A step towards Torwali machine translation: an analysis of morphosyntactic challenges in a low-resource language

Naeem Uddin
Torwali Research Forum
naeemuddinhadi@gmail.com

Jalal Uddin
Torwali Research Forum
torwalipk@gmail.com

Abstract

Torwali is an endangered language spoken in the north of Pakistan. It is a computationally challenging language because of its RTL Perso-Arabic script, non-concatenative nature and distinct words alterations. This paper discusses issues and challenges regarding grammatical structure, divergence in terms of lexicon as well as morphological make-up for the machine translation of a less studied language. It includes creation of NLP tools such as parts of speech (POS) tagger and morphological analyser with HFST which is based on the idea of building lexicon and morphological rules using finite state devices. This work, on which this paper is based, will be a source of Torwali finite state morphology and its future computational growth as electronic dictionaries are usually equipped with morphological analyser and it will also be helpful for developing language pairs.

Key words: Machine Translation, Low-resource language, Morphological analysis, language pairs

1 Introduction

Torwali belongs to the Kohistani sub-group of the Indo-Aryan Dardic languages, spoken in the upper reaches of district Swat of northern Pakistan. It has two dialects (the Bahrain and Chail dialects), with a total of approximately 90,000 to 100,000 speakers.

Torwali is written in a cursive, context sensitive Perso-Arabic script from left to right having unique grammar (morphology + syntax). Being a marginalized and low resource language, there are no robust morphology sources which hinders progress in NLP (Natural Language Processing) tools for Torwali thou there is a digital Torwali dictionary available along with some structured data. This paper discusses an attempt to create a morphological analyzer using HFST from scratch.

In NLP, morphological analysis is used to identify the morpheme and affixes of words in a language and individual words are analyzed into their components. Apart from computational linguistics, there are other uses that require morphological analysis e.g text processing, information retrieval and user interfaces.

Morphologies nowadays are commonly written by using special purpose languages based on finite state technology, one of them is HFST which is based on regular expressions.

2 Goals

The goal of this study is to create a baseline system that paves way for machine translation of Torwali which will cover:

- POS tagging
- Creating a lexc
- Basic inflection rules using twol (two level rule)

3 Unicode and input method

Unicode UTF-8 encoding is used as an encoding scheme as XFST/HFST files are always treated as UTF-8.

As an Input tool TRF phonetic keyboard (TRF 2L V1.0) is used which is developed so that users can easily input texts without going on-screen.

4 Morphological analysis using HFST

HFST-Helsinki finite-state Technology is a framework for compiling and applying linguistic descriptions with finite state methods. Finite-state transducers methods are useful for solving problems involving language identification via morphological processing and POS tagging. There are two principle files in a morphological transducer in HFST, a lexc file which is concerned with morphotactics i-e about the way morphemes are joined together in a word and twol file is used to describe phonological and orthographical alternation rules i-e about what happened when the morphemes are joined together.

For morphological analysis of Torwali HFST/finite state transducers are chosen because Torwali language is quite immature for statistical machine translation and also this implementation is done using Apertium, which is an open source machine translation platform in which HFST can be used.

4.1 LEXC

Lexicon **C**ompiler or LEXC is a finite-state compiler also called a lexical transducer that reads set of morphemes and their morphotactic combinations in order to create finite-state transducer of a lexicon. LEXC contains morphemes grouped in sub-lexicon sets which in turn contain finite strings separated by ‘:’ and a continuation class (a lexicon name).

4.2 TWOLC

TWOLC, **Two-L**evel **C**ompiler is a two level rule compiler used for compiling grammars of two levels into finite state transducers sets. Two level rules are constraints on lexical word forms corresponding to surface forms, It describes morphological alternations such as ژینگ:ژینگو (*weep: weeping*), بن:بنو (*say: saying*). It takes surface forms produced by LEXC and applies rules on them; the rules vary depending on morphological alteration of stem, morphologically or phonologically conditioned deletion of suffix, morphologically or phonologically conditioned insertion, morphologically or phonologically conditioned symbol change.

5 Torwali Morphology

Torwali has a unique morphology because it is basically a fusional language which uses several strategies like stem modification, reduplication and existence of words in inflected form, derived form, compound form and root form. The morphological analyzer separates root and suffix morphemes in all lexical entries i-e in أمیژیل and لناچا, /أمیژ/,/لن/ are roots and /یل/,/چا/ are suffixes. The purpose of this section is to discuss Torwali morphology and its implementation in HFST for main grammatical categories of Torwali i-e nouns, verbs, adjectives and pronouns.

