LX-Tokenizer

1. BASIC INFORMATION

1.1 Tool name

LX-Tokenizer.

1.2 Overview and purpose of the tool

The present tool, that was built to deal with Portuguese-specific issues concerning a few non-trivial cases that involve tokenization-ambigous strings, segments text into lexically relevant tokens, using whitespace as the separator. Note that, in these examples, the | (vertical bar) symbol is used to mark the token boundaries clearier:

```
um exemplo → |um|exemplo|
```

Expands contractions. Note that the first element of an expanded contraction is marked with an _ (underscore) symbol:

```
do → |de_|o|
```

Marks spacing around punctuation or symbols. The * and the */ symbols indicate a space to the left and a space to the right, respectively:

```
um, dois e três → |um|,*/|dois|e|três|
5.3 → |5|.|3|
1. 2 → |1|.*/|2|
8 . 6 → |8|\*.*/|6|
```

Detaches clitic pronouns from the verb. The detached pronoun is marked with a `-` (hyphen) symbol. When in mesoclisis, a `-CL-` mark is used to signal the original position of the detached clitic. Additionally, possible vocalic alterations of the verb form are marked with a # (hash) symbol:

```
dá-se-lho → |dá|-se|-lhe|-o|
afirmar-se-ia → |afirmar-CL-ia|-se|
vê-las → |vê#|-las|
```

This tool also handles ambiguous strings. These are words that, depending on their particular occurrence, can be tokenized in different ways. For instance:

```
deste → |deste| when occurring as a Verb
deste → |de|este| when occurring as a contraction (Preposition +
Demonstrative)
```

This tool achieves a f-score of 99.72%.

LX-Tokenizer 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 tokenizer 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/tokenizer/run-Tokenizer.sh >
output.txt
```

The input text must encoded using UTF-8. The output uses the same encoding.

The tokenizer must resort to a POS tagger to properly handle token-ambiguous strings (strings that can be tokenized as one or as two tokens, depending on their category). The POS tagger that is used is MXPOST, by Adwait Ratnaparkhi. A model for the POS tagger is provided, but you must get the POS tagger yourself since we cannot redistribute that software. The MXPOST tagger can be downloaded at the following address: `ftp://ftp.cis.upenn.edu/pub/adwait/jmx/jmx.tar.gz`.

You must set the path to your local installation of MXPOST. For this, edit the Tagger/run-Tagger.sh script and set the MXPOST_JAR variable.

2.4 Input/Output data formats

Input is in raw text (see 3.1).

Output is in raw text (see 3.2) with paragraphs and sentences and ponctuation marks.

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

```
Esta frase serve para testar o funcionamento de_ a suite .*/
Esta outra frase faz o mesmo .*/
```

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 free for research purposes, with attribution and no redistribution nor derivatives allowed. It is 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.