DEVELOPING A POS TAGGED RESOURCE OF URDU

Tahira Asif, Aasim Ali, Kamran Malik
Punjab University College of Information Technology (PUCIT), University of the Punjab, Lahore Pakistan
aasim.ali@pucit.edu.pk,

ABSTRACT: *Part of speech (POS) is an important linguistic information which is fundamental in several advanced stages of text processing, like, Named Entity Recognition and Statistical Machine Translation. Several existing POS tagsets are analyzed to define a tagset that has maximal tags. Consequently, 46 POS and 4 morphological tags are used to tag 440,000 tokens in above 20,000 sentences of Urdu corpus of religious text, using bootstrapping assisted by a statistical tagger for human reviewed tagging. Increase in the data size shows gradual improvement in the accuracies for both, seen and unseen vocabulary, with an overall best match of 95.59%.*

1 INTRODUCTION

Part of speech (POS) tagging is a method of identifying the appropriate POS category for a sequence of words in a running text. POS tagged corpus is such a foundation that may be used to understand the advanced features of a language such as syntax, semantic, pragmatics, speech, and others. This paper presents the attempt of *developing a POS tagged resource*. In addition, we also tagged morphological features of each word.

We selected Urdu translation of religious text for this work. Urdu is such a language for which POS tagging is not done on a significant amount of data and if it is done then very little tagged data is freely available.

A supervised approach (using a statistical tagger) is used to assist the tagging process of around 440,000 tokens. Various POS tagsets have been tested on different data sizes, to analyze the impact of each.

2 POS TAGGING FOR URDU LANGUAGE

There have been several efforts made for POS tagging of Urdu data. We have named the tagsets developed in those efforts as T0 [12], T1 [27], T2 [20], and T3 [32] to decide on POS Tagset for this work. Muaz et al used statistical taggers for tagging of news data [20]. Sajjad et al used the statistical tagging approach with and without external dictionary [32]. Hardie used rule based approach and developed a morphologically induced POS tagset, thus having a huge list of tags, for tagging of text from a book and transcription of speech data [12]. POS tagger trained on Hindi text has also been used to tag Urdu text [37]. A larger tagset for Urdu POS tagging has been used to show the reduction in ambiguity [4].

3 FEATURES OF URDU LANGUAGE

Urdu language has various morphological features in different POS categories such as: noun, pronoun, verb, and adjective.

3.1 Noun

In Urdu grammars, generally noun is classified with respect to its structure, meaning, number), and gender.

Nouns are also inflected to show the case such as: nominative, oblique or vocative.

3.2 Verb

It is divided with respect to following types: root, imperfective participles, perfective participles, and infinitive. Verbs can also be categorized as: **(i)** Transitive, **(ii)** Intransitive. Verbs in Urdu language have rich inflectional features. Around 60 inflected form of verb are present [1], [34].

3.3 Adjectives

In case of gender adjective, there are no particular oblique suffixes to handle the plural. When two or more nouns appear in a sentence, then the adjective in gender and number will be according to that head noun which is nearest to that adjective in natural reading order [1], [34].

3.4 Morphological features

Urdu words may have following morphological features:

Gender

Urdu has only two possible values for gender: male and female. The gender male is also used as default when the gender of the word/concept is not available.

Number

There are only two possible values for number in Urdu: singular and plural.

Case

Urdu nouns have three cases at the level of morphology: nominative, oblique, and vocative. When a noun is used to call someone, then it is in its vocative case. When noun is followed by a semantic marker, then noun appears in its oblique case, otherwise it is in its nominative case.

Honor

There are several levels of showing in Urdu. We have noted them as H0, H1, H2, and H3 where H3 denotes the highest level of honor.