5.1 Nouns

In Torwali, nouns are inflected for number and case and the stem can be joined by an optional plural suffix and an optional oblique case marker. Torwali uses several strategies to mark plurality but the primary morphological method is tone along with verb agreement like for most of the singular nouns have a tone with rising pitch from low-to-high and their plural counterparts have a tone with low pitch. Due to the issue related to representing tone, Torwali words which use tone to mark plurality the following approach is used where singular/plural for masculine and feminine are handled in a single paradigm.

Table 1: tags

tag	description
N-M	Noun masculine
N-F	Noun feminine
N-MF	Noun masculine/feminine

```
Multichar_Symbols
! Part of speech categories
%<n%>     ! Noun

! Number morphology
%<pl%>  ! Plural
%<sg%>  ! Singular

! Gender
%<m%>   ! Masculine
%<f%>   ! Feminine

LEXICON Root
          NounRoot ;
LEXICON N-M
%<n%>%<m%>%<sg%>: # ;
%<n%>%<m%>%<pl%>: # ;
LEXICON N-F
 %<n%>%<f%>%<sg%>: # ;
 %<n%>%<m%>%<pl%>: # ;
```

And, words are added in the lexicon in the following way:

```
LEXICON NounRoot

   آن:آن N-M  ;
   أر:أر N-F  ;
```

If we compile the lexicon with `hfst-lexc` and test it with `hfst-fst2strings`, it spits out the following result:

```
    آن<n><m><sg>:آن
     آن<n><m><pl>:آن
     أر<n><f><sg>:أر
     آن<n><f><pl>:أر
```

The above output marks singularity and plurality for words having tonal change but have no description about the tone's pitch. To make plural oblique of nouns a suffix /e/, /ے/ is added to the stem as in the words; /خار/, /خارے/ and /شان/, /شانے/.

Reduplication is another strategy to communicate plurality and intensity but not in the same way as tone does. For instance, /گال مال/, /میل گیل/, /چُن چُن/, /پہٹ پہٹ/, /چئ مئ/.

Torwali noun forms can be derived from adjectives which can be implemented by adding the suffix /اچا/ making noun root follow a continuation class:

```
%<%n>%<nder%>%<m%>:%>اچا # ;
%<%n>%<nder%>%<f%>:%>اچا # ;
```

Where `%<nder%>` is tag for derived noun.

For noun inflection which undergoes stem modifications, there is a lot of complexity regarding standard rule formation for them; here is a general conclusion:

For majority of masculine nouns the vowel changes form \a\ to \ə\ and for feminine nouns \a\ to \æ\ but some masculine and feminine nouns behave differently. For morphological alteration of the stem the following rules must be implemented using twolc, taking the surface forms produced by lexc.

- Delete (a,ا) for making plurals of masculine singular nouns, when (a,ا) follows a consonant as in /ناد/ ,/دن/ and /ژان/ ,/ژن/
- Replace (a,ا) by (æ,أ) for making plurals of feminine nouns when the (a,ا) follows a consonant as in /بات/ , /بأت/ and /ژات/ , /ژأت/ .
- For noun inflections relating sizes; replace (a,ا) by (æ,أ) to mark small size of large-size nouns e.g. /لہاڑ/ ,/لہأڑ/.

5.2 Verbs

Torwali verbs inflect for tense, aspect, mood and gender and most of the verb forms make gender and number distinction only, no distinction for person. Torwali has three tenses: present, past and future. The suffix /i/, /ی/ can be used to mark feminine singular forms and present tense on feminine singular forms, /u/, /و/ as masculine singular suffix and present tense on masculine singular forms with the suffix /i/, /ی/ being used for present tense on plural forms too. For infinite verbs the suffix /u/ is added to the stem.

To make a test, only present tense on masculine and feminine, infinitives, transitive and intransitive forms of verb are selected and in the continuation class the suffixes are added to mark inflection associated with each of them which are defined with suitable tags as shown below.

```
Multichar_Symbols
! Part of speech categories
%<v%>     ! Verb

! Number morphology
%<pl%>  ! Plural
%<sg%>  ! Singular

! Gender
%<m%>   ! Masculine
%<f%>   ! Feminine

! Verb forms
%<pres%>  ! Pres
%<inf%>   ! Infinitives
%<vt%>    ! Verb Transitive
%<vi%>    ! Verb Intransitive
! Other symbols
%>      ! Morpheme boundary

LEXICON Root
              Verbs ;

 LEXICON V-INF
 %<v%>%<inf%>:%>و # ;
 %<v%>%<m%>%<pres%>:%>دو # ;
 %<v%>%<f%>%<pres%>:%>جی # ;
 LEXICON VT
 %<v%>%<vt%>:%>ؤ # ;
 LEXICON VI
 %<v%>%<vi%>:%>ہوؤ # ;
```

