Analisador Morfológico

João Tiago Aparício; Student number: 97155; Group 86;
Msc. Information Systems and Computer Engineering - 2019
Instituto Superior Técnico, Universidade de Lisboa

I. Options taken

In this report, ”fst” denotes ”finite-state transducer” and a set of alphanumeric characters concatenated with ”.fst” denotes a file with the ”.fst” extension (meaning a binary file).

The following fst's were written in the respective txt files: lemma2noun.fst, lemma2adverb.fst, lemma2verbip.fst, lemma2verbis.fst and lemma2verbif.fst.

The lemma2verb.fst was generated by union of some of the previously mentioned fst's (lemma2verbip.fst, lemma2verbis.fst and lemma2verbif.fst).

The lemma2word.fst was generated by union of the ”lemma2verb.fst” with lemma2noun.fst, lemma2adverb.fst.

An couple of auxiliary fst's are generated (lemma2verbAUX.fst, lemma2wordAUX.fst). Their purpose is to help the generation of lemma2verb.fst and lemma2word.fst. But they are erased shortly after being created in the run.sh script, in order to avoid confusion (and evidently following the rules set by the professor).

The ”word2lemma.fst” was generated by inverting the ”lemma2word.fst”. This option was taken due to the fact that their function is in fact the inverse of one another.

II. Comments on the solution developed

A set of tests were developed to test the functionality of the transducers. Every test outputs a .fst file and the corresponding visualizations. The empty .pdf files (in the FINALexamples folder) correspond to words that were not accepted by the transducer in the corresponding file name (according to notation rules).

The word2lemma fst developed works well in many cases, providing a morphological classification. But it fails for many words on the Portuguese language. Every fst on the tests cases, that consists only in letters form a to z (and their variations), is accepted in the word2lemma fst as verb, on the third singular person of the present.

This happens due to the fact that no character is added to this particular case (only the r is removed) on the lemma2verbip fst, that is used to generate the lemma2word and subsequently the word2lemma. I have included two tests to prove that empirically (test5 and test6).