Part-Of-Speech Tagging Of Urdu in Limited Resources Scenario

Ms. M. Humera Khanam
Dept. of Computer Science and Engineering
S.V. University
Tirupati, A.P, India.
humera_svec@yahoo.co.in

Prof. K. V. Madhu Murthy
Dept. of Computer Science and Engineering
S. V. University
Tirupati, A. P, India.
kvmmurthy@gmail.com

***Abstract*—** we address the problem of Part-of-Speech (POS) tagging of Urdu. POS tagging is the process of assigning a part-of-speech or lexical class marker to each word in the given text. Tagging for natural languages is similar to tokenization and lexical analysis for computer languages, except that we encounter ambiguities which are to be resolved. It plays a fundamental role in various Natural Language Processing (NLP) applications such as word sense disambiguation, parsing, name entity recognition and chunking. POS tagging, particularly plays very important role in processing free-word-order languages because such languages have relatively complex morphological structure. Urdu is a morphologically rich language. Forms of the verb, as well as case, gender, and number are expressed by the morphology. It shares its morphology, phonology and grammatical structures with Hindi. It shares its vocabulary with Arabic, Persian, Sanskrit, Turkish and Pashto languages. Urdu is written using the Perso-Arabic script. POS tagging of Urdu is a necessary component for most NLP applications of Urdu. Development of an Urdu POS tagger will influence several pipelined modules of natural language understanding system, including machine translation; partial parsing and word sense disambiguation. Our objective is to develop a robust POS tagger for Urdu. We have worked on the automatic annotation of part-of-speech for Urdu. We have defined a tag-set for Urdu. We manually annotated a corpus of 10,000 sentences. We have used different machine learning methods, namely Hidden Markov Model (HMM), Maximum Entropy Model (ME) and Conditional Random Field (CRF). Further, to deal with a small-annotated corpus, we explored the use of semi-supervised learning by using an additional un-annotated corpus. We also explored the use of a dictionary to provide to us all possible POS labeling for a given word. Since Urdu is morphologically productive. Hence we augmented Hidden Markov Model, Maximum Entropy Model and Conditional Random Field with morphological features, word suffixes and POS categories of words to develop robust POS tagger for Urdu in the limited resources scenario.

***Keywords* -** POS Tagging; Urdu; Stochastic approach; Limited Resources Scenario

*****

I. INTRODUCTION

Part-of-speech tagging consists of three main steps, i.e. tokenization, assigning potential tags to tokens, disambiguation by choosing a single most appropriate tag for each token. The words are separated utilizing white space as word boundary in the first step. One or more tags are allocated to each word based on morphological information extracted in the second step. In the third step, ambiguities are resolved and a unique tag is allocated to each word. Several methods are used for this purpose. Figure 1 describes the most common methodologies for part-of-speech tagging.

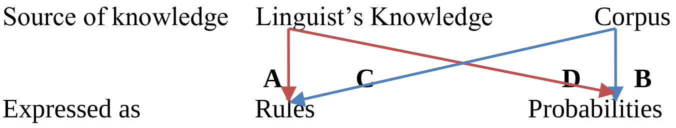


Figure 1: Methodologies for part-of-speech tagging.

The expertise of a linguist is used to formulate rules in method A [1]. Supervised machine learning techniques are used to estimate tag probabilities utilizing annotated corpus in method B [2]. Method C utilizes machine learning techniques to extract contextual information from the annotated corpus and to define the rules to disambiguate the tags. Recent work on this type of technique is done by Eric Brill [3]. No work has been found under method D. This may be due to the reason that the viewpoint of different experts may not be uniform regarding tag probabilities.

II. RELATED WORK

One of the earliest works on rule based part-of-speech tagging was reported by Klein and Simmons [1]. Their program, Computational Grammar Coder (CGC), tags the word using the lexicon and the suffix information. Voutilainen and Karlsson [4] used a Constraint Grammar (CG) approach. They designed an English tagger based on the Constraint Grammar (CG) architecture ENGTWOL (English TWO Level tagger). This was based on early rule based systems of two stage architecture, although both lexicon and rules were more complicated.

