Automatic Derivation of Nouns from Adjectives in Pashto

Tariq Naeem
Muhammad Abid Khan
Outline

• Introduction
• In the World
• Problem Statement
• Data Collection / Analysis
• Modeling of Data
• Implementation
• Error Analysis
• Conclusion
Introduction

• Derivations, in computational linguistics, are also called lexeme formation and word formation.
• Derivational affixes change completely the grammatical category of the root words to which they are connected. For e.g.

<table>
<thead>
<tr>
<th>Adjectives</th>
<th>Nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wicked</td>
<td>Wickedness</td>
</tr>
<tr>
<td>Dark</td>
<td>Darkness</td>
</tr>
<tr>
<td>Foolish</td>
<td>Foolishness</td>
</tr>
</tbody>
</table>
In the World..

- There are quite a lot of Natural Language Processing (NLP) applications.
- Morphological analysis is a pre-imperative in different areas of NLP, for instance:
  - Lemmatizers
  - Parsers
  - Stemmers
  - Spell checkers
  - Part of Speech taggers
  - Speech Recognition systems
We looked into the work of several other researchers in the field of derivation in other languages.

1. Paumari verbs as derivational additions depicting particular affixes.

2. Lexemes in Malay were inspected to find the association between morphologically stems which can adapt to derivational morphology in pairs.
Problem Statement

• Does class changing derivational morphological analysis exist in Pashto language?

• Under what circumstances, words in Pashto language change their grammatical category from adjectives to nouns?
3. Arabic language showed that part of words are deduced from stems by insertion of affixes.

4. Turkish grammar was examined for dealing with productive derivational strategies.

5. Hindi original root words were examined to find patterns arranged by understanding the properties of the affixes.
Data Collection

• Developing corpus of Pashto language was the first and foremost activity.

• Data was collected from different areas, like, news, memos, letters, research articles, books, fiction, sports and magazines, making it a representative corpus.
Data Analysis

- Pashto corpus was created using the corpus improvement tool XML Aware Indexing and Retrieval Architecture (XAIRA).

- Tagging of Pashto text was done in Extensible Markup Language (XML)

- XAIRA tagged files were used in SQL Server 2012 tables.
The table has entry against every grammatical class.
The derivations were broke down into stems and affixes.
Modeling of Data

- We model our corpus data into Finite State Transducers (FSTs).

- We used FSTs because they are mathematically derived models which efficiently compute many useful NLP functions and weighted transitions on strings.

- They are efficient and more effective.
The eight (8) unique rules defined were modeled by FSTs.

1. The adjective تور (Black) to a noun توروالي (Blackness) and adjective كلك (hard) to a noun كلك والي (hardness). It is shown as:
3. تور [tor] ‘dark’ or ‘black’ تیاره [tiyarah] or تیارا [tiyara] ‘darkness’ or ‘blackness’.
Contd..

6. وږي ([wagey]) ‘hungry’ or لور (lawaga) ‘hunger’
7. بیل [bael] ‘separate’, بیلتون [baeltoon] ‘separation’;
Design and Flowchart of System

Start

User Input

IF Adjective

No

If not, Prompt Msg

Yes

Change it to Noun

Display Output

Exit

Input

Derivational Morphological Analyzer for Pashto

SQL Server Database

Pashto Corpus

Output
Implementation

- We used four (4) programming languages/tools.
  1. We used XEROX LEXC compiler to draw finite state diagrams from the lexicon of adjectives and nouns in Pashto.
  2. After finite state diagrams, we used XEROX XFST to compile the .txt files generated by LEXC.
  3. MS SQL Server was used for corpus and
  4. MS CSharp was used as a front end to fetch data from the database.
Error Analysis

- A sample of 174 nouns and 169 adjectives were collected from the written Pashto data and given to the system as input.
- Out of this input 147 nouns and 142 adjective words were accurately examined.
- Consequently, the complete accuracy of the overall system is:

\[
\frac{(147+142)}{(174+169)} \times 100 = 84.25\%
\]
Conclusion

• The collation feature for Pashto language is not available in older versions of MS SQL Server.

• It was made available in MS SQL Server 2012 to store Pashto text.

• Derivation is sensitive because a slight change of َ [zabbar], َ [zeir] and ُ [paish] can be disastrous.
Future Work

• Further work can be done on class maintaining derivatives in Pashto language.

• Also, the corpus size can be increased to achieve higher accuracy for derivational rules.
Questions please.