Tagging Urdu Text with Parts of Speech: A Tagger Comparison

Hassan Sajjad
Universität Stuttgart
Stuttgart. Germany
sajjad@ims.uni-stuttgart.de

Helmut Schmid
Universität Stuttgart
Stuttgart, Germany
schmid@ims.uni-stuttgart.de

Abstract

In this paper, four state-of-art probabilistic taggers i.e. TnT tagger, TreeTagger, RF tagger and SVM tool, are applied to the Urdu language. For the purpose of the experiment, a syntactic tagset is proposed. A training corpus of 100,000 tokens is used to train the models. Using the lexicon extracted from the training corpus, SVM tool shows the best accuracy of 94.15%. After providing a separate lexicon of 70,568 types, SVM tool again shows the best accuracy of 95.66%.

1 Urdu Language

Urdu belongs to the Indo-Aryan language family. It is the national language of Pakistan and is one of the official languages of India. The majority of the speakers of Urdu spread over the area of South Asia, South Africa and the United Kingdom[1].

Urdu is a free order language with general word order SOV. It shares its phonological, morphological and syntactic structures with Hindi. Some linguists considered them as two different dialects of one language (Bhatia and Koul, 2000). However, Urdu is written in Perso-arabic script and inherits most of the vocabulary from Arabic and Persian. On the other hand, Hindi is written in Devanagari script and inherits vocabulary from Sanskrit.

Urdu is a morphologically rich language. Forms of the verb, as well as case, gender, and number are expressed by the morphology. Urdu represents case with a separate character after the head noun of the noun phrase. Due to their separate occurrence and their place of occurrence, they are sometimes considered as postpositions. Considering them as case markers, Urdu has nominative, ergative, accusative, dative, instrumental, genitive and locative cases (Butt, 1995: pg 10). The Urdu verb phrase contains a main verb, a light verb describing the aspect, and a tense verb describing the tense of the phrase (Hardie, 2003; Hardie, 2003a).

2 Urdu Tagset

There are various questions that need to be answered during the design of a tagset. The granularity of the tagset is the first problem in this regard. A tagset may consist either of general parts of speech only or it may consist of additional morpho-syntactic categories such as number, gender and case. In order to facilitate the tagger training and to reduce the lexical and syntactic ambiguity, we decided to concentrate on the syntactic categories of the language. Purely syntactic categories lead to a smaller number of tags which also improves the accuracy of manual tagging[2] (Marcus *et al.*, 1993).

Urdu is influenced from Arabic, and can be considered as having three main parts of speech, namely noun, verb and particle (Platts, 1909; Javed, 1981; Haq, 1987). However, some grammarians proposed ten main parts of speech for Urdu (Schmidt, 1999). The work of Urdu grammar writers provides a full overview of all the features of the language. However, in the perspective of the tagset, their analysis is lacking the computational grounds. The semantic, morphological and syntactic categories are mixed in their distribution of parts of speech. For example, Haq (1987) divides the common nouns into situational (smile, sadness, darkness), locative (park, office, morning, evening), instrumental (knife, sword) and collective nouns (army, data).

In 2003, Hardie proposed the first computational part of speech tagset for Urdu (Hardie,

[1] http://www.ethnologue.com/14/show_language.asp?code=URD

[2] A part of speech tagger for Indian languages, available at http://shiva.iiit.ac.in/SPSAL2007 /iiit_tagset_guidelines.pdf

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 692–700,
Athens, Greece, 30 March – 3 April 2009.

2003a). It is a morpho-syntactic tagset based on the EAGLES guidelines. The tagset contains 350 different tags with information about number, gender, case, etc. (van Halteren, 2005). The EAGLES guidelines are based on three levels, major word classes, recommended attributes and optional attributes. Major word classes include thirteen tags: noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unassigned, residual and punctuation. The recommended attributes include number, gender, case, finiteness, voice, etc.[3] In this paper, we will focus on purely syntactic distributions thus will not go into the details of the recommended attributes of the EAGLES guidelines. Considering the EAGLES guidelines and the tagset of Hardie in comparison with the general parts of speech of Urdu, there are no articles in Urdu. Due to the phrase level and semantic differences, pronoun and demonstrative are separate parts of speech in Urdu. In the Hardie tagset, the possessive pronouns like میرا /mera/ (my), تمہارا /tumhara/ (your), ہمارا /humara/ (our) are assigned to the category of possessive adjective. Most of the Urdu grammarians consider them as pronouns (Platts, 1909; Javed, 1981; Haq, 1987). However, all these possessive pronouns require a noun in their noun phrase, thus show a similar behavior as demonstratives. The locative and temporal adverbs (یہاں /yahan/ (here), وہاں /wahan/ (there), اب /ab/ (now), etc.) and, the locative and temporal nouns (صبح /subah/ (morning), شام /sham/ (evening), گھر /gher/ (home)) appear in a very similar syntactic context. In order to keep the structure of pronoun and noun consistent, locative and temporal adverbs are treated as pronouns. The tense and aspect of a verb in Urdu is represented by a sequence of auxiliaries. Consider the example[4]:

ہے	رہا	جا	کرتا	کام	جان
Hai	raha	Ja	kerta	kam	Jan
Is	Doing		Kept	Work	John
John is kept on doing work					

"Table 1: The aspect of the verb کرتا /kerta/ (doing) is represented by two separate words جا /ja/ and رہا /raha/ and the last word of the sentence ہے /hai/ (is) shows the tense of the verb."

The above considerations lead to the following tagset design for Urdu. The general parts of speech are noun, pronoun, demonstrative, verb, adjective, adverb, conjunction, particle, number and punctuation. The further refinement of the tagset is based on syntactic properties. The morphologically motivated features of the language are not encoded in the tagset. For example, an Urdu verb has 60 forms which are morphologically derived from its root form. All these forms are annotated with the same category i.e. verb.

During manual tagging, some words are hard for the linguist to disambiguate reliably. In order to keep the training data consistent, such words are assigned a separate tag. For instance, the semantic marker سے /se/ gets a separate tag due to its various confusing usages such as locative and instrumental (Platts, 1909).

The tagset used in the experiments reported in this paper contains 42 tags including three special tags. Nouns are divided into noun (NN) and proper name (PN). Demonstratives are divided into personal (PD), KAF (KD), adverbial (AD) and relative demonstratives (RD). All four categories of demonstratives are ambiguous with four categories of pronouns. Pronouns are divided into six types i.e. personal (PP), reflexive (RP), relative (REP), adverbial (AP), KAF (KP) and adverbial KAF (AKP) pronouns. Based on phrase level differences, genitive reflexive (GR) and genitive (G) are kept separate from pronouns. The verb phrase is divided into verb, aspectual auxiliaries and tense auxiliaries. Numerals are divided into cardinal (CA), ordinal (OR), fractional (FR) and multiplicative (MUL). Conjunctions are divided into coordinating (CC) and subordinating (SC) conjunctions. All semantic markers except سے /se/ are kept in one category. Adjective (ADJ), adverb (ADV), quantifier (Q), measuring unit (U), intensifier (I), interjection (INT), negation (NEG) and question words (QW) are handled as separate categories. Adjectival particle (A), KER (KER), SE (SE) and WALA (WALA) are ambiguous entities which are annotated with separate tags. A complete list of the tags with the examples is given in appendix A. The examples of the weird categories such as WALA, KAF pronoun, KAF demonstratives, etc. are given in appendix B.

3 Tagging Methodologies

The work on automatic part of speech tagging started in early 1960s. Klein and Simmons

[3] The details on the EAGLES guidelines can be found at: http://www.ilc.cnr.it/EAGLES/browse.html

[4] Urdu is written in right to left direction.

(1963) rule based POS tagger can be considered as the first automatic tagging system. In the rule based approach, after assigning each word its potential tags, a list of hand written disambiguation rules are used to reduce the number of tags to one (Klein and Simmons, 1963; Green and Rubin, 1971; Hindle, 1989; Chanod and Tapanainen 1994). A rule based model has the disadvantage of requiring lots of linguistic efforts to write rules for the language.

Data-driven approaches resolve this problem by automatically extracting the information from an already tagged corpus. Ambiguity between the tags is resolved by selecting the most likely tag for a word (Bahl and Mercer, 1976; Church, 1988; Brill, 1992). Brill's transformation based tagger uses lexical rules to assign each word the most frequent tag and then applies contextual rules over and over again to get a high accuracy. However, Brill's tagger requires training on a large number of rules which reduces the efficiency of machine learning process. Statistical approaches usually achieve an accuracy of 96%-97% (Hardie, 2003: 295). However, statistical taggers require a large training corpus to avoid data sparseness. The problem of low frequencies can be resolved by applying different methods such as smoothing, decision trees, etc. In the next section, an overview of the statistical taggers is provided which are evaluated on the Urdu tagset.

