LX-Chunker

1. BASIC INFORMATION

1.1 Tool name

LX-Chunker.

1.2 Overview and purpose of the tool

The present tool, that was built to deal with specific issues concerning orthographic conventions adopted for Portuguese, marks sentence boundaries with `<s>`…`</s>`, and paragraph boundaries with `<p>`…`</p>`. Unwraps sentences split over different lines.
A f-score of 99.94% was obtained when testing on a 12,000 sentence corpus accurately hand tagged with respect to sentence and paragraph boundaries.

LX-Chunker was developed and is maintained at University of Lisbon by the NLX-Natural Language and Speech Group of the Department of Informatics.

1.3 A short description of the algorithm

The rule-based algorithm is described in Silva, 2007.

2. TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

Linux.

2.2 Installation

Not applicable.

2.3 Execution instructions

The sentence chunker works as a command line filter tool (it reads input from stdin and writes to stdout). Accordingly, it is meant to be used as part of pipe constructs in the (UNIX/Linux) command line.

Example:

```
$ cat input.txt | /path/to/chunker/run-Chunker.sh > output.txt
```

Note that:

1. The tool needs the following files to be executable in order to run:

```
./chunker-one
```

```
./run-Chunker.sh
```

2. This tool use pre-compiled C code. Make sure you have the ia32-libs package installed for 32bit support, this happened to us too when we made the switch to 64bit.

2.4 Input/Output data formats

Input is in raw text (see 3.1).

Output in also in raw text with <s> (for sentences) and <p> (for paragraphs) labels (see 3.2).

2.5 Integration with external tools

Not applicable.

3. CONTENT INFORMATION

3.1 A test input file

```
Esta frase serve para testar o funcionamento da suite.
Esta outra frase faz o mesmo.
```

3.2 The output file

```
<p><s> Esta frase serve para testar o funcionamento da suite.
</s>
<s> Esta outra frase faz o mesmo. </s></p>
```

3.3 Approximation of the time necessary to process the test input file.

Not applicable.

4. ADMINISTRATIVE INFORMATION

4.1 Contact person

Name: António Branco
Address: Departamento de Informática NLX - Grupo de Fala e Linguagem Natural, Faculdade de Ciências da Universidade de Lisboa, Edifício C6,
Campo Grande 1749-016 Lisboa
Position: Assistant professor
Affiliation: Faculty of Sciences, University of Lisbon
Telephone: +351 217 500 087
Fax: +351 217 500 084

E-mail: antonio.branco@di.fc.ul.pt

5. LICENSE

This tool is a free license-based for research purposes and a free license-based for commercial purposes, with attribution and no redistribution nor derivatives allowed. It will be available on the META-SHARE platform.

6. RELEVANT REFERENCES AND OTHER INFORMATION

Branco, António and João Silva (2004). "Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese." In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, Paris, ELRA, ISBN 2-9517408-1-6, pp. 507-510.

Silva, João, 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. MSc thesis, University of Lisbon. Published as Technical Report DI-FCUL-TR-07-16.