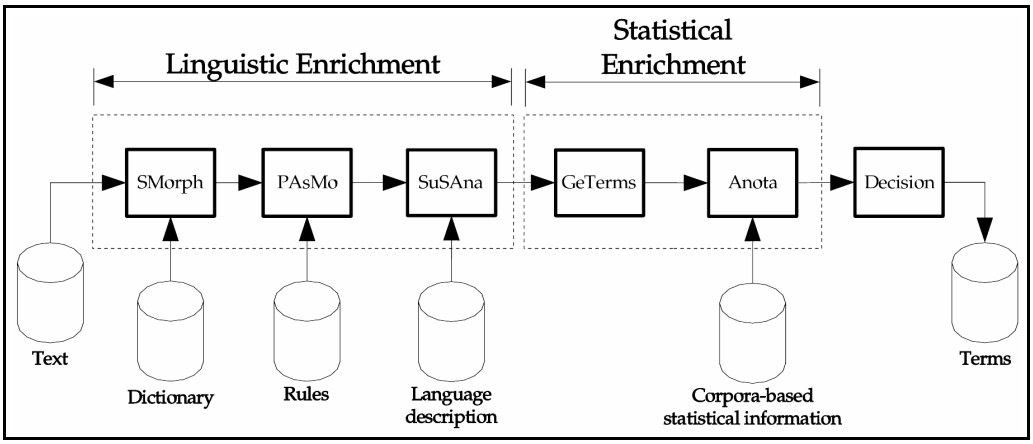

Figure 1: ATA's architecure

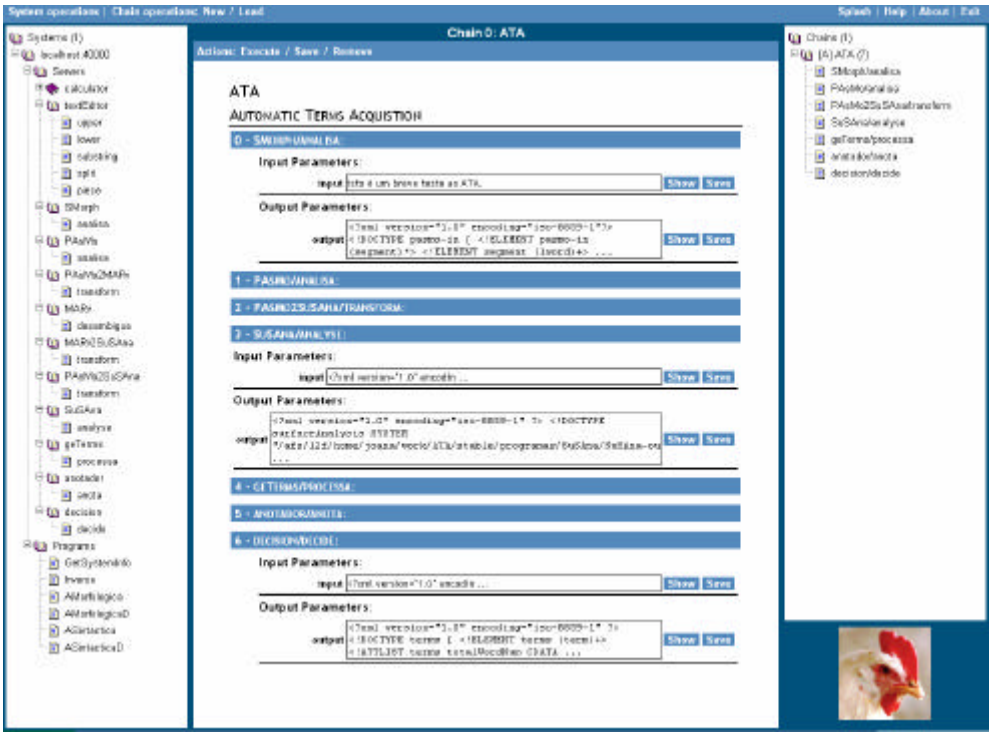

Figure 2: Galinha as a graphical interface for ATA.

2. Galinha

Galinha is a web-based user interface for building modular applications that enables users to access and compose modules using a web browser. Since Galinha works with chains, to include ATA we had to write the corresponding chain and to connect the modules. The main linguistic analysis modules used by ATA were already available through Galinha. Since they accept/produce different data formats, two additional modules were needed to provide data format conversion.

All that was needed to connect the two new modules was to include an existing XSLT (W3C) processor into Galinha. For that, we had only to write a wrapper to call external applications. The wrapper was so simple we were able to generalize it to use any future application we may need.

In figure 2, we show Galinha with ATA's definition: on the left, we have available systems; on the right, we have a chain where all the relevant services are connected and can be executed; in the middle, we can give the text that we want to analyse and - after the results are produced - browse each module's input and output. Since the final result depends on intermediate results, their availability makes evaluation easier.

3. Conclusions

We wanted to automatically extract terms and to create some graphical interface to the system. After designing ATA and its modules, we used Galinha to integrate the modules and provide a graphical interface. As Galinha is easy to use, and adding new modules is also easy, we presented a graphical interface to our automatic terms acquisition system that gives us access to intermediate results and, besides that, can be made available to anyone on the web.

4. References

Aït-Mokhtar, S. (1998). *L'analyse Présyntaxique en une seule étap*. Ph. D. thesis, Université Blaise Pascal, GRIL, Clermont-Ferrand, France.

Batista, F. and Mamede, N. (2002). SuSAna: Módulo Multifuncional de Análise Sintáctica de Superfície. In J. Gonzalo and A. Peñas and A. Ferrández (eds.) *Proc. Multilingual Information Access and Natural Language Processing Workshop*, IBERAMIA 2002, Sevilla, Spain, 29-37.

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

Matos, D. M. de et al. (2002). Empowering the User: a Data Oriented Application-building Framework. In *Adj. Proc. of the 7th ERCIM Workshop "User Interfaces for All"*. Chantilly, France. European Research Consortium for Informatics and Mathematics, 37-44.

Paulo, J. L. (2001). PAsMo – Pós-Análise Morfológica. *Technical Report*. Lisboa, Portugal.

Paulo, J. L. et al. (2002). Using Morphological, Syntactical, and Statistical Information for Automatic Term Acquisition. In E. Ranchhod and N. Mamede (eds.), *Advances in Natural Language Processing, Third International Conference, Portugal for Natural Language Processing (PorTAL)*. Faro, Portugal. Springer-Verlag, LNAI 2389: 219-227.
See: `www.w3.org/Style/XSL`.

World Wide Web Consortium (W3C). The Extensible Stylesheet Language (XSL).