Morphossyntactic Tagging as a Case Study of Linguistic Resources Reuse

Ricardo Ribeiro[1], Nuno J. Mamede[2], and Isabel Trancoso[2]

[1] INESC-ID Lisboa/ISCTE
[2] INESC-ID Lisboa/IST
Spoken Language System Lab
R. Alves Redol, 9, 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. We have implemented a morphossyntactic tagging system with a modular architecture. What are the consequences of changing a module of this system? How difficult would it 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 [1–3].

In [4, 5] is presented a morphossyntactic tagger that follows a modular approach. The strategy adopted by this system consists of two sequential steps: morphological analysis and ambiguity resolution. Given such an architecture, one would expect that replacing one of the modules would not be a difficult task. But as it is described bellow, such modifications may trigger a chain of problems that has as source the difficulties in the reuse of existing resources.

This document is organized as follows: Sect. 2 introduces the reusability problems; Sections 3 to 5 describe the involved resources (tools and corpora); Finally, before the concluding remarks, Sect. 6 discusses some steps taken to overcome the problems and points some work directions that define the chosen approach.

2 Reusability Problems

There are several reasons for it being difficult to reuse existing natural language resources. For instance, at some stage in the development of a morphossyntactic tagger, a decision about the granularity of the information, or the principles that command the used tagset, has to be made. These strategies or principles can compromise the usability of the tool in a context different from the one of its development [4]. Even in the same context, when integrating several tools to develop a complex application, connecting the interfaces of all used modules can become a more complex task than expected [1]. The same problem occurs in the development of a lexicon: if the lexicon is built to be used by language interpretation applications, generally it is not suitable to be used directly for language generation. A generation lexicon is usually indexed by semantic concepts whereas an interpretation lexicon is indexed by words [2].

Concerning this case study, the reusability problem appeared when trying to use a morphossyntactic disambiguation module – MARv[3] – in an automatic term acquisition system – ATA (see Sect. 5). In the ATA system, morphological analysis is performed by SMorph [6] and followed by the post morphological analysis tool PAsMo [7], whilst the morphological analysis module of the morphossyntactic tagging system to which MARv belongs is Palavroso (see Sect. 3). 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.

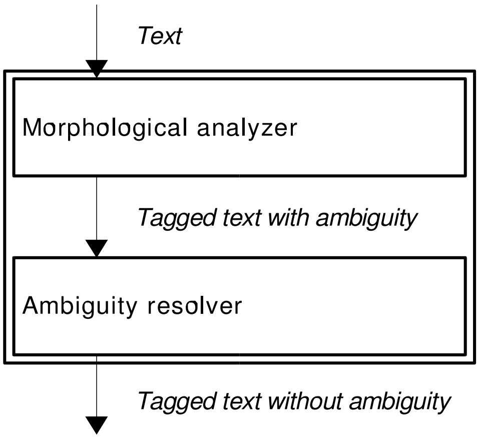

Fig. 1. Morphossyntactic tagger architecture.

[3] MARv is the disambiguation module of the morphossyntactic tagger mentioned in the Introduction and is described in Sect. 3.

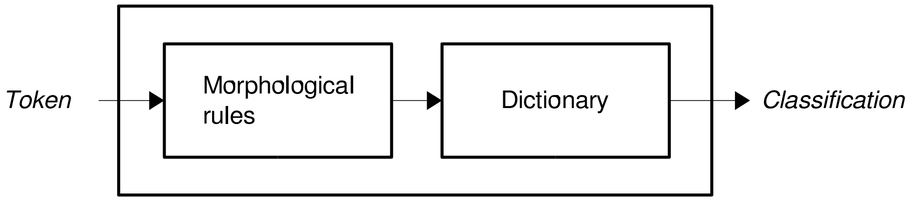

Fig. 2. Palavroso architecture.

3 Morphossyntactic Tagging System

The morphossyntactic tagging process we have implemented [4] consists of the two sequential steps shown in figure 1. This approach was motivated by the fact that neolatin languages, such as Portuguese, are highly inflectional when compared with English. In this sense, morphological analysis can be relevant. In fact, on the one hand, linguistic oriented systems are usually based on the elimination of the ambiguity previously introduced by a lexical analysis process, and, on the other hand, in data-driven approaches, information is derived from corpora and, due to data sparseness, word forms may not appear with all possible tags or may even not occur at all [8, 9].