Work on statistical part-of-speech tagging started in late 1970's. Some initial work was done by Bahl, Mercer and Debili [5]. However, significant work on POS tagging started after Garside, Leech and Beale [6] provided the probabilistic formulation of the disambiguation problem. Church and Kempe [2] used second order Markov Models for disambiguation. Training of their system was done by using large hand-tagged corpora. They were able to tag 96% of words correctly using this method. Limited size of annotated training data results in a sparse data problem where in some tag sequence probabilities may be close to zeros. Sparse data problem makes certain types of computations infeasible. Jelinek and Cutting et al. [7] overcame the sparse data problem by training their taggers on large un-tagged data sets using Baum-Welch algorithm. The results provided by them were comparable with those of Church and Kempe.

Brill outlined the advantages of transformation-based approach over rule-based and stochastic approaches. According to Brill, in rule-based approaches, it is difficult to construct rules and in probabilistic approaches, much space is required to store the tables of frequencies. Transformation-based approach overcomes these issues by providing for automatic extraction of rules. The space required to store these rules is less than that for storing the probabilistic information. Another advantage described by Brill was that it was easy to use with different tag sets.

A. Brief survey of work done on part-of-speech tagging of Indian languages:

Smriti [8] described a technique for morphological based part-of-speech tagging of Hindi in a limited resources scenario. Her system used decision tree based learning algorithms to handle word sense ambiguities and unknown words. A manually annotated corpus and a set of 23 tags were used in the experiment. An accuracy of 93.5% was reported when 4-fold cross validation was used on modest-sized corpora (around 16,000 words). Another reasonably accurate part-of-speech tagger for Hindi was developed by Dalal [9], using Maximum Entropy Markov Model. He used linguistic suffix and POS categories of a word along with other contextual features. He also used the same set of tags as in [8] and an annotated corpus for training the system. An average per-word tagging accuracy of 94.4% and sentence accuracy of 35.2% were reported when 4-fold cross validation was used.

Shrivastav et al. [10] presented a Conditional Random Field **(**CRF) based statistical tagger for Hindi. They used different features such as lexical features and spelling features to generate the model parameters. They experimented on a corpus of around 12,000 tokens and annotated with a tag set of size 23. The reported accuracy was 88.95% with 4-fold cross validation.

Sanchez Leon and Nieto Serrano [11] suggested that a potentially free-word-order language could lead to greater ambiguity, i.e. it becomes harder to guess the tag of a word on the basis of its context. Dandapat et al. [12] implemented a HMM based tagger for Bengali, which is a free-word-order language, and reported an accuracy of 89%.

B. Overview of Urdu language, work done on part-of-speech tagging of Urdu Language

Urdu is one of the official languages of India and the national language of Pakistan. It is an Indo-Aryan language and belongs to Indo-European family of languages. It has about 104 million speakers, including those who speak it as a second language. Other than Pakistan and India, the majority of its speakers live in UAE, USA and UK. 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. Urdu shares its morphology, phonology and grammatical structures with Hindi. It shares its vocabulary with Arabic, Persian, Sanskrit, Turkish and Pashto languages. Urdu is written using the Perso-Arabic script.

Only one rule based part-of-speech tagger exits for Urdu. Hardie developed it in 2003 [13]. Its tag-set was developed using the grammar of Urdu by Schmidt with EAGLE guideline [14] for morphosyntactics annotation of corpora. It uses uni-rule disambiguator having approximately 270 written rules and 380 tags.

In 2007**,** Waqas Anwar [15] presented Hidden Markov Model (HMM) based approach to solve the part-of-speech tagging problem in Urdu. The general HMM based method did not perform well, so he combined smoothing techniques such as Laplace estimation, Lidstone estimation, Expected likelihood and Witten-Bell estimation with an Hidden Markov Model to resolve the data sparseness problem. This approach gave significant performance improvement.

