INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

Computational Linguistics 1

Project Report

Submitted by:-
Aamir Farhan (20161078)
Ojaswi Binnani (20161006)

Title

Verb Argument Patterns in Urdu

Introduction

The way through which we share our contents or feelings have always great importance in understanding and processing of language. Parsing is the most suited approach in identifying and scanning what the available sentences express. Parsing is the process in which syntactic structure of sentence is identified using grammatical tags. Phrase and Dependency are two main structure formalisms for parsing natural language sentences. We explored that dependency parsing is more appropriate for Urdu and other free word order languages.

The Shakti Standard Format (SSF) format is generated with the help of a morph analyser and chunking. SSF is a highly readable representation for storing language analysis. It is designed to be used as a common format or common representation on which all modules of a system operate.

Problem Statement

Extracting verb argument patterns from Urdu treebanks.

Method Followed

The project was divided into three phases which were as follows:-

Phase 1 – We built our own parser in python which takes the folder containing ssf files and extracts the verb frames and their frequency corresponding to each verb.

Phase 2 – We used the ssfAPI and we ran it on the ssf files to extract the TAM information for each verb present in the folder.

Phase 3 – Now, using the frequency data generated in phase 1, we give score to each verb frame and from this score we predict the dependencies of the verb used at some other instance.

Discussion

The sample SSF format from an Urdu treebank is shown below :-

```
<body>
<tb number="1" segment="no" bullet="no">
<foreign language="select" writingsystem="RTL"> </foreign>
<text>

<Sentence id='1'>
1	((	NP	<fs name='NP' drel='k2:VGF'>
1.1	سابق	JJ	<fs af='سابق,adj,any,any,,d,,' posn='10' name='سابق'>
1.2	عالمی	JJ	<fs af='عالمی,adj,any,any,,o,,' posn='20' name='عالمی'>
1.3	نمبر	NN	<fs af='نمبر,n,m,sg,3,o,0,0' posn='30' name='نمبر'>
1.4	اےک	QC	<fs af='اےک,num,any,any,,any,,' posn='40' name='اےک'>
1.5	فیڈرر	NNP	<fs af='فیڈرر,n,m,sg,3,o,0,0' posn='50' name='فیڈرر'>
1.6	کو	PSP	<fs af='کو,psp,,,,,,' posn='60' name='کو'>
	))
2	((	NP	<fs name='NP2' drel='r6:NP3'>
2.1	ٹینس	NN	<fs af='ٹینس,n,m,sg,3,d,0,0' posn='70' name='ٹینس'>
2.2	تاریخ	NN	<fs af='تاریخ,n,f,sg,3,o,0,0' posn='80' name='تاریخ'>
2.3	کا	PSP	<fs af='کا,psp,m,sg,,d,,' posn='90' name='کا'>
	))
3	((	NP	<fs name='NP3' drel='k2s:VGF'>
3.1	بہترین	JJ	<fs af='بہترین,adj,any,any,,d,,' posn='100' name='بہترین'>
3.2	کھلاڑی	NN	<fs af='کھلاڑی,n,m,sg,3,d,0,0' posn='110' name='کھلاڑی'>
	))
4	((	JJP	<fs name='JJP' drel='pof:VGF'>
4.1	قرار	JJ	<fs af='قرار,adj,any,any,,,,' posn='120' name='قرار'>
	))
5	((	VGF	<fs name='VGF' stype='declarative' voicetype='active'>
5.1	دیا	VM	<fs af='دے,v,m,sg,any,,یا,yA' posn='130' name='دیا'>
5.2	جاتا	VAUX	<fs af='جا,v,any,any,any,,تا,wA' posn='140' name='جاتا'>
5.3	ہے	VAUX	<fs af='ہے,v,any,sg,3,,ہے,hE' posn='150' name='ہے'>
5.4	۔	SYM	<fs af='۔,punc,,,,,,' posn='160' name='۔'>
	))

</Sentence>

</text>
</tb>
</body>
```

The above is the ssf format of the Urdu sentence "*Sabik aalami number ek Federer ko tennis tarikh ka behtareen khiladi karar diya jata hai*"

In the above sentence, the main verb is "diya". We ran the ssfAPI on this sentence and we got the TAM information about the verb, which is, "jata hai" (VAUX).

Next, we ran a self coded parser on a folder containing the ssf files and we extracted the verb frames and its frequency corresponding to a particular verb. Now, we assigned a score value to each of these frames and we can use these values to predict the dependencies this verb would have at any other instance. Thus, the verb argument patterns are used in form of the verb frame frequencies.

Conclusion

We successfully extracted the Verb frames, its frequencies and the TAM information from the given Urdu tree banks in form of ssf files. We can use this frequency data and scores assigned to each verb frames for predicting the dependencies a verb will have at an instance.

Papers Read and References

SSF: Shakti Standard Format Guide, LTRC, IIIT-H.
Formal specifications of Morphology, HCU
A review on Urdu Parsing, Muhammad Javed Arslan Ali Raza, Gomal University, Pakistan.