Now if we add some verbs,

```
LEXICON Verbs
ژینگ:ژینگ V ;
بن:بن V-INF ;
سُلُو:سُلُو VT ;
سُورَئ:سُورَئ VI ;
```

The analyzer analyses these verb forms in the following way when compiled with `hfst-lexc` and tested with `hfst-fst2strings`:

```
$ hfst-fst2strings trw.lexc.hfst
```

```
ژینگ<v>:ژینگ
بن<v><inf>:و<بن
بن<v><m><pres>:دو<بن
بن<v><f><pres>:جی<بن
سلُو<v><vt>:ۆ< سلُو
سُورَئ<v><vi>:ہوۆ< سُورَئ
```

From the above output it is concluded that the following implementations can be done. These rules can be applied to verbs whose stem ends with a consonant.

- Adding a suffix /دُد/ to represent past tense on infinitive verb.
- Adding a suffix /نین/ to mark future tense on finite verb.
- Adding suffix /سأت/ to mark inceptive of infinite verb.
- For present perfective on masculine singular adding suffix /و/ and /ی/ for both plurals and feminine singular.
- Suffix /ودو/ for masculine singular, /یجی/ for feminine singular and /یدی/ for plurals to mark present perfective on finite verbs
- Suffix /وشو/ for masculine singular and /یشی/ for both feminine singular and plurals to mark past perfective on finite verb.

Verbs ending with a vowel inflect differently they sometimes behave like verbs with consonant ending stems with a minor modification some plural forms tend to have /اُ/with the stem before applying the plural suffix; however most of the verbs with vowel-final stem follow different configurations.

5.3 Adjectives, Pronouns, Adverbs and closed classes

In a similar way Adjectives, Pronouns, Adverbs, Postpositions, Conjunctions and Interjections have been implemented with the same level of detail.

6 Result and Conclusions

This work presents a straight forward implementation of Towali morphology analyzer using HFST, which implemented the basic inflections of nouns, verbs and other POS like adjectives, adverbs, pronouns and postpositions being tagged. We found HFST a good choice for implementing Torwali Morphology for now as this is the first ever attempt to implement Torwali morphology using FST. However; to develop a full fledge Morphological analyzer more work has to be done. The major problem which we have to face is the random stem changes in nouns, variation in nouns using change of tone, distinct behavior of vowel ending verbs and nouns. There is a need to learn more about Torwali Morphology regarding the affixes and varying stems and more rules needs to be defined.

7 Future work

This work could further be enhanced to following extensions depending upon the possibility:

- Addition of missing diacritic marks to words
- Technique to interpret tone of nouns to indentify singular/plural nouns by its tone.
- Algorithms to differentiate phonetically similar words.
- A comprehensive implementation of Torwali syntax.

8 References

Beesley, K.R, Karttunen, L.: Finite State Morphology. CSLI Publications, Palo Alto (2003)

K. Lindén, A. Erik, H. Sam, A. P. Tommi, and S. Miikka, "*HFST-Framework for Compiling and Applying Morphologies.*" International Workshop on Systems and Frameworks for Computational Morphology, Springer Berlin Heidelberg, 2011.

K. Linden, E. Axelson , S. Drobac, S. Hardwick, J. Kuokkala, J. Niemi, T. Pirinen and Silfverberg "*HFST—A System for Creating NLP Tools.*", International Workshop on Systems and Frameworks for Computational Morphology, Springer Berlin Heidelberg, 2013.

Ullah, Inam (2019) "Digital Dictionary Development for Torwali, A Less-studied Language: Process and Challenges," *Proceedings of the Workshop on Computational Methods for Endangered Languages*: Vol. 2 , Article 3.

Grierson, George A. 1929. *Torwali: An Account of a Dardic Language of the Swat Kohistan*. Royal Asiatic Society. London.

K. Linden, M. Silfverberg, , T. Pirinen, "HFST Tools for Morphology—An Efficient Open Source Package for Construction of Morphological Analyzers”, In: Mahlow, Piotrowski (eds.) , pp. 28–47,2009.

Lunsford, Wayne. 2001. *An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan. M.A. Thesis, University of Texas at Arlington.*

Ullah, Inam. 2004. *‘Lexical Database of the Torwali Dictionary.’ In The Asia lexicography conference. Chiangmai: Payap University.*