UZBEK AFFIX FINITE STATE MACHINE FOR STEMMING

Sharipov Maksud
Salaev Ulugbek
Yuldashev Ollabergan
Sobirov Jasur

Urgench State University
OUTLINE

- Proposal
- Introduction
- Analysing
- Design
- Results
- Summary
- References
PROPOSAL

• Develop a morphological analysis tool for Uzbek stemming:
  o Morphotactic rules will be considered to achieve right stem
  o Finite State Machines (FSMs) will be designed
  o Avoid including lexicon
  o Perform morphologic analysis of a words from a large amount of text in high speed
Introduction

- In linguistic morphology and information retrieval, stemming is the process of reducing inflected or derived words to their word stem, base or root form.
- *Stemming* is an algorithm that extract the morphological root of a word.
- A **computer program** may be called a stemming program, stemming algorithm, or stemmer.

Source: http://mlwiki.org/index.php/Stemming
Introduction

Uzbek language spoken in general in Uzbekistan (and some other places in Central Asia). It is left to right written language. Uzbek language is an agglutinative, it is a language whose words are generated by adding affixes to the root forms. In such languages, given a word in its root form we can drive a new word by adding affix, and so on. Thus in many cases, a single Uzbek word may correspond to a many-word sentence or phrase in a non-agglutinative language. The language has a rich morphological structure.

Word form structure

Surface form of a word

Prefix + Root + Derivational suffixes + Lexical suffixes + Grammatical suffixes

Inflectional suffixes

Source: Azim Hojiyev, O’zbek tili morfologiyasi, morfemikasi va so’z yasalishining nazariy masalalari. Toshkent-2010
ANALYSING

● In Uzbek, there are also suffixes that are written in the same form, but have different meanings. Herewith, there are suffixes that are formed by adding two or more suffixes side by side. For example, there are suffixes –chi, –lik, but there is also a monolithic form suffix –chilik according to the rule of separation of morphemes. When analyzing of the suffix –chilik can be a combination of 2 suffixes side by side ([–chi] +[–lik]) or mono form suffix (–chilik) depending to the content.

● gul+[–chi]+[–lik] gulchi (eng flower), gulchilik (eng floriculture)

● dehqonchilik (dehqonchi, it is incorrect word)

● There is also a suffix that is contain three suffixes side by side:

[–gar]+[–chi]+[–lik], odamgarchilik (eng humanity)
In the affix stemmer, while the lexicon is not used, it is planned to looking for a longer suffix (an undivided suffix) instead of one of the adjacent suffixes.

In Uzbek, inflectional suffixes generally come after derivational suffixes. Some exceptional derivational suffixes that come after inflectional suffixes are: like o’ch/ir/gich (–gich - derivational suff., –ir - inflectional suff.).

The suffix “–lar” besides of plural has a another meaning which is indicating his/her greeting to pointed person. Indeed, when the suffix used in a word, it formed by a different morphological structure:

- dada/m/lar: root(dada) + Possessive (m) + Greeting (lar) → my father
- olma/lar/im: root(olma) + Plural (lar) + Possessive(im) → my apples
DESIGN

System design

Analyzing affixes → Classification affixes → Develop FSMs → Integrate FSMs → Testing
All affixes are classified into classes through their role and coming order.

<table>
<thead>
<tr>
<th>Class #</th>
<th>Class name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tense &amp; Person suffixes</td>
<td>Inflectional</td>
</tr>
<tr>
<td>2</td>
<td>Verb suffixes</td>
<td>Inflectional</td>
</tr>
<tr>
<td>3</td>
<td>Relative verb suffixes</td>
<td>Inflectional</td>
</tr>
<tr>
<td>4</td>
<td>Derivational suffixes</td>
<td>Derivational</td>
</tr>
<tr>
<td>5</td>
<td>Noun suffixes</td>
<td>Inflectional</td>
</tr>
<tr>
<td>6</td>
<td>Number suffixes</td>
<td>Inflectional</td>
</tr>
<tr>
<td>7</td>
<td>Prefixes</td>
<td>Derivational</td>
</tr>
</tbody>
</table>
METHODOLOGY

An affix in Uzbek can have multiple allomorphs in order to provide sound harmony (as the phonological rules) in the word to which it is affixed.

- G: g, k, q
- Y: a, y
- K: k, g
- Q: k, g, g’, q
- T: t, d
- (): the letter between parentheses can be omitted

The number of affixes of all derivational types is 77 and the number of with allomorphs is 87. The number of all affixes of inflectional types is 95 and the number of with allomorphs is 135.

<table>
<thead>
<tr>
<th>Class #</th>
<th>Affixes</th>
<th>Allomorphs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>31</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td>222</td>
</tr>
</tbody>
</table>
DESIGN

- Identifying all affixes and perform morphologic analysis of a word from;
- Classified affixes into classes;
- Designed FSMs for each class which make a reversed order (right to left) analysis of a word. Performed following stages turn by turn to create FSMs:
  - Designing a left to right FSM;
  - Identification the affixes;
  - Inverting the left to right FSM and obtaining a non-deterministic finite state automaton (NFA);
  - Converting NFA to a deterministic finite automaton (DFA) and constructing the right to left FSM;
Relative verb suffixes left to right FSM
Relative verb suffixes right to left FSM

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–lat</td>
<td>6</td>
<td>–(i)n</td>
</tr>
<tr>
<td>2</td>
<td>–Tir</td>
<td>7</td>
<td>–(i)l</td>
</tr>
<tr>
<td>3</td>
<td>–ir</td>
<td>8</td>
<td>–t</td>
</tr>
<tr>
<td>4</td>
<td>–sat</td>
<td>9</td>
<td>–(i)sh</td>
</tr>
<tr>
<td>5</td>
<td>–ar</td>
<td>10</td>
<td>–Qaz</td>
</tr>
</tbody>
</table>
DESIGN

Relationship of the FSMs

- Prefixes
- Relative Verb Suffixes
- Derivational Suffixes
- Verb Suffixes
- Number Suffixes
- Noun Suffixes
- Tense & Person Suffixes
The output taken after the analysis of the word “bajartirilmayaptimi” (“Is it not being performed?”) by the main FSM.

<table>
<thead>
<tr>
<th>Word, affix</th>
<th>Definition</th>
<th>Affix class</th>
</tr>
</thead>
<tbody>
<tr>
<td>bajar</td>
<td>stem, verb</td>
<td></td>
</tr>
<tr>
<td>–tir</td>
<td>relative verb suffix</td>
<td>Relative verb suffix</td>
</tr>
<tr>
<td>–il</td>
<td>relative verb suffix</td>
<td>Relative verb suffix</td>
</tr>
<tr>
<td>–ma</td>
<td>negative verb suffix</td>
<td>Verb suffix</td>
</tr>
<tr>
<td>–yap</td>
<td>continuous tense suffix</td>
<td>Tense &amp; Person suffix</td>
</tr>
<tr>
<td>–ti</td>
<td>3(^{rd}) single person suffix</td>
<td>Tense &amp; Person suffix</td>
</tr>
<tr>
<td>–mi</td>
<td>question suffix</td>
<td>Tense &amp; Person suffix</td>
</tr>
</tbody>
</table>
SUMMARY

● Examined all possibly affixes in Uzbek language and analyzed their morphotactic rules
● Classified the affixes into class respectively their role
● Created the FSMs for each class and main combining FSM
● Aviability moderating structure by affix addition and list exceptions in xml file
● Created the complete software for Uzbek stemming according the algorithm
REFERENCES

- http://dx.doi.org/10.14257/ijca.2016.9.11.13
Thank you!