The morphological analysis module adopted in [4] is Palavroso. Presented in [10], this broad coverage morphological analyzer was developed to address specific problems of the Portuguese language, such as compound nouns, enclitic pronouns or adjective degrees. As output it gives all possible part-of-speech tags for a given word. If a word is not known, Palavroso has the possibility to guess possible part-of-speech tags, always giving an answer.

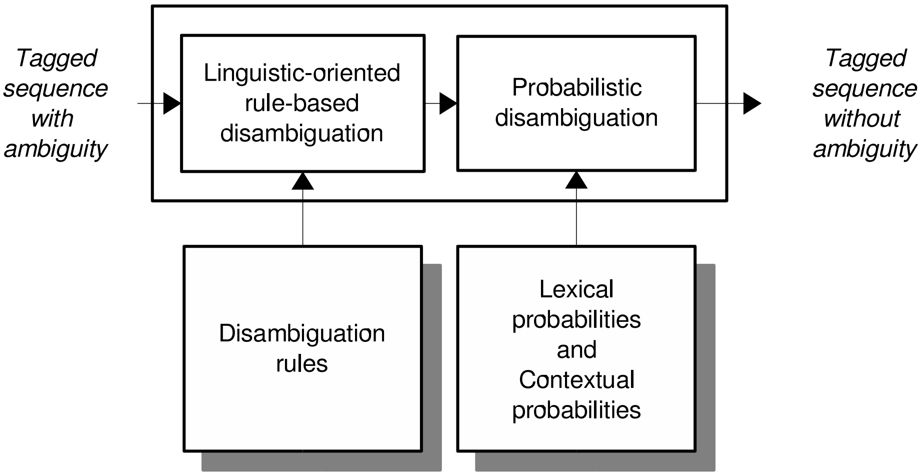
Tagged sequence with ambiguity
Linguistic-oriented rule-based disambiguation
Probabilistic disambiguation
Tagged sequence without ambiguity
Disambiguation rules
Lexical probabilities and Contextual probabilities

Fig. 3. MARv architecture.

The architecture of the disambiguation module – MARv – comprehends two submodules: a linguistic-oriented disambiguation rules module and a probabilistic disambiguation module. The ambiguity is first reduced by the disambiguation rules module and then the probabilistic module produces a fully disambiguated output. The disambiguation rules module is based on a set of contextual rules developed specifically for Portuguese. The rules have the following structure: an input trigger section (where a word or an ambiguity class can be specified); an *if*-condition (where the applicability context referring to surrounding ambiguities, words or tags, is specified); and an action section (select or remove tags). Figure 4 shows an example of a rule. The probabilistic-based disambiguation module is based on Markov models and uses the Viterbi algorithm to find the most likely sequence of tags for the given sequence of words. The forward algorithm is used to compute the lexical probabilities. The contextual probabilities are given by smoothed trigrams.

```
Input: AMB = "A= Nc V="
If
     (-1/TAG = "S=")
then
     SELECT "Nc"
```

A=	Adjective
Nc	Common noun
V=	Verb
S=	Preposition

Fig. 4. Disambiguation rule.

4 The Corpus

The corpus used to develop the models needed by the probabilistic-based disambiguation module was built in the LE-PAROLE European project [11].

The objective of LE-PAROLE was to built harmonized reference corpora and generalist lexica according to a common model for the 12 European languages involved. These resources were developed to be part of the core of a set of written language resources for the European Community countries. In other words, their main purpose is to be reused.

The used corpus is a subset of about 290,000 running words of the collected 20 million running words corpus for European Portuguese. This subset was morphossyntactically tagged using Palavroso and manually disambiguated. The tagset had about 200 tags with information that varied from grammatical category to morphological features. The tags could be combined to form composed tags, increasing in this way the size of the tagset to about 400 different tags. This tagset was harmonized between all the languages involved.

This corpus was subdivided into training and test subsets. The training corpus has about 230,000 running words and it covers about 25,000 different word forms. The test corpus has about 60,000 running words, covering 10,000 different word forms.

5 ATA

ATA is an automatic term acquisition system: given an input text from a specific field, it produces a list of possible terminological units found in the text for that field [12,13].

