Urdu Part of Speech Tagset

December 07, 2007

CENTER FOR RESEARCH IN URDU LANGUAGE PROCESSING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES, LAHORE
PAKISTAN

Table of Contents

1 Introduction....4
2 Urdu Parts of Speech Classification....5

Revision History

Name	Change Date	Version	Description of Changes
Hassan Sajjad	07-12-2007	1.0.0.0	Initial document

1 Introduction

Tagset of a language caters main parts of speech as well as morphological information of the language. A tagset may be consisted either of syntactic categories or it may be consisted of morpho-syntactic categories. Considering the efficiency in machine learning process and to reduce lexical and syntactic ambiguity, it was decided to concentrate on the syntactic categories of language.

There were three types of corpus available for analysis i.e. literature, news and poetry corpus. For the design of tagset, only literature and news corpus was analyzed. The corpus was based on the most recent available vocabulary used by local people.

2 Urdu Parts of Speech Classification

Demonstrative:
Demonstratives are divided into four categories. All four categories of demonstratives have ambiguity with four categories of pronoun. Phrase level analysis was done to distinguish between demonstrative and pronoun. Following are some examples of demonstratives.

Personal demonstrative (PD)	This category includes the elements of demonstrative and personal demonstratives. Following is an example of it.
ہم، تم، آپ، یہ، وہ، اس	یہ <PD> مسجدیں <NN> ہماری <G> پہچان<NN> ہیں <VB>_<SM>
Relative demonstrative (RD)	جو<RD> لڑکا<NN> صبح<NN> آیا<VB> تھا<TA> وہ<PP> میرا<G>
جو، جن، جنہوں	دوست<NN> ہے<VB>_<SM>
Kaf demonstrative (KD)	کن<KD> لوگوں<NN> کو<P> آم<NN> اچھا<ADJ> لگتا<VB> ہے<TA>_<SM>
کن، کوئی	کمرے<NN> میں<P> کوئی<KD> لڑکا<NN> نہیں<NEG> ہے<VB>_<SM>
Adverbial demonstrative (AD)	میں <PP> ایسا <AD> کام <NN> نہیں<NEG>کر <VB>سکتا<AA>_<SM>
اب، تب، ادھر، یہاں	

Nouns:
Nouns are divided into two categories. First category consists of simple nouns which are represented by NN in the tagset. However, there are other nouns that show adverbial nature like time, place, manner, etc. These are also catered under noun. The proper nouns are kept in a separate category. Following are some examples of different types of nouns.

Noun (NN)	یہ <PD> مسجدیں <NN> ہماری <G> پہچان<NN> ہیں <VB>_<SM>
جہاز، زمین، درخت، لڑکا، اوپر، اندر، سمیت، طرح، طرف	چھت<NN> کے <P> اوپر <NA> حامد<PN> ہے<VB>_<SM>
Proper noun (PN)	لاہور<PN> باغات <NN> کا <P> شہر <NN> ہے<VB>_<SM>
لاہور، پشاور، پاکستان	

Pronouns:
Pronouns are divided into six categories based on their syntactic structure. Most of the categories are consistent with the types provided by Urdu grammarians. Following are some examples of the types of pronouns.

Personal pronoun (PP) میں، ہم، تم، آپ، یہ، وہ، اس	میں <PP> تمھارا <G> دوست <NN> ہوں <VB>_<SM>
Reflexive pronoun (RP) خود، آپ	میں <PP> اپنا <GR> کام<NN> خود<RP> کروں<VB> گا<TA>_<SM>
Relative pronoun (REP) جو، جن، جنہوں	علی<PN> جو <REP> حامد<PN> کا <P> بھائی <NN> ہے<VB> میرا <G> دوست <NN> ہے<VB>_<SM>

Adverbial pronoun (AP):
The adverbial pronouns occur at the place of nouns with adverbial nature and show the property of time, place, manner, etc. They are represented by AP in the tagset. Consider the following examples:

Example: اب، تب، ادھر، یہاں	علی<PN> نے <P> اب <AP> کھانا <NN> کھایا <VB> ہے<TA>_<SM>

Kaf pronoun (KP):
Kaf pronouns add interrogative property in the sentence. They are divided into two categories. Kaf pronouns, represented by KP, are used to ask question about a noun. The second category includes adverbial kaf pronouns which are used at the place of nouns with adverbial nature. Following are their examples:

Kaf pronoun (KP) کون، کوئی، کن	کمرے<NN> میں<P> کون <KP> ہے<VB>_<SM>
Adverbial kaf pro (AKP) کدھر، کب، کیسا	علی <PN> کدھر <AKP> گیا <VB> ہے<TA>_<SM>
Genitive reflexive (GR) اپنا	اپنا<GR> کام<NN> خود<RP> کرنا<VB> میرا<G> فرض<NN> ہے<VB>_<SM>
Genitives (G) میرا، تمھارا، ہمارا، تیرا	Consider the example of genitive reflexive
Verb (VB): At sentence level, any word showing action in any form is considered as verb. No further categorization is done. Consider the following examples of verb:	
Example: لکھنا، کھاتا، جاتا، کرنا	وہ <PP> روٹی<NN> کھا <VB>رہا<AA> ہے<TA>_<SM>

Auxiliaries:
Based on the syntactic nature of language, auxiliaries are divided into two categories. Aspectual auxiliaries always occur after main verb of the sentence. Tense auxiliaries are used to show the time of the action. They occurred at the end of the verb phrase. Consider the examples of aspectual and tense auxiliaries:

Aspectual auxiliary (AA)	Consider the example of verb.
رہا، کرنا، چکھ	
Tense auxiliary (TA) ہے، ہیں، ہوں، تھا، تھے، تھیں، گا، گی، گے، ہو، ہوں	Consider the above describe examples.

Adjective (ADJ):
Adjectives are catered as one category. The information related to the degree of adjective is not taken into account. Following are given some examples of adjectives.

ظالم، خوبصورت، کمزور، بیکار، سمجھدار، نفیس	حامد <PN> بہت <ADV> ظالم <ADJ> لڑکا <NN> ہے<VB>_<SM>

Adverb (ADV):
Adverbs are handled as one category in the tagset. Consider the following examples of adverbs.

Example: بہت، نہایت، بڑا	وہ <PP> بڑا <ADV> محنتی <ADJ> لڑکا <NN> ہے<VB>_<SM>

Quantifier (Q):
Consider following examples of quantifier:

Example: کچھ ، چند،تمام، اتنے، سب، تھوڑا، تھوڑے، کئی، بعض، کل	سب <Q> لوگ<NN> تھوڑا <Q> انتظار<NN> کریں <VB>_<SM>

Numerals:
Numerals are divided into four categories based on their syntactic structure. Cardinal (CA), ordinal (OR), fractional (FR) and multiplicative (MUL) are types included in the tagset. Following are the examples of each category.

Cardinal (CA) ایک، دو، تین، چار بیالیس، انسٹھ، ننانوے، ہزار، دو ہزار	پہلے <OR> دو <CA> لڑکوں <NN> کو<P> بلاؤ <VB> _<SM>
Ordinal (OR)	Consider the example of cardinal.

پہلا، دوسرا، تیسرا، چوتھا، پانچواں، چھٹا، ساتواں، آٹھواں، آخری	
Fractional (FR) چوتھائی، ڈھائی، اڑھائی	ڈھائی <FR> کلو<U> دودھ<NN> دینا<VB>_<SM>
Multiplicative (MUL) گنا، دگنا، دہرا، تہرا	علی <PN>حامد<PN> سے<P> دگنا<MUL> موٹا<ADJ> ہے<VB>_<SM>

Measuring unit (U):
They are frequently used with numerals. However, they have a different syntactic structure than numerals. Consider the example of fractional to see the occurrence of measuring units.

Example: پون، پائو، کلو، سیر	ڈھائی <FR> کلو<U> دودھ<NN> دینا<VB>_<SM>

Conjunction:
Conjunctions are divided into coordinating and subordinating conjunctions. Following are their examples:

Coordinating (CC) یا ,اور	حامد<PN> اور<CC> علی<PN> اچھے<ADJ> دوست<NN> ہیں<VB>_<SM>
Subordinating (SC) کیونکہ ,کہ	حامد <PN> سے<P> کہو<VB>کہ<SC>مجھ <PP> سے<P>ملے<VB>_<SM>

Intensifier (I):
There are only three words in this category. Consider their following examples:

Example: ہی، بھی، تو	میں <PP> بھی <I> آؤں <VB>گا<TA>_<SM>

Adjectival particle (A):
This category includes only one word sa with its two inflection forms. This particle is normally used for comparison. Consider the following examples of adjectival particle.

Example: سا، سے، سی	مینڈک<NN> ایک<CA> عجیب<ADJ> سا<A> جانور<NN> ہے<VB>_<SM>

KER particle (KER):
These particles normally occur in verb phrase. There are only two entities in this class. Consider the following examples:

Example: کے،کر	گھر<NN> پہنچ<VB> کر<KER> فون<NN> کر<VB> دینا<AA>۔<SM>

Title:
Titles are divided into two categories based on their pre and post occurrence around a proper noun. Consider their examples below.

Pre-title (PRT) حضرت، میاں	میاں<PRT> سرمد<PN> صاحب<POT> اچھے<ADJ> انسان<NN> ہیں<VB>۔<SM>
Post-title (POT) جی، صاحب	Consider the example of pre-title above.

Semantic Marker (P):

Following are the list of particles included into this category. However, the entity سے is kept as separate category due to its ambiguous usage.

کا، کو، کی، کے، نے، میں، تک تلک، پر،	حامد<PN> کو<P> علی<PN> نے<P> چھڑی<NN> سے <SE> مارا<VB>۔<SM>
SE (SE): سے	Consider the above example

Wala (WALA):
This category contains one word wala and its inflections. Consider its examples:

Example: والا، والی، والے	پھل<NN> بیچنے<VB> والا<WALA> آدمی<NN> آیا<VB> ہے<TA>۔<SM>

Negation (NEG):
Consider the following examples of negation.

Example: نہ، نہیں	میں <PP> ایسا <AD> کام <NN> نہیں<NEG> کر <VB> سکتا<AA>۔<SM>

Interjection (INT):
Interjections normally occur at the start of the sentence. They are kept as separate category in the tagset. Following are its examples:

Example:	واہ<INT> کیا<ADV> اچھی<ADJ> بات<NN> کی<VB> ہے<TA>۔<SM>
واہ, سبحان اللہ، اچھا	

Question words (QW):
There are some words instead of kaf pronouns that are used for the interrogation in the sentence. However, these words cannot be replaced by a noun or pronoun. A separate category of question words has been formed for these words. Consider their examples below:

Example:	کیا <QW>علی <PN> سکول <NN> جائے <VB> گا<TA>۔<SM>
کیا،کیوں	

Punctuation marks: In this tagset, punctuation marks are divided into two categories. Sentence markers mark the boundary of the sentence. Phrase markers are used inside the sentence but never used at the end of sentence. Consider their examples below:

Sentence marker (SM)	'.', '?'
Phrase marker (PM)	',' , ';'

DATE	2007, 1999

Expression (Exp):
Any word or symbol which is not handled in this tagset will be catered under expression. It can be mathematical symbols, digits, etc.