3.1 Probabilistic Disambiguation

The Hidden Markov model is the most widely used method for statistical part of speech tagging. Each tag is considered as a state. States are connected by transition probabilities which represent the cost of moving from one state to another. The probability of a word having a particular tag is called lexical probability. Both, the transitional and the lexical probabilities are used to select the tag of a particular word.

As a standard HMM tagger, The TnT tagger is used for the experiments. The TnT tagger is a trigram HMM tagger in which the transition probability depends on two preceding tags. The performance of the tagger was tested on NEGRA corpus and Penn Treebank corpus. The average accuracy of the tagger is 96% to 97% (Brants, 2000).

The second order Markov model used by the TnT tagger requires large amounts of tagged data to get reasonable frequencies of POS trigrams. The TnT tagger smooths the probability with linear interpolation to handle the problem of data sparseness. The Tags of unknown words are predicted based on the word suffix. The longest ending string of an unknown word having one or more occurrences in the training corpus is considered as a suffix. The tag probabilities of a suffix are evaluated from all the words in the training corpus (Brants, 2000).

In 1994, Schmid proposed a probabilistic part of speech tagger very similar to a HMM based tagger. The transition probabilities are calculated by decision trees. The decision tree merges infrequent trigrams with similar contexts until the trigram frequencies are large enough to get reliable estimates of the transition probabilities. The TreeTagger uses an unknown word POS guesser similar to that of the TnT tagger. The TreeTagger was trained on 2 million words of the Penn-Treebank corpus and was evaluated on 100,000 words. Its accuracy is compared against a trigram tagger built on the same data. The TreeTagger showed an accuracy of 96.06% (Schmid, 1994a).

In 2004, Giménez and Màrquez proposed a part of speech tagger (SVM tool) based on support vector machines and reported accuracy higher than all state-of-art taggers. The aim of the development was to have a simple, efficient, robust tagger with high accuracy. The support vector machine does a binary classification of the data. It constructs an N-dimensional hyperplane that separates the data into positive and negative classes. Each data element is considered as a vector. Those vectors which are close to the separating hyperplane are called support vectors[5].

A support vector machine has to be trained for each tag. The complexity is controlled by introducing a lexicon extracted from the training data. Each word tag pair in the training corpus is considered as a positive case for that tag class and all other tags in the lexicon are considered negative cases for that word. This feature avoids generating useless cases for the comparison of classes.

The SVM tool was evaluated on the English Penn Treebank. Experiments were conducted using both polynomial and linear kernels. When using n-gram features, the linear kernel showed a significant improvement in speed and accuracy. Unknown words are considered as the most ambiguous words by assigning them all open class POS tags. The disambiguation of unknowns uses features such as prefixes, suffixes,

[5] Andrew Moore:
http://www.autonlab.org/tutorials/svm.html

upper case, lower case, word length, etc. On the Penn Treebank corpus, SVM tool showed an accuracy of 97.16% (Giménez and Màrquez, 2004).

In 2008, Schmid and Florian proposed a probabilistic POS tagger for fine grained tagsets. The basic idea is to consider POS tags as sets of attributes. The context probability of a tag is the product of the probabilities of its attributes. The probability of an attribute given the previous tags is estimated with a decision tree. The decision tree uses different context features for the prediction of different attributes (Schmid and Laws, 2008).

The RF tagger is well suited for languages with a rich morphology and a large fine grained tagset. The RF tagger was evaluated on the German Tiger Treebank and Czech Academic corpus which contain 700 and 1200 POS tags, respectively. The RF tagger achieved a higher accuracy than TnT and SVMTool.

Urdu is a morphologically rich language. Training a tagger on a large fine grained tagset requires a large training corpus. Therefore, the tagset which we are using for these experiments is only based on syntactic distributions. However, it is always interesting to evaluate new disambiguation ideas like RF tagger on different languages.