The architecture of ATA comprehends three main components: a linguistic enrichment component, a statistical enrichment component and a decision component.

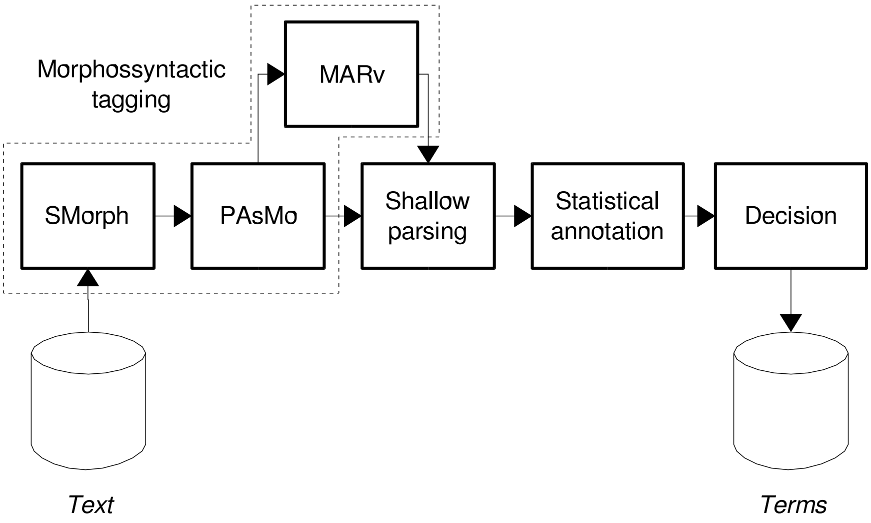

Fig. 5. ATA's architecture.

The linguistic component is responsible for the identification of nouns and noun phrases. The text is initially submitted to a morphological analyzer (SMorph) which produces a tokenized, lemmatized and tagged (with ambiguity) text. SMorph's output is passed to PAsMo which is responsible for the regrouping of tokens and tag conversions. The resulting text is submitted to the shallow parser in order to identify all possible noun phrases. Either to produce a smaller (and more reliable) list of lexical units that can be considered terms or just to improve the shallow parsing step, a morphossyntactic disambiguation module (MARv) could be introduced between the morphological analyzer and the shallow parser.

After the linguistic enrichment component, the statistical component then adds corpora-based frequency information to the identified lexical units (nouns and noun phrases).

The last component of ATA is in charge of the decision process: it produces a list of possible terminological units, given the statistical information previously added.

6 Adopted Approach

As mentioned before, MARv's architecture includes a linguistic-oriented rule-based disambiguation module and a probabilistic disambiguation module. Considering the tokenization and tagset differences between the two morphological analyzers, replacing Palavroso with SMorph/PAsMo in the morphossyntactic tagging stage of the ATA system (figure 5), demanded some changes in both of MARv's 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. Rule adaptation was a manual work. And since MARv's rules were originally crafted to be used with SMorph, their adaptation worked as expected. On the other hand, in order to develop new models for the probabilistic module of MARv, the LE-PAROLE corpus was reused. But since this corpus was tagged with Palavroso, the tokenization and the tagset problems previously identified appear in the corpus reuse.

The approach to these problems was a semi-automatic solution to develop a new corpus from the original, tokenized and tagged accordingly to the SMorph/PAsMo strategy. The process involved four steps:

- Tagging of the corpus using SMorph/PAsMo;
- Identification of the situations where contraction or expansion of tokens identified by Palavroso occur. These situations result in misalignments between the original corpus and the new version of the corpus tagged with SMorph/PAsMo. Some examples of these tokenization differences are shown in table 1.
- 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. Another negative issue of this

Table 1. Tokenization differences.

Text	SMorph/PAsMo	Palavroso
... é sintetizada ...	“é sintetizada”	“é” + “sintetizada”
... cidade - campo ...	“cidade - campo”	“cidade” + “-” + “campo”
... do ...	“de” + “o”	“do”

idea is that the resulting corpus would still be difficult to reuse in a different processing environment, since it would suffer from the same problems of the original one.

A new approach to the problem was delineated and is now in progress. Instead of building a new corpus following a specific strategy for the morphossyntactic level, we decided to develop a new corpus that would allow to acommodate different strategies. Two main items were taken in account:

1. Changing the tokenization (maximizing the number of tokens);
2. Increasing the granularity of the morphossyntactic descriptions.