In 2009, Hassan Sajjad compared four state-of-the-art probabilistic taggers i.e. TnT tagger, Tree tagger, RF tagger and SVM tool, in the context of Urdu language. A syntactic tag-set was proposed for the purpose of the experiments. 100,000 tokens of corpus were used for training the model. An accuracy of 94.15% was obtained when text data extracted from same corpus was used and an accuracy of 95.66% was obtained otherwise. It was concluded that in case of known words, SVM tool gives the best accuracy and in case of unknown words, CRF tagger gives the best results.

In 2012, Fareena Naz worked on Brill's Transformation-Based Learning (TBL) approach. This method automatically deduced rules from a training corpus and gave accuracy comparable to other statistical techniques. This POS tagger achieved an accuracy of around 84%, for a tag set size of 36, trained on Urdu corpus of 123775 tokens

The following reasons are identified from the literature, for non-availability of robust part-of-speech tagger for Urdu:

- The existing part-of-speech tagging techniques show that the development of a reasonably good part-of-speech tagger requires either developing an exhaustive set of linguistic rules or a large amount of annotated corpus.
- Designing robust taggers require availability of adequate quantity of good quality annotated text corpus (called gold data). However, a very limited amount of such data is available at present for Urdu.
- Preparing gold data is costly, as it requires high degree of linguistic expertise. It is also time consuming, as it is a manual process.
- The performance of part-of-speech taggers depends on the type and size of the tagset. However, no standard tag sets are available in Urdu.
- There is a dearth of good morphological analyzers and machine readable dictionaries, which form important tools for developing part-of-speech taggers in Urdu.
- Machine learning based techniques for design of POS taggers show good promise. However, they require large annotated corpus, which is not available in Urdu.

The primary objective is to develop a robust part-of-speech tagger for Urdu. We identify the following goals to address this broad objective.

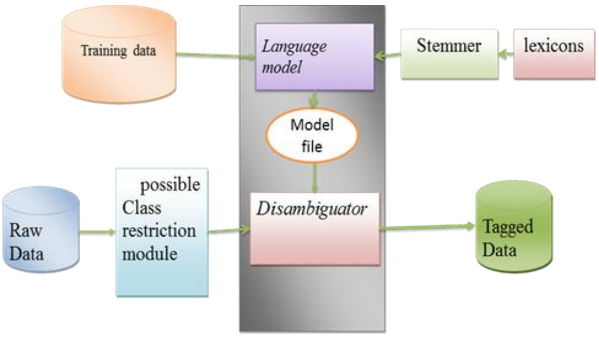
- We wish to investigate different machine learning algorithms to develop a robust part-of-speech tagger for Urdu.
- Large annotated text corpora in Urdu are not available at present, and the situation may not improve in near future. However, relatively large quantities of un-annotated Urdu text corpora are available. We wish to work with methods that effectively utilize the available resources and give good performance.
- Urdu is a morphologically rich language. We wish to

exploit the morphological features of words, as well as word suffixes to develop a robust part-of-speech tagger even with limited corpus.

- We wish to develop manually annotated Urdu corpus of at least 1000 sentences, to use in our experiments.
- Finally, we aim to do a comparative study of the performance of different POS tagging methods, in combination with different machine learning techniques.

III. SYSTEM ARCHITECTURE

The proposed architecture integrates the main components, namely Language model, Disambiguator, possible class restriction module and stemmer. Figure shows the architecture of our approach. These components are explained in detail in the following sections.



A. Language Model

We used statistical based HMM, mem and crf as a language model in our POS tagging system as in [66]. The language model is the representation of the linguistic knowledge. The knowledge may come from several resources, such as annotated text corpus, and can be encoded in various representations.

B. Disambiguator

The main function of disambiguator is to decide the best possible tag for each word in a sentence according to the language model.

C. Possible Class Restriction Module

Possible class restriction module estimates the set of possible tags {*T*}, for every word in a sentence. We use Morphological Analyzer (MA) as a possible class restriction module. This module consists of a list of lexical units associated with the list of possible tags. In this approach the assumption is made that every word can be associated with all the tags in the tagset (i.e. A set of 46 tags in the tagset {T}). Further, assume that the POS tag of a word *w* can take the values from the set $T_{MA}(w)$, where $T_{MA}(w)$ is computed by the Morphological Analyzer.

The Language model, disambiguator and possible class restriction module are related and combine them into a single tagger description. The input to the disambiguation algorithm takes the list of lexical units with the associated list of possible tags. The disambiguation module provides the output tag for each lexical unit using the encoded information from the language model.

IV. METHODOLOGY

The following methods are proposed in this paper.

A. Augmented Hidden Markov Model based part-of-speech tagging for Urdu in limited resources scenario.

We used statistical based HMM as a language model in our POS tagging system as in [15]. In HMM, the language model is represented by the model parameters μ= (π, *A*, *B*).Where π, A and B represent initial, transition and observation probabilities, respectively. We aimed to estimate the model parameters μ= (π, *A*, *B*) of the HMM using annotated text corpora. This is called supervised learning.

This approach thus assumes that we try to compute for each sentence the most probable sequence of tags s = t_1, t_2,…,t_n in set of tag sequences S, given a sequence of words w_1,…,w_n in the sentence (W):

$$\hat{s} = \arg\max_{s \in S} P(s|W)$$

It would be difficult to collect statistics for this equation directly. Instead, we rewrite it in the usual Bayesian manner as follows:

$$\hat{s} = \arg\max_{s \in S} \frac{P(W|s)P(s)}{P(W)}$$

Since we are looking for the most likely tag sequence for a sentence given a particular word sequence, the probability of the word sequence P(W) will be the same for each tag sequence and we can ignore it.

$$\hat{s} = \arg\max_{s \in S} P(W|s)\,P(s)$$

We can estimate the probability of an entire word sequence given a tag sequence by the product of the probabilities of its individual words given that tag sequence.

$$P(W|s) = \prod_{i=1}^{n} P(w_i|s)$$

$$\hat{s} = \arg\max_{s \in S} P(s)\prod_{i=1}^{n} P(w_i|s)$$

First, we make the simplifying assumption that the probability of a word is dependent only its tag:

$$\boldsymbol{P(w_i|s) = P(w_i|t_i)}$$

Next, we make the assumption that the tag history can be approximated by the most recent two tags:

$$\boldsymbol{P(s) = P(t_i|t_{i-2}t_{i-1})}$$

Thus we choose the tag sequence that maximizes:

$$\boldsymbol{P(t_1)P(t_2|t_1)\prod_{i=3}^{n} P(t_i|t_{i-2}t_{i-1})[\prod_{i=1}^{n} P(w_i|t_i)]}$$

We can use maximum likelihood estimation from relative frequencies to estimate probabilities:

$$\boldsymbol{P(t_i|t_{i-2}t_{i-1}) = \frac{Count\ (t_{i-2}t_{i-1}t_i)}{Count\ (t_{i-2}t_{i-1})}}$$

$$P(w_i \mid t_i) = \frac{Count\ (w_i,t_i)}{Count\ (t_i)}$$

This model can be smoothed to avoid zero probabilities.

HMM models do not work well when training data is limited, as it would result in large variances in the estimation of model parameters. We assume the availability of a small amount of annotated text along with a relatively large amount of un-annotated text. We proposed a method where the model parameters are initially estimated using the limited amount of annotated Urdu text. Then we used a semi-supervised learning technique where in Baum-Welch algorithm is used to re-estimate the model parameters using un-annotated text as in the case of [7]. This improves the estimation accuracy.

Our tagging system consists of two stages. We used HMM based model with bigram and trigram probabilities. Here supervised and semi-supervised methods are used to assign appropriate tags to the words. However, some words may remain untagged in the first stage either due to ambiguities or because they are unknown. These are handled in the second stage, by augmenting morphological ending based rules.

We observed that the first stage alone did not perform well due to lack of sufficient size of the corpus. However, when the morphological ending based technique was augmented, the performance improved. This could be due to the reason of high inflection and partially free-word-order features of Urdu.

In this model, each word in the test data is assigned the tag, which occurred most frequently for that word in the training data. Two taggers have been implemented based on bigram HMM model using tagset of size 46.

B. Augmented Maximum Entropy Model based part-of-speech tagging for Urdu in limited resources scenario.

We treat part-of-speech tagging as a stochastic sequence labeling task in which, given an input sequence of words W = w_1, w_2,…., w_n the task is to construct a label sequence T = t_1, t_2,…., t_n where t's belong to the set of part-of-speech tags. The label sequence T generated by the model is the one which has the highest probability among all the possible label sequences for the input word sequence W. That is

$$\boldsymbol{T = argmax\ \{P\ (T^L|W)\}}$$

Where T^L is the list of possible tag sequences. We employ a feature driven, exponential model for tagging. The underlying model is a Maximum Entropy Model (MEM). The general formulation of MEM model is given as in [16],

$$P(t|c) = \frac{1}{Z} exp \sum_{i=1}^{n} \lambda_i f_i(c,t)$$

Where Z is the normalization factor, p (t | c) is the probability of tag 't' being assigned for a context c, f_i (c, t) is a binary valued feature function on the event (c, t). A set of such feature functions is defined to capture relevant aspects of the language. The model parameters $\lambda'_i\ s$ are determined through Generalized Iterative Scaling (GIS) algorithm. The language model and the disambiguator are the heart of the system. The language model puts together all the information and generates the model. This model is then used by the disambiguator to tag the raw data.

Feature functions

A crucial aspect of feature based probabilistic modeling is to identify the appropriate characteristics of the data. We have developed a rich set of features, capturing lexical and morphological characteristics of the language. The feature set was arrived at after an exhaustive analysis of an annotated corpus. The morphological aspects of the language are addressed by features based on information retrieved from dictionary and stemmer.

Contextual features

An important problem in computational linguistics is Word Sense Disambiguation (WSD). The ambiguity can be resolved by using the context of the usage in the majority of the cases. Contextual features define the baseline system in our model. It is a tagger with just contextual features. Consider an example Urdu statement.

اج سہنے کا دام کے ے ؟
? hai kyaa daam kaa sone aaja
Is what price of gold today?

What is the price of gold today? The word [sone] can take two forms, noun (gold) and verb (sleep). The ambiguity between the two forms can be resolved only when word دام [daam] (price) is seen. We define a feature set within a context window to resolve such ambiguities.

Morphological features

Another typical problem in computational linguistics is tagging of unseen words. These are sets of words which are not observed in the training data and hence there are no context based events within the model to facilitate correct tagging. Our system uses a stemmer, which uses the dictionary to identify suffixes of a given word. These suffixes are used to generate morphological features. Consider the suffix نا [*naa*]. The following words having نا [*naa*] as suffix belong to the verb class.

تےرنا	tairanaa (swimming).
چلنا	chalanaa (walking).
گانا	gaana (singing).

The features are binary valued functions which associate a tag with various elements of the context. An example is shown below:

$$\boldsymbol{f_j(h,t) = \begin{cases} 1\ if\ suffix\ (w_i) = \text{نا}\ (na)\ and\ t = VRB \\ \quad 0\ otherwise \end{cases}}$$

Categorical features

Our approach uses the lexical properties of words in feature functions. This is achieved by collecting categorical information from the Morphological analyzer (MA). It is known that part-of speech for a word is restricted to a limited set of tags. For example, the word اچھا [*achha*] has one of the two possible POS categories, adjective (good) and adverb (well). We use this restricted set of POS categories for a word as a feature. This boosts the probability of assigning a POS tag belonging to the restricted category list as a tag for the word.

This feature is crucial for unseen words where there is no explicit bias for a word in the built model and we introduce an artificial bias with the help of limited tag set. A special case of this feature is when the restricted category list has exactly one POS tag, which implies that the word would be tagged with that particular tag.

Context window

The best context window was determined empirically. Our initial context window consisted of current word (w_i), previous two words (w_{i-1} and w_{i-2}) and the previous two tags (t_{i-1} and t_{i-2}). The best per-word-tagging accuracy of the tagger for this context window without any other feature function was 80.73%. We experimented with the context window by trying different combinations of surrounding words and their POS tags. The best tagging accuracy of 85.59% was obtained with the context window consisting of the POS tag of the previous word, the current word and the next word.

Influence of feature functions

We have designed and implemented a tagger with an appropriate context window and obtained a per-word-tagging accuracy of up to 85.59%. We call this as the baseline tagger. The addition of linguistic features boosts the performance of the baseline system. Improvement in performance with the addition of each feature function is summarized.

Handling of unknown words is an important issue in POS tagging. For words, which were not seen in the training corpus, p (t_i | w_i) is estimated based on the features of the unknown words, such as whether the word contains a particular suffix. A list of 534 suffixes has been prepared. The probability distribution of a particular suffix with respect to POS tags is calculated from all words in the training set that share the same suffix.

In this work, we have shown that contextual, morphological and lexical features of a language, when used judiciously, can deliver high performance for a morphologically rich language like Urdu. We have also discussed the exact nature of various features and their role in boosting the tagging accuracy for an MEM based tagger. Our system reached the best accuracy of 94.89% and an average accuracy of 94.38%. We have developed a stochastic tagger to which morphological and linguistic features can easily be augmented through resources like stemmer, dictionary and lexicon. Our methods have a distinct advantage over pure stochastic and rule-based linguistic systems as they provide a simple way for embedding linguistic information within a stochastic model. Rule based systems are strongly coupled with the language specific properties and the associated tagset, whereas pure stochastic systems fail to capture language specific peculiarities. Our method overcomes the shortcomings of both these approaches and can be easily extended to other morphologically rich languages, just by building resources like lexicon and stemmer, for them.

C. *Augmented Conditional Random Field based part-of-speech tagging of Urdu in limited resources scenario:*

Conditional Random Field (CRF) [17] is a probabilistic framework for labeling sequential data. A CRF is an undirected graphical model that defines a single exponential model over label sequence given the particular observation sequence. The primary advantage of the CRF over HMM is the conditional nature, resulting in the relaxation of the independence assumption required by HMM. CRF also avoids the label bias problem of the Maximum Entropy model and other directed graphical models. Thus, CRF outperform HMM and MEM models on a number of sequence labeling tasks.

CRF is used to compute the conditional probability of a state sequence Y = $y_1,..,y_T$ (tag sequence) given an observation sequence X = $x_1,..., x_T$ (word sequence) [19].

$$P(Y|X)= \frac{1}{Z_0} exp(\sum_{t=1}^{T} \sum_k \lambda_k f_k (y_{t-1,} y_{t,} x, t))$$

Where, f_k (y_{t-1}, y_t , x, t) is a feature function whose weight λ_k is to be learned via training. The values of the feature functions may range between -∞ to +∞, but typically they are binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor.

$$Z_0 = \sum_y \exp\left(\sum_{t=1}^{T} \sum_k \lambda_k f_k\left(y_{t-1}, y_t, x, t\right)\right)$$

A feature function f_k (y_{t-1}, y_t, x, t) has a value of 0 for most cases and is set to 1, when $\mathbf{y_{t-1}}$, $\mathbf{y_t}$ are exact tags and the words has solid properties.

Feature selection

Feature selection plays a crucial role in the CRF frame work. Experiments were carried out to find out the most suitable features. The main features for the part-of-speech tagging have been identified based on the different possible combinations of available word and tag contexts. The features also include prefix and suffix for all words. The use of prefix/suffix information works well for highly inflected languages. We have considered different combinations from the following set to identify the best feature set for part-of-speech tagging:

$$\mathbf{F=\{w_{i-m},\ldots.w_{i-1,}w_i,w_{i+1},\ldots,w_{i+n},t_{i-},t_i,|prefix|\leq n,|suffix|\leq n\}}$$

A set of features that have been applied to the POS tagging are as follows:

Context feature: The surrounding words of a particular word