4 Experiments

A corpus of approx 110,000 tokens was taken from a news corpus (www.jang.com.pk). In the filtering phase, diacritics were removed from the text and normalization was applied to keep the Unicode of the characters consistent. The problem of space insertion and space deletion was manually solved and space is defined as the word boundary. The data was randomly divided into two parts, 90% training corpus and 10% test corpus. A part of the training set was also used as held out data to optimize the parameters of the taggers. The statistics of the training corpus and test corpus are shown in table 2 and table 3. The optimized parameters of the TreeTagger are context size 2, with minimum information gain for decision tree 0.1 and information gain at leaf node 1.4. For TnT, a default trigram tagger is used with suffix length of 10, sparse data mode 4 with lambda1 0.03 and lambda2 0.4. The RF tagger uses a context length of 4 with threshold of suffix tree pruning 1.5. The SVM tool is trained at right to left direction with model 4. Model 4 improves the detection of unknown words by artificially marking some known words as unknown words and then learning the model.

	Training corpus	Test corpus
Tokens	100,000	9000
Types	7514	1931
Unknown Tokens	--	754
Unknown Types	--	444

"Table 2: Statistics of training and test data."

Tag	Total	Unknown	Tag	Total	Unknown
NN	2537	458	PN	459	101
P	1216	0	AA	379	0
VB	971	81	TA	285	0
ADJ	510	68	ADV	158	21

"Table 3: Eight most frequent tags in the test corpus."

In the first experiment, no external lexicon was provided. The types from the training corpus were used as the lexicon by the tagger. SVM tool showed the best accuracy for both known and unknown words. Table 4 shows the accuracies of all the taggers. The baseline result where each word is annotated with its most frequent tag, irrespective of the context, is 88.0%.

TnT tagger	TreeTagger	RF tagger	SVM tagger
93.40%	93.02%	93.28%	94.15%
Known			
95.78%	95.60%	95.68%	96.15%
Unknown			
68.44%	65.92%	68.08%	73.21%

"Table 4: Accuracies of the taggers without using any external lexicon. SVM tool shows the best result for both known and unknown words."

The taggers show poor accuracy while detecting proper names. In most of the cases, proper name is confused with adjective and noun. This is because in Urdu, there is no clear distinction between noun and proper name. Also, the usage of an adjective as a proper name is a frequent phenomenon in Urdu. The accuracies of open class tags are shown in table 5. The detailed discussion on the results of the taggers is done after providing an external lexicon to the taggers.

Tag	TnT tagger	Tree-Tagger	RF tagger	SVM tagger
VB	93.20%	91.86%	92.68%	94.23%
NN	94.12%	96.21%	93.89%	96.45%
PN	73.20%	66.88%	72.77%	68.62%
ADV	75.94%	72.78%	74.68%	72.15%
ADJ	85.67%	80.78%	86.5%	85.88%

"Table 5: Accuracies of open class tags without having an external lexicon"

In the second stage of the experiment, a large lexicon consisting of 70,568 types was provided[6]. After adding the lexicon, there are 112 unknown tokens and 81 unknown types in the test corpus[7]. SVM tool again showed the best accuracy of 95.66%. Table 6 shows the accuracy of the taggers. The results of open class words significantly improve due to the smaller number of unknown words in the test corpus. The total accuracy of open class tags and their accuracy on unknown words are given in table 7 and table 8 respectively.

TnT tagger	Tree-Tagger	RF tagger	SVM tool
94.91%	95.17%	95.26%	95.66%
Known			
95.42%	95.65%	95.66%	96.11%
Unknown			
56.25%	58.04%	64.60%	61.61%

"Table 6: Accuracies of the taggers after adding the lexicon. SVM tool shows the best accuracy for known word disambiguation. RF tagger shows the best accuracy for unknown words."

Tag	TnT tagger	Tree-Tagger	RF tagger	SVM tool
VB	95.88%	95.88%	96.58%	96.80%
NN	94.64%	95.85%	94.79%	96.64%
PN	86.92%	79.73%	84.96%	81.70%
ADV	82.28%	79.11%	81.64%	81.01%
ADJ	91.59%	89.82%	92.37%	88.26%

"Table 7: Accuracies of open class tags after adding an external lexicon."

[6] Additional lexicon is taken from CRULP, Lahore, Pakistan (www.crulp.org).

[7] The lexicon was added by using the default settings provided by each tagger. No probability distribution information was given with the lexicon.

