XML rules for enclitic segmentation ⋆

Fco. Mario Barcala, Miguel A. Molinero, and Eva Domínguez

Centro Ramón Piñeiro, Ctra. Santiago-Noia km. 3, A Barcia,
15896 Santiago de Compostela, Spain
{fbarcala,molinero,edomin}@cirp.es

Extended Abstract

Sentence word segmentation is an important task in robust part-of-speech (POS) tagging systems. In some cases this is relatively simple, since each textual word (or token) corresponds to one linguistic component. However, there are many others where segmentation can be very hard, such as those of contractions, verbal forms with enclitic pronouns, etc., where the same token contains information about two or more linguistic components. There are two main approaches to solving these difficult cases:

1. To treat tokens as a whole, and to extend the tagset to represent these phenomena. For instance, the Galician word `colléullelo` (*he or she caught it from them*) could be tagged as `Vei3s0+Rad3ap+Raa3ms`, that is, verb (`V`), past simple (`ei`), 3rd person (`3`) singular (`s`) with an atonic pronoun (`Ra`), dative (`d`), 3rd person (`3`) masculine or feminine (`a`), plural (`p`) and another atonic pronoun (`Ra`), accusative (`a`), 3rd person (`3`), masculine (`m`), singular (`s`).
2. To segment compound tokens, separating the components. For example, the same word can be broken into three parts: `colléu`, `lle` and `lo`. In this case, the components could be tagged separately as `Vei3s0`, `Rad3ap` and `Raa3ms` respectively.

Although EAGLES [1] points towards using the second approach when these phenomena occur frequently, most papers and systems are based on the first approach [2] [3]. It is the simplest one, because there is no need to change current taggers [4], and it performs well with languages that hardly present these cases, for example English or French, but it presents several problems with others which have many occurrences of several linguistic components within the same word.

Such is the case of Galician, and to a lesser extent, other romance languages such as Spanish. In these cases, using the first approach, the tagset size would be greatly increased, and sometimes the POS tagging tasks would be impractical due to the need for an extremely large training corpus in order to obtain an

⋆ This paper results from a project developed at Centro Ramón Piñeiro para a Investigación en Humanidades and is partially supported by Ministerio de Educación y Ciencia (TIN2004-07246-C03-01), Xunta de Galicia (PGIDIT05PXIC30501PN) and Universidade de Vigo.

acceptable tagger behaviour [5]. Therefore, works which choose the first option do not explain how to solve other POS related problems: how to assign lemmas to compound tokens, how to use the system inside a more complex one (for instance, a translation machine or a parser), etc. which in other works concerning the second alternative are trivial processes.

In this paper we follow guidelines indicated by EAGLES, and we explain the internals of a highly configurable system which segments compound tokens into its components. It must be used in combination with a tagger which solves segmentation ambiguities [6] [7] so as to perform jointly all POS tagging related tasks.

We will center our attention on the hardest item in segmentation, that is, the processing of verbal forms with enclitic pronouns. Contrary to other ad-hoc systems, to do this we use easily configurable high level XML [8] rules outside the system code, which will allow linguists to adapt the system to their particular needs and languages. We have applied it to the Galician language [9], but the system is generic enough to be applied to most romance languages and any tagset.

References

1. Expert Advisory Group on Language Engineering Standards (EAGLES). Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages. In *EAGLES Document EAG-CLWG-MORPHSYN/R*, May, 1996.
2. José L. Aguirre Moreno, Alberto Álvarez Lugrís, Xavier Gómez Guinovart. Aplicación do etiquetario morfosintáctico do SLI ó corpus de traducción TECTRA. In *Viceversa*, pages 207-231, 2002.
3. Carreras, X., I. Chao, L. Padró, M. Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC'04), Lisbon, Portugal, 2004.
4. Brants, T. A statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'2000)*, Seatle, 2000.
5. Jorge Graña. Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural. In *Doctoral thesis, Universidad de La Coruña*, Spain, 2000.
6. Jorge Graña, Miguel A. Alonso, Manuel Vilares. A Common Solution for Tokenization and Part-of-Speech Tagging: One-Pass Viterbi Algorithm vs. Iterative Approaches. In Petr Sojka, Ivan Kopecek and Karel Pala (eds.), *Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Artificial Intelligence*, pp. 3-10, Springer-Verlag, Berlin-Heidelberg-New York, 2002.
7. Jorge Graña, Fco. Mario Barcala, Jesús Vilares. Formal Methods of Tokenization for Part-of-Speech Tagging. In *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pp. 240-249, Springer-Verlag, Berlin-Heidelberg-New York, 2002.
8. World Wide Web Consortium. http://www.w3c.org, [25/10/2006].
9. Rosario Álvarez, Xosé Xove. Gramática da Lingua Galega. In *Editorial Galaxia*, Vigo, Spain, 2002.