HFST Practical

Lorenzo Tosi

21 ноября 2018 г.

Содержание

1 Phonological Rules 1

2 Productive derivation 1

3 Numerals and abbreviations 1

4 Coverage 2

5 Guesser 2

6 Weighting 2

7 Files Included in the Package 2

1 Phonological Rules

I leave the rule for %{м%} as an exercise for the reader:
"Non surface {м} if following %{Ă%}: followed by н"
%{м%}:0 <=>_ %>: %{Ă%}: н ;

What does minimisation do?
Minimisation tries to reduce as much as possible the steps in the transducer diagram (epsilon removal, path reworking). The goal is to make it as deterministic as possible using .

2 Productive derivation

In order to produce the mor file we need to invert our gen.hfst transducer, otherwise we will not be able to see to weight parameter through the lookup command.

3 Numerals and abbreviations

The .lexc file given is incorrect as it marks саккăр and саккăр as having a front vowel when they have a back one %{є%} instead of %{л%}, this has been corrected in my version.
To implement phonological rules I implemented the archiphoneme %{N%}:p %{N%}:т to deal with the variation and in the .lexc file was added this class in CASES: (this to make it virtually working with all classes to augment coverage) %<abl%>:%{N%}%{A%}н # . The back vowel harmony was added in the previously created rule in order to make it scalable for classes other than numerals and make the .twol file smaller.
"Back vowel harmony for archiphoneme A"
%{A%}:a <=>BackVow: [Cns: | %>: | %{л%}: | %- | %{N%}:]+ _ ;
I added the archiphoneme for back and front vowel in the related letter set and I created a new set for н,л and р, the letter involved in the ablative alternation. The set is called NumCns. The rule to implement it looks like this:
"Consonant agreement N for ablative"
%{N%}:т <=>[NumCns | %{л%}:] [BackVow: | FrontVow: | %-]* _ ;
Note that we need to add NumCns as both forms (without the :) as otherwise will break the rule.

4 Coverage

Coverage is only 0.01939944035553883415 for the transducer without guesser, even if rules look all working as expected through testing, this can be due to a more recent and bigger Wikipedia Dump used in comparison to the one of the practical. With the guesser included it raises to 9.38403837562019422955.

5 Guesser

You will note how it will generate a guess analysis even if the stem is in the lexicon. How do you think you might be able to avoid this analysis?
This can be achieved through weighting as the example in the first section, by giving guessed forms an arbitrary weight, higher than the other weights but lower than the "zero-form"weight.
Another option would be to move the guesser under the Nouns LEXICON section, making it separate to other nouns. Recompiling the transducer in this fashion works exactly as expected.

6 Weighting

The results are the following:
echo "область hfst-lookup -qp chv.surweights.hfst
область область 11.423800

```
echo "облаç hfst-lookup -qp chv.surweights.hfst
облаç облаç 10.027500
```

7 Files Included in the Package

The file included should be enough to reproduce all the cases of this report and the practical by recompiling them as needed. (Other file used are exactly the same as the ones given in the practical, or are produced from those one by compiling them as per the tutorial.)

ava.lexc

ava.twoc

chv.freq

chv.lexc

chv.twol

fin.lexc