Tag	TnT tagger	Tree-Tagger	RF tagger	SVM tool
VB	28.57%	0.00%	42.86%	42.86%
NN	74.47%	95.74%	80.85%	80.85%
PN	68.18%	54.54%	63.63%	50.00%
ADV	8.33%	0.00%	8.33%	0.00%
ADJ	30.00%	20.00%	70.00%	80.00%

"Table 8: Accuracies of open class tags on unknown words. The number of unknown words with tag VB and ADJ are less than 10 in this experiment."

The results of the taggers are analyzed by finding the most frequently confused pairs for all the taggers. It includes both the known and unknown words. Only those pairs are added in the table which have an occurrence of more than 10. Table 9 shows the results.

Confused pair		TnT tagger	Tree-Tagger	RF tagger	SVM tool
NN	ADJ	85	87	87	95
NN	PN	118	140	129	109
NN	ADV	12	15	13	15
NN	VB	14	17	12	12
VB	TA	12	0	0	0
KER	P	14	14	14	0
ADV	ADJ	11	14	13	11
PD	PP	26	26	30	14

"Table 9: Most frequently confused tag pairs with total number of occurrences."

5 Discussion

The output of table 9 can be analyzed in many ways e.g. ambiguous tags, unknown words, open class tags, close class tags, etc. In the close class tags, the most frequent errors are between demonstrative and pronoun, and between KER tag and semantic marker (P). The difference between demonstrative and pronoun is at the phrase level. Demonstratives are followed by a noun which belongs to the same noun phrase whereas pronouns form a noun phrase by itself. Taggers analyze the language in a flat structure and are unable to handle the phrase level differences. It is interesting to see that the SVM tool shows a clear improvement in detecting the phrase level differences over the other taggers. It might be due to the SVM tool ability to look not only at

the neighboring tags but at the neighboring words as well.

(a)

گے	گائیں	گانا	لوگ	وہ
Gay	gayain	Gana	log	Voh
TA	VB	NN	NN	PD
Will	sing	Song	people	Those

Those people will sing a song.

(b)

گے	گائیں	گانا	وہ
Gay	Gayain	gana	Voh
TA	VB	NN	PP
Will	Sing	Song	those

Those will sing a song.

"Table 10: The word وہ /voh/ is occurring both as pronoun and demonstrative. In both of the cases, it is followed by a noun. But looking at the phrases, demonstrative وہ has the noun inside the noun phrase."

The second most frequent error among the closed class tags is the distinction between the KER tag کے /kay/ and the semantic marker کے /kay/. The KER tag always takes a verb before it and the semantic marker always takes a noun before it. The ambiguity arises when a verbal noun occurs. In the tagset, verbal nouns are handled as verb. Syntactically, verbal nouns occur at the place of a noun and can also take a semantic marker after them. This decreases the accuracy in two ways; the wrong disambiguation of KER tag and the wrong disambiguation of unknown verbal nouns. Due to the small amount of training data, unknown words are frequent in the test corpus. Whenever an unknown word occurs at the place of a noun, the most probable tag for that word will be noun which is wrong in our case. Table 11 shows an example of such a scenario.

(a)

بعد	کے	کرنے	کام
baad	Kay	kernay	kam
NN	P	VB	NN
after	--	doing	work

After doing work

(b)

کے	کر	کام
kay	ker	kam
KER	VB	NN
--	Doing	work

(After) doing work

"Table 11: (a) Verbal noun with semantic marker, (b) syntactic structure of KER tag."[8]

All the taggers other than the SVM tool have difficulties to disambiguate between KER tags and semantic markers.

(a)

دو	خوراک	کو	لوگوں	ضرورتمند
do	khoraak	Ko	log	zarorat-mand
VB	NN	P	NN	ADJ
give	food	To	people	needy

Give food to the needy people

(b)

دو	خوراک	کو	ضرورتمند
do	khoraak	ko	zaroratmand
VB	NN	P	NN
give	food	To	needy

Give food to the needy

"Table 12: (a) Occurrence of adjective with noun, (b) dropping of main noun from the noun phrase. In that case, adjective becomes the noun."

Coming to open class tags, the most frequent errors are between noun and the other open class tags in the noun phrase like proper noun, adjective and adverb. In Urdu, there is no clear distinction between noun and proper noun. The phenomenon of dropping of words is also frequent in Urdu. If a noun in a noun phrase is dropped, the adjective becomes a noun in that phrase (see table 12). The ambiguity between noun and verb is due to verbal nouns as explained above (see table 11).

