Reusing Linguistic Resources: a Case Study in Morphossyntactic Tagging

Ricardo Ribeiro†, Nuno J. Mamede*, Isabel Trancoso*

†INESC-ID Lisboa/ISCTE
*INESC-ID Lisboa/IST
Spoken Language Systems Lab
R. Alves Redol, 1000-029 LISBON, Portugal
{Ricardo.Ribeiro, Nuno.Mamede, Isabel.Trancoso}@inesc-id.pt

Abstract

This paper describes several issues concerning the reusability of linguistic resources, with special emphasis on morphossyntactic tagging. Ribeiro (2003) presents a morphossyntactic tagging system with a modular architecture. What are the consequences of changing a module of this system? How difficult would be to integrate the morphossyntactic tagger in other systems? These are some of the questions that are addressed by this paper, where possible approaches to the problems that may appear are also discussed.

1. Introduction

One of the major problems related to natural language processing is the availability of manually annotated resources. In fact, this question can be posed concerning all kinds of resources: corpora, lexica and tools. Yet, nowadays, the relevance of this problem, even for the Portuguese language, seems to be diminishing, but a new one arising: the usability of the existing resources (Matos et al.,2003; Jing and McKeown, 1998; Olsson et al., 1998).

In (Ribeiro, 2003; Ribeiro et al., 2003) is presented a morphossyntactic tagger that followed a modular approach. The strategy adopted by this system, motivated by the fact that neolatin languages, such as Portuguese, are highly inflectional when compared with English, consists of two sequential steps: morphological analysis and ambiguity resolution. Given such architecture, one would expect that replacing one of the modules would not be a difficult task.

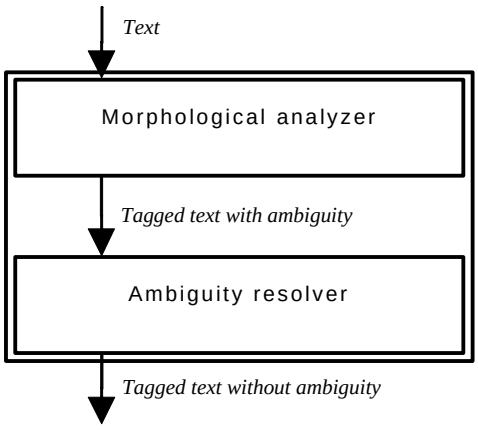

Figure 1. Morphossyntactic tagger architecture.

2. Reusability problem

The reusability problem appeared when we tried to use MARv (Ribeiro, 2003), the morphossyntactic disambiguation module, in the automatic term acquisition (ATA) system presented in (Paulo, 2003). In the ATA system, the morphological analysis is performed by SMorph (Aït-Mokhtar, 1998) and followed by the post morphological analysis tool PAsMo (Paulo, 2001), whilst the morphological analysis module of the morphossyntactic tagging system is Palavroso (Medeiros, 1995). Since there are some conceptual differences between these two systems some adaptations were needed. Two major problems were identified:

- the tokenization performed by the two systems was different;
- the tagsets, besides being different, were ruled by divergent principles.

MARv's architecture comprehends two submodules: a linguistic-oriented disambiguation rules module and a probabilistic disambiguation module. Considering the differences between the two morphological analyzers, substituting Palavroso by SMorph/PAsMo demanded some changes in both modules. Concerning the disambiguation rules module, the focus was on rule adaptation. Concerning the probabilistic disambiguation module, the modifications consisted in the development of new probabilistic models.

3. Used corpus

The corpus used to develop these models was built in the LE-PAROLE European project (Bacelar et al., 1997) in which harmonized reference corpora and generalist lexica were built according to a common model for the 12 European languages involved. This corpus was morphossyntactically tagged using Palavroso and manually disambiguated. The tagset had about 200 tags with information that varied from grammatical category to morphological features that could be combined to form composed tags (resulting in about 400 different tags). This corpus was developed to be part of the core of a set of written language resources for the European Community countries. In other words, its main purpose is to be reused.

4. Adopted approach

In order to develop new models for the probabilistic module of MARv, the LE-PAROLE corpus was used.

But since this corpus was tagged with Palavroso, the tokenization and the tagset problems previously identified arouse in the corpus reuse.

The approach to these problems was a semi-automatic solution that comprehends four steps:

- Tagging of the corpus using SMorph/PAsMo;
- Identification of the situations where occur contraction or expansion of tokens identified by Palavroso. For example, SMorph/PAsMo gives "é sintetizada" or "cidade - campo" as tokens, where Palavroso gives "é", "sintetizada" and "cidade", "-", "campo" as tokens;
- Identification of a mapping between the tagsets;
- Development of an interface based on a rule set obtained from the previously identified situations. Whenever it was not possible to apply a rule the automatic process was interrupted and the user was queried about how to solve that particular situation.

Although effective, this approach was very slow, since the rule set did not cover several situations and it was not possible to define a function from the Palavroso tagset to the SMorph/PAsMo tagset.

5. References

Aït-Mokhtar, S., (1998). *L'analyse présyntaxique en une seule étape*. PhD Thesis, Université Blaise Pascal, Clermont-Ferrand, GRIL.

Bacelar, F., J. Bettencourt, P. Marrafa, R. Ribeiro, R. Veloso and L. Wittmann (1997). LE-PAROLE – Do corpus à modelização da informação lexical num sistema multifunção. In *Actas do XIII Encontro da APL*. Portugal.

Jing, H. and K. McKeown (1998). Combining multiple, large-scale resources in a reusable lexicon for natural language generation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (pp. 607–613).

Matos, D., J. L. Paulo and N. Mamede (2003). Managing Linguistic Resources and Tools. In Mamede, N., J. Baptista, I. Trancoso and M. das Graças Volpe Nunes, editors, *Proceedings of the 6th International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003)*, volume 2721 of *Lecture Notes in Artificial Intelligence* (pp. 135–142). Springer.

Medeiros, J. C. (1995). *Processamento Morfológico e Correcção Ortográfica do Português*. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Olsson, F., B. Gambäck and M. Eriksson (1998). Reusing Swedish Language Processing Resources in SVENSK. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, volume *Workshop on Minimizing the Effort for Language Resource Acquisition* (pp. 27–33). ELRA.

Paulo, J. L. (2001). *PAsMo – Pós-Análise Morfológica*. Technical report, L²F – INESC-ID Lisboa, Portugal.

Paulo, J. L. (2003). *Aquisição Automática de Termos*. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal. (to appear).

Ribeiro, R. (2003). *Anotação Morfossintáctica Desambiguada do Português*. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal.

Ribeiro, R., L. Oliveira and I. Trancoso (2003). Using Morphossyntactic Information in {TTS} Systems: Comparing Strategies for European Portuguese. In Mamede, N., J. Baptista, I. Trancoso and M. das Graças Volpe Nunes, editors, *Proceedings of the 6th International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003)*, volume 2721 of *Lecture Notes in Artificial Intelligence* (pp. 143–150). Springer.