FORMA

1. BASIC INFORMATION

1.1 Tool name

Tagger FORMA 1.1.

1.2 Overview and purpose of the tool

FORMA 1.1 is a probabilistic tool for morphological tagging and lemmatization of text. The purpose of this tool is to obtain annotated text to be processed by other NLP tools (see Gonzalez *et al.*, 2006).

1.3 A short description of the algorithm

FORMA is a probabilistic tool which uses auxiliary data sets that were created from a lemmatized-tagged training corpus. The auxiliary data sets constitute three automata for compounds, accents, and suffixes, and two probabilistic matrices *BP* and *AP*. The automaton for compounds (with 577 compounds) is a left-right automaton that detects multi-word units, i.e., compound prepositions and compound adverbs. The automaton for accents (with 302 accentuated words) is a left-right automaton that detects words where the accent is a distinctive feature (which is important when dealing with Portuguese). The automaton for suffixes (with 43,476 entries) is a right-left automaton which analyses word suffixes for determining lemma and tag probabilities. While the three automata obtain lemmas and tag probabilities, the matrices *BP* and *AP* estimate the probability of occurrence of a tag concerning the text. The matrix *BP* = {*bpmj*} is a 21x21 matrix where each element is bp_{mj} = Pr (*j* | *m*), i.e., the probability of occurrence of the tag j given a prior occurrence of the tag m in the corpus. On the other hand, AP = {*apnj*} is a 21x21 matrix where each element is ap_{nj} = Pr (*j* | *n*), i.e., the probability of occurrence of the tag *j* given a posterior occurrence of the tag *n* in the corpus.

2. TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

Linux.

2.2 Installation

N/A.

2.3 Execution instructions

A) Proceedings for folders and files preparation:
1) Copy the “forma_1_1.bat” file to a folder (here named “pai”)

2) Create a folder below the “pai” archive with the name "forma_1_1"
3) In the folder "forma_1_1", copy the files "acentos.let", "forma_1_1.c", "locucoes.let", "sufixato.let" and "sufixos.let"
4) Compile (linguage C) the programe-source "forma_1_1.c", generating the executable "forma_1_1.exe" in the folder "forma_1_1"
B) Proceedings for the execution (Linux)
In a comand line, in the "pai" folder, execute:
./forma_1_1.bat ARQ1 ARQ2

in which:
ARQ1 = input
ARQ2 = output

For more information, see leia_me.txt (readme) file.

2.4 Input/Output data formats

Input format (ARQ1) is plain text.

Output format (ARQ2) is plain text with one token per line. Each token is followed by its lemma and morphological category. The tag set adopted is: *_AD* and *_AI* (definite and indefinite articles), *_AJ* (adjective), *_AP* (participle), *_AV* (adverb), *_CC* and *_CS* (coordinate and subordinate conjunctions), *_IN* (interjection), *_NC* and *_NO* (cardinal and ordinal numbers), *_PS*, *_PD*, *_PI*, *_PL*, and *_PP* (possessive, demonstrative, indefinite, relative, and personal pronouns), *_PN* (punctuation), *_PR* (preposition), *_SU* (noun), *_VA* (auxiliary verb), *_VB* (verb), and *_VG* (comma, parentheses, dash).

2.5 Integration with external tools

N/A.

3. CONTENT INFORMATION

3.1 A test input file

...programas que realizam tarefas complexas...

3.2 The output file

programas programa _SU
que que _PL
realizam realizar _VB
tarefas tarefa _SU
complexas complexo _AJ

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

N/A.

4. ADMINISTRATIVE INFORMATION

4.1 Contact person

Name: Marco Gonzalez
Address:
PUCRS – Pontificia Universidade Católica do Rio Grande do Sul,
Faculdade de Informática
Avenida Ipiranga, 6681 – Prédio 32 – FACIN
90610-001 Porto Alegre - Brasil
Affiliation: PUCRS – Pontificia Universidade Católica do Rio Grande do Sul,
Faculdade de Informática
Position: Professor
Telephone: +51 33203558
Fax: +51 33203758
e-mail: marco.gonzalez@pucrs.br

5. LICENSE

This tool is free licensed-based for both research and commercial purposes under a GNU LGPL license. It will be available on the META-SHARE platform.

6. RELEVANT REFERENCES AND OTHER INFORMATION

Gonzalez, M.; Lima, V. L. S. de; Lima, J. V de (2006). “Tools for Nominalization: An Alternative for Lexical Normalization.” In *Workshop on Computational Processing of Portuguese – Written and Spoken, 7, PROPOR*, Vol. 3960, Springer-Verlag, pp. 100-109.