6 Conclusion

In this paper, probabilistic part of speech tagging technologies are tested on the Urdu language. The main goal of this work is to investigate whether general disambiguation techniques and standard POS taggers can be used for the tagging of Urdu. The results of the taggers clearly answer this question positively. With the small training corpus, all the taggers showed accuracies around 95%. The SVM tool shows the best accuracy in

[8] One possible solution to this problem could be to introduce a separate tag for verbal nouns which will certainly remove the ambiguity between the KER tag and the semantic marker and reduce the ambiguity between verb and noun.

disambiguating the known words and the RF tagger shows the best accuracy in detecting the tags of unknown words.

Appendices

Appendix A. Urdu part of speech tagset

Following is the complete list of the tags of Urdu. There are some occurrences in which two Urdu words are mapped to the same translation of English. There are two reasons for that, either the Urdu words have different case or there is no significant meaning difference between the two words which can be described by different English translations.

Tag	Example
Personal demonstrative (PD)	ہم (we) ، تم (you) ، آپ (you)[9]، یہ (this)، وہ (that)، اس (that)
Relative demonstrative (RD)	جو(that) ، جن(that)، جنہوں(that)
Kaf demonstrative (KD)	کن(whose) ، کوئی (someone)
Adverbial demonstrative (AD)	اب (now) ، تب (then) ، ادھر (here) ، یہاں (here)
Noun (NN)	جہاز (ship) ، زمین (earth) ، لڑکا (boy) ، اوپر (above)، اندر (inside) ، سمیت (with) ، طرح (like)
Proper noun (PN)	جرمنی (Germany) ، پاکستان (Pakistan)
Personal pronoun (PP)	میں(I) ، ہم (we) ، تم(you) ، آپ (you) ، یہ (he) ، وہ (he)، اس(he)
Reflexive pronoun (RP)	خود (myself) ، آپ (myself)
Relative pronoun (REP)	جو(that)، جن(that)، جنہوں(that)
Adverbial pronoun (AD)	اب (now) ، تب (then) ، ادھر (here) ، یہاں (here)
Kaf pronoun (KP)	کون (who) ، کوئی (someone)، ، کن (which)
Adverbial kaf pro (AKP)	کدھر (where) ، کب (when)، کیسا (how)
Genitive reflexive (GR)	اپنا (my)
Genitives (G)	میرا (my) ، تمہارا (your) ، ہمارا (our) ، تیرا (your)
Verb (VB)	لکھنا (write) ، کھاتا (eat) ، جاتا (go) ، کرنا (do)
Aspectual auxiliary (AA)	رہا، کرنا، چکہ[10]
Tense auxiliary (TA)	ہے (is) ، ہیں (are) ، تھا (was)، تھے (were)
Adjective (ADJ)	ظالم (cruel) ، خوبصورت (beautiful)، کمزور (weak)
Adverb (ADV)	بہت (very) ، نہایت (very) ، بڑا (very)
Quantifier (Q)	کچھ (some) ،تمام (all) ، اتنے (this much) ، کل (total)
Cardinal (CA)	ایک(one) ، دو (two) ، تین (three)
Ordinal (OR)	پہلا (first) ، دوسرا (second)، آخری (last)
Fractional (FR)	چوتھائی(one fourth) ، ڈھائی (two and a half)
Multiplicative (MUL)	گنا(times) ، دگنا (two times)
Measuring unit (U)	کلو (kilo)
Coordinating (CC)	اور (and) ,یا (or)
Subordinating (SC)	کہ (that),کیونکہ (because)
Intensifier (I)	ہی، بھی، تو
Adjectival particle	سا (like)
KER	کے، کر
Pre-title (PRT)	حضرت(.Mr) ، میاں(.Mr)
Post-title (POT)	جی ، صاحب(.Mr)
Case marker (P)	کا ، کو ، کی ، کے ، نے، میں، تک تلک ، پر،
SE (SE)	سے
WALA (WALA)	والا، والی، والے
Negation (NEG)	[نہ ، نہیں(not/no)]
Interjection (INT)	واہ (hurrah), سبحان اللہ، اچھا(Good)
Question word (QW)	کیا (what) ، کیوں (why)
Sentence marker (SM)	’?‘, ’.‘
Phrase marker (PM)	’,‘ , ’;‘
DATE	2007, 1999
Expression (Exp): Any word or symbol which is not handled in the tagset will be catered under expression. It can be mathematical symbols, digits, etc.	