Concerning the change in the tokenization, maximizing the number of tokens will allow to acommodate different tokenization strategies. In fact, any other tokenization strategy specifies which tokens are regrouped. On the other hand, the increase of the granularity of the morphossyntactic descriptions is done by adding the information from SMorph's dictionaries to the information already present in the corpus. Tags are treated as matrices, so is only a matter of adding a new field to receive the new information. This process can be used whenever new information is available.

```
<w msd="S=ffs">da</w>
<w msd="Ncfs">poesia</w>
<w msd="A=pfs">francesa</w>
```

Fig. 6. Extract from the original corpus.

Figure 6 shows an extract of the original corpus and figure 7 shows an extract of the new corpus. It is possible to observe both the changes in the tokenization and in the tagset.

7 Final Remarks

This paper analyzes several issues concerning the reuse of natural language resources. This study allowed us to understand what kind of problems may appear

```
<w msd="S=s">de</w> <w msd="Tdfs">a</w>
<w msd="Ncfs">poesia</w>
<w msd="A3pfs">francesa</w>
```

Fig. 7. Extract from the new corpus.

when trying to reuse existing resources, in particular at the morphossyntactic level. A discussion about how to overcome the problems is presented and a solution is introduced, underlining how this strategy improves the reusability of these resources.

References

1. Matos, D., Paulo, J.L., Mamede, N.: Managing linguistic resources and tools. In Mamede, N., Baptista, J., Trancoso, I., das Graças Volpe Nunes, M., eds.: Proceedings of the 6^{th} International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003). Volume 2721 of Lecture Notes in Artificial Intelligence., Springer (2003) 135–142
2. Jing, H., McKeown, K.: Combining multiple, large-scale resources in a reusable lexicon for natural language generation. In: Proceedings of the 36^{th} Annual Meeting of the Association for Computational Linguistics and the 17^{th} International Conference on Computational Linguistics. (1998) 607–613
3. Olsson, F., Gambäck, B., Eriksson, M.: 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., ELRA (1998) 27–33
4. Ribeiro, R.: Anotação morfossintáctica desambiguada do português. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal (2003)
5. Ribeiro, R., Oliveira, L., Trancoso, I.: Using morphossyntactic information in TTS systems: Comparing strategies for european portuguese. In Mamede, N., Baptista, J., Trancoso, I., das Graças Volpe Nunes, M., eds.: Proceedings of the 6^{th} International Workshop on Computational Processing of the Portuguese Language (PROPOR 2003). Volume 2721 of Lecture Notes in Artificial Intelligence., Springer (2003) 143–150
6. Aït-Mokhtar, S.: L'analyse présyntaxique en une seule étape. PhD thesis, Université Blaise Pascal, Clermont-Ferrand, GRIL (1998)
7. Paulo, J.L.: PAsMo – Pós-Análise Morfológica. Technical report, L^2F – INESC-ID, Lisboa (2001)
8. Laporte, .: Resolução de ambiguidades. In: Tratamento das Línguas por Computador. Caminho (2001)
9. Oravecz, C., Dienes, P.: Efficient stochastic part-of-speech tagging for hungarian. In: Proc. of the Third LREC, Las Palmas, Espanha, ELRA (2002) 710–717
10. Medeiros, J.C.: Processamento morfológico e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, Portugal (1995)

11. Bacelar, F., Bettencourt, J., Marrafa, P., Ribeiro, R., Veloso, R., Wittmann, L.: LE-PAROLE – Do corpus à modelização da informação lexical num sistema multifunção. In: Actas do XIII Encontro da APL, Portugal (1997)
12. Paulo, J.L.: Aquisição automática de termos. Master's thesis, Instituto Superior Técnico – Universidade Técnica de Lisboa, Portugal (2003) (to appear).
13. Paulo, J., Matos, D., Mamede, N.: Easy automatic terms acquisition with ata and galinha. In Branco, A., Mendes, A., Ribeiro, R., eds.: Tagging and Shallow Processing of Portuguese: workshop notes of TASHA'2003. Volume TR–03–28 of Technical Reports., Lisboa, Portugal, Faculdade de Ciências da Universidade de Lisboa (2003) 29–30