4 PROPOSED TAGSET FOR URDU

4.1 Part of Speech (POS) tags

The tagset which is proposed here is modified version of T1 [27]. T1 was designed in order to develop the English-Urdu parallel corpus [20], and is very close to the Penn Treebank tagset of English. Here proposed tagset referred to as "PTM" (Proposed POS

Table 1: Sorted list of POS tags and descriptive titles

Tags POS	Tag Titles
AUXA	Aspectual auxiliary
AUXT	Tense auxiliary
CC	Coordinating conjunction
CD	Cardinal
CM	Semantic case marker
DM	Demonstrative
DMRL	Relative demonstrative
FR	Fractional
FW	Foreign word
I	Intensifier
INJ	Interjection
ITRP	Intensifier particle
JJ	Adjective
JJRP	Adjectival Particle
KER	Serial verb joiner
MOPE	Pre-Mohmil
MOPO	Post-Mohmil
MUL	Multiplicative
NN	Noun
NNC	Combined noun / Noun continued
NNCM	Prepositional noun/ Noun after case marker
NNCR	Combined noun continue / Noun continuation terminated
NNP	Proper noun
NNPC	Proper noun continue
OD	Ordinal
PM	Phrase marker
PR	Personal Pronun
PRP$	Personal possessive pronoun
PRRF	Reflexive pronoun
PRRFP$	Reflexive possessive pronoun
PRRL	Relative pronoun
Q	Quantifier
QW	Question word
RB	Adverb
RBRP	Adverbial particle
SC	Subordinating conjunction
SM	Sentence marker
SYM	Symbol
U	Measuring unit
UNK	Unknown
VB	Verb bare form
VBI	Infinitive verb
VBL	Light verb
VBLI	Infinitive light verb
VBT	Verb to-be
WALA	The word 'wala'

Tagset with Morphological marking). The modification in the T1 is the addition of morphological tags and one additional tag in the POS category. This modification was required in order to make it suitable for the selected data, and to provide additional grammatical information about the words. There are couple of tags in T1 which have been decided to not include in the proposed tagset due to the reasons described in the subsection 5.2 below. Table 1 lists the proposed POS tagset.

4.2 DISCUSSION ON POS TAGS

Differences between demonstrative (DM) and pronouns (PR) are found on the phrase level study. Word is tagged as DM when a demonstrative is followed by a noun in the same noun phrase whereas a pronoun forms a phrase by itself or pronoun appears without a noun as subsequent word.

Adjective either follows the noun or is followed by nouns. Most of the proper nouns are derived from adjective In Urdu language. Similarly, the inflected forms of adjective also come as a noun [34]. Some examples are:

Tag (VBT) and tag (AUXT) occur at the same position in a sentence and sometimes are tagged ambiguously in automatic tagging process [20]. A light verb with VBL tag is added to handle the complex predicates. It is such a verb that does not give a complete meaning in a sentence without the help of a noun or adjective or even a verb. Hence a light verb makes a compound verb by combining a noun, or an adjective, or a verb and gives complete meaning in sentence [1]. Tag (VBI) is used to handle the infinitive verbs. Tag (VBLI) is also used to handle the complex predicates and infinitive light verb makes a compound verb by combining a noun, or an adjective, or a verb and gives complete meaning in sentence [1].

It is a word that joins two or more verb phrases and shows the completion of previous verbs in a sentence. In some sentences, a semantic marker 'kay' is also tagged with tag (KER). For example:

Mohmils are those words that do not have their own meanings. In a sentence, Mohmil cannot occur lonely and always come before/after with a meaningful word.

4.3 Morphological Tags

Table 2 lists the proposed morphological tags. Morphological features are: gender with its two values as masculine and feminine; number with its two values as singular and plural; case with its three values as nominative, oblique, vocative; and honor with its four values as H0, H1, H2, H3. POS tags for foreign word (FW) to deal with cross language words (e.g. Arabic); and unknown (UNK) to provide training space for the out of vocabulary words in the training corpus.

Table 2: List of Morphological tags categorized in according to morphological features.

Morphological Tags			
Gender		Number	
F	Feminine	P	Plural
M	Male	S	Singular
Case		Honor	
NOM	Nominative	H0	Honor Level 0
OBL	Oblique	H1	Honor Level 1
VOC	Vocative	H2	Honor Level 2
		H3	Honor Level 3

4.4 Discussion on Morphological Tags

Nominative case can either be case of subject-verb agreement or object-verb agreement. When subject is in nominative form, then subject will agree with the verb and subject can be

noun or pronoun [2]. If subject is in non-nominative case and if object is in nominative case then object starts to link with verb. Consider below example for object-verb agreement. Nominative case is also observed in different types of sentences [2]. A word is in oblique case, if it is followed by case marker (CM), and it may be noun/pronoun/verb or a word with a special tag WALA. Vocative case of a word is used to call a person. It sometimes plays a role of interjections [34].

5 EXPERIMENTAL SETUP

We perform experiments using TnT tagger [6] on six different tagsets using different training and testing data as mentioned in Table 3.

Accuracies of Know words, Unknown word and Known + Unknkown are calculated against each tagset. Known words are all those words which are part of relevant language model, whereas unknown words are those words that do not exist in language model.

Table 3: Count of words in each version for training data and test data

Version	Training Data	Test Data
I	56415	100044
II	106448	101543
III	184221	55487

The reason of conducting these various experiments is to analyze the results of different tagsets.

Dataset which is tagged using PTM tagset is our basic dataset for experimentation. Using basic dataset we derive dataset with T1, T2, T2M, T3, and T3M tagsets.

For our first experiment we build our model on 56415 words and test on 100044 words. The detail results are mentioned in Table 4.

Table 4: Accuracies on different tagset using Version I data

	T1 (%)	PTM (%)	T2 (%)	T2M (%)	T3 (%)	T3M (%)
Known and Unknown	92.65	78.02	94.32	79.16	93.89	78.70
Known	94.23	80.44	95.64	81.53	95.21	81.03
Unknown	63.05	32.48	69.50	34.63	69.05	34.79

Results in Table 4 shows that the best accuracy rate is achieved on the dataset tagged with T2 that is 94.32 %, whereas on 'T3' accuracy rate is 93.89%, and on T1 based data set is 92.65%. Similarly, after adding the morphological features to T1, T2, 'T3', we again train and test the tagger on Version I data. After adding morphological information accuracies of PTM, T2M and T3M on Known and Unknown are 78.02%, 79.16% and 78.70% respectively. Above experiments show that by adding morphological information accuracy decreases.

For second experiment we used 106448 words for training and 101543 words for testing. We use same tagsets for accuracies and build six models. In this experiment training data is much larger than the previous one, which causes higher accuracies than previous one. The detailed results are given in Table 5. As the results show that by building model on large training data all tagsets produce better results than the previous one.

Table 5: Accuracies on different tagset using Version II data

	T1 (%)	PTM (%)	T2 (%)	T2M (%)	T3 (%)	T3M (%)
Known and Unknown	93.93	79.89	95.21	80.73	94.95	80.35
Known	94.89	81.44	95.98	82.25	95.71	81.83
Unknown	63.16	30.43	70.26	32.31	70.62	33.06

For third experiment 184221 words and 55487 words are taken for training and testing respectively. Model is trained and tested on using all six datasets with different tagsets. The detail results are shown in Table 6.

Table 6: Accuracies on different tagset using Version II data

	T1 (%)	PTM (%)	T2 (%)	T2M (%)	T3 (%)	T3M (%)
Known and Unknown	94.58	78.22	95.60	78.90	95.31	78.44
Known	95.17	79.20	96.06	79.80	95.77	79.35
Unknown	68.82	35.13	75.50	39.40	75.26	38.76

After analyzing the accuracies, it is observed that some of the dataset accuracies increases and some of the dataset accuracies decreases.

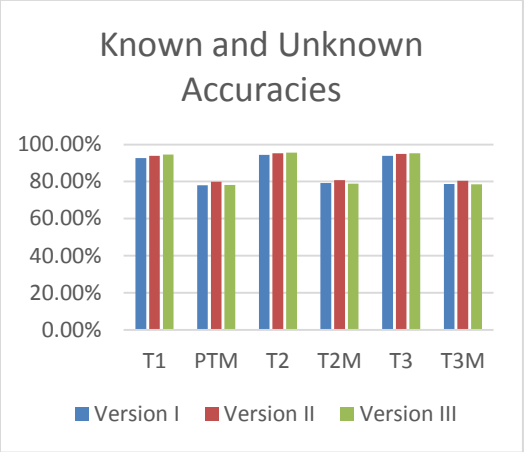

Figure 1: Known and Unknown Accuracies with respect to each tagset.

Details of accuracies of Known + Unknown, Known and Unknown with respect to each Tagset is mentioned in Figure1, Figure 2 and Figure 3 respectively.

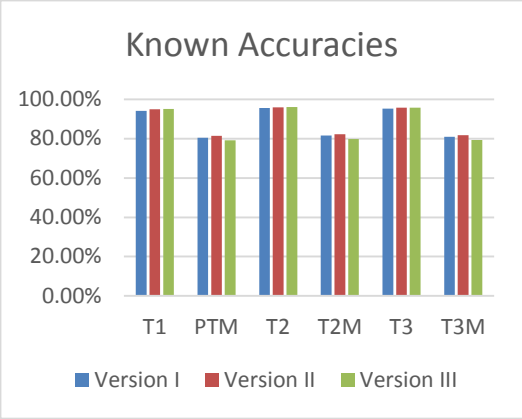

Figure 2: Known Accuracies with respect to each tagset

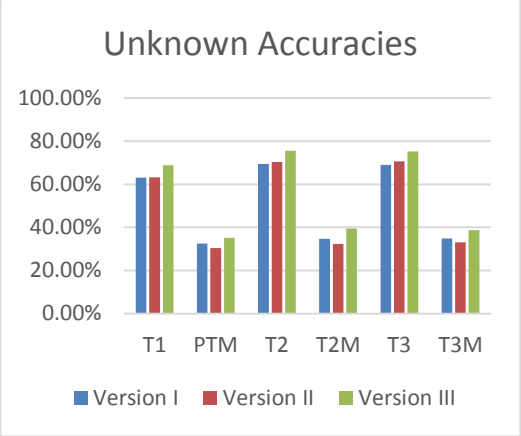

Figure 3: Unknown Accuracies with respect to each tagset

6 CONCLUSION

A quick view of all results with respect to overall results, known words, unknown words cases on combined test set is presented in above table: 5.0.

In this study, originally three models were built as our basic language models. These models were varied from each other with respect to their knowledge. Later on, more fifteen models were built with the help of basic models in this chapter. As a consequence, this chapter covers total eighteen (18) language models with three versions. All these models were applied on the chunk test data as well as on the combined test data and accuracies were achieved with differences in their rates.

Table 7: Misclassified Tags in corpus, based on PTM during TnT tagging

Assigned Tag	Correct Tag	Assigned Tag	Correct Tag
NNPC	NNP	VBT	AUXT
NN	NNP	PRRL	DMRL
NNP	NN	DMRL	PRRL
PR	SC	PR	DM
VBL	VB	DM	PR
NNPC	VB	NN	VBI
VB	NN	VBI	NN
NN	VB	VBI	VBLI
VB	VBL	NN	VBLI
QW	VBL	CC	VOC
VB	AUXA	CD	VB
PRRF	PR	VB	CD
PR	CM	DM/ PR	CD
CM	PR	RBRP	JJRP
VBL	KER	NN	U
Q	CC	NN	RB

This incorrect tagging became the cause of degradation in accuracy rates. After removing the incorrect tagging problem in data set, we reached in the experiment phase. In that phase, we performed In first three versions, we identified the tags that were confused with other tags during tagging using the PTM based data set as our basic data set. Following table represents some confused tags in pairs and shows that which tag was incorrect and what was its correct tag in corpus. These confusions between tags were identified during the post editing of all TnT tagged files based on our basic data set (i.e., PTM based data set).

several experiments and got the diverse accuracy rates on different data sets.

Here we can analyze that what were the reasons of low and high accuracy rates on different data sets having the same text with same statistics in each version. So, the reasons which we identified are following:

Tagsets which we chosen were syntactic based. Some tagsets among them have sub-classes in tags of one POS class, whereas other tagsets have not such classification in that particular POS class. These sub-classifications in tags were not different syntactically and affected on accuracy rates due to incorrect tagging. Similarly, the addition of morphological tags also affected the accuracy rates. These tags only increase the language information in a corpus.

A simple example on accuracy rate variation:

We take the T2 that has only one tag (VB) in Verb POS class and one tag (NN) in noun POS class, whereas T1 have four

sub-classes in Verb POS class, three sub-classes in Noun POS class. All the four sub-classes of Verb POS class are map able onto the single Verb POS class of T2 and syntactically are not different. During manual editing of POS tags on our basic data set, we identified that TnT tagger was confused during the tagging of such POS sub-classes that have no difference at syntax level (shown in above Table 7.0) and affected the accuracy rates. For example: confusion between noun and verb classes in T1 based data set affected the accuracy rates, whereas no such confusion was found on T2 based data set.

Similarly, if we consider the case of additional morphological tags, we can see that accuracy rates on T2M, PTM, T3M based data sets became low than the T1, T2, T3 based data sets in all versions. So, the data sets which are tagged with original tagsets means without morphological tags also have the good accuracy rates as compare to those data sets that have such extra information. So, if we want to increase the language information in data sets, then we have to face the low accuracy rates.

Future Work

The tagset for this work is designed with the view of its direct mapping on other tagsets used in this study. It may be investigated for its mapping to other tagsets like another POS tagset of Urdu [41].

REFERENCES

1. Hardie, A. 2003. Developing a tagset for automated part-of-speech tagging in Urdu. Archer, D, Rayson, P, Wilson, A, and Mc Enery, T (eds.) *Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16. Department of Linguistics, Lancaster University, UK.*
2. Muaz, A., Ali, A., Hussain, S. Analysis and Development of Urdu POS Tagged Corpora, Proceedings *of the 7th Workshop on Asian Language Resources, IJCNLP'09*, Suntec City, Singapore, 2009.
3. Hussain, S. 2008. Resources for Urdu Language Processing. *Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08*, IIIT Hyderabad, India.
4. Sajjad, H. 2007. Statistical Part of Speech Tagger for Urdu. Unpublished MS Thesis, National University of Computer and Emerging Sciences, Lahore, Pakistan.
5. Srivastava, K. A. 2008. Unsupervised Approaches to Part-of-Speech Tagging (Five methodologies survey).
6. Anwar, W., Wang, X., Li. L., Wang. X. A Statistical based Part of Speech Tagger for Urdu Language. *Preceedings of the Sixth International Conference on Machine Learning and Cybernetics*, Hong Kong, 19-22 August 2007.
7. Ali, A. 2010. Study of Morphology of Urdu Language, for its Computational Modeling. Pub: VDM.
8. Schmidt, R. 1999. *Urdu: An Essential Grammar*. Routledge, London, UK.
9. Ali, A. 2011. Syntax of Urdu Language (A survey of Urdu Language syntax). LAP, Lambert Academic Publishing.
10. Brants, T. 2000. TnT – A statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000* Seattle, WA, USA.
11. Urooj, S., Hussain, S., Mustafa, A., Parveen, R., Adeeba, F., Ahmed, T., Butt, M., and Hautli, A. (2014). The CLE Urdu POS Tagset. In LREC proceedings (pp. 2920-2925).