“Table 13: Tagset of Urdu”

[9] Polite form of you which is used while talking with the elders and with the strangers

[10] They always occur with a verb and can not be translated standalone.

Appendix B. Examples of WALA, Noun with locative behavior, KAF pronoun and KAF demonstrative and multiplicative.

WALA والا:

Attributive	Demonstrative	Occupation
عزت والا Respectable	یہ والا This one	دودھ والا Milk man

Manner	Possession	Time
آہستہ والا The one with the manner "slow"	کانٹوں والا پھول Flower with thorns	صبح والا اخبار Morning newspaper

Place	Doer	--
باہر والا جوتا Shoes which is bought from some other country	پڑھنے والا The one whose study	-- --

"Table 14: Examples of tag WALA"

Noun with locative behavior:

Adverb	Noun
نیچے والی دکان Down shop	نیچے سے آنا Coming from downstairs

Postposition	Noun
میز کے نیچے Under the table	نیچے جانا Goes down

"Table 15: Examples of noun with locative behavior

Multiplicative:

وہ مجھ سے دگنا (دوگنا) موٹا ہے۔ He is two times fatter than me.

"Table 16: Example of Multiplicative

KAF pronoun and KAF demonstrative:

KAF pronoun
کن لوگوں کو آم اچھے لگتے ہیں ؟ Which people like mangoes?

KAF Demonstrative
کن کو آم اچھے لگتے ہیں ؟ Which one like mangoes?

Adverbial KAF pronoun
وہ کدھر گیا ہے ؟ Where did he go?

"Table 17: Examples of KAF pronoun and KAF demonstrative

References

Bahl, L. R. and Mercer, R. L. 1976. Part of speech assignment by a statistical decision algorithm, *IEEE International Symposium on Information Theory*, pp. 88-89.

Bhatia, TK and Koul, A. 2000. Colloquial Urdu. London: Routledge.

Brants, Thorsten. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000* Seattle, WA.

Brill, E. 1992. A simple rule-based part of speech tagger, Department of Computer Science, University of Pennsylvania.

Butt, M. 1995. The structure of complex predicates in Urdu. CSLI, Stanford.

Chanod, Jean-Pierre and Tapananinen, Pasi 1994. Statistical and constraint-Based taggers for French, Technical report MLTT-016, RXRC Grenoble.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted test, In the proceedings of *2nd conference on Applied Natural Language Processing*, pp. 136-143.

Giménez and Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proceedings of the IV International Conference on Language Resources and Evaluation (LREC' 04)*, Lisbon, Portugal.

Green, B. and Rubin, G. 1971. Automated grammatical tagging of English, Department of Linguistics, Brown University.

Haq, M. Abdul. 1987. صرف و نحو اردو, Amjuman-e-Taraqqi Urdu (Hind).

Hardie, A. 2003. Developing a tag-set for automated part-of-speech tagging in Urdu. In *Archer, D, Rayson, P, Wilson, A, and McEnery, T (eds.)* Proceedings of *the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16.* Department of Linguistics, Lancaster University, UK.

Hardie, A. 2003a. The computational analysis of morphosyntactic categories in Urdu, PhD thesis, Lancaster University.

Hindle, D. 1989. Acquiring disambiguation rules from text, Proceedings of *27th annual meeting of Association for Computational Linguistics.*

van Halteren, H, 2005. Syntactic Word Class Tagging, Springer.

Javed, Ismat. 1981. اردو قواعد نئ, Taraqqi Urdu Bureau, New Delhi.

Klein, S. and Simmons, R.F. 1963. A computational approach to grammatical coding of English words, JACM 10: pp. 334-347.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: the Penn Treebank Computational Linguistics 19, pp. 313-330

Platts, John T 1909. A grammar of the Hindustani or Urdu language, London.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision tree, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

Schmid, H. 1994a. Part-of-speech tagging with neural networks, In the Proceedings of *International Conference on Computational Linguistics*, pp. 172-176, Kyoto, Japan.

Schmid, H. and Laws, F. 2008. Estimation of conditional Probabilities with Decision Trees and an Application to Fine-Grained POS tagging, *COLING 2008*, Manchester, Great Britain.

Schmidt, RL 1999. Urdu: an essential grammar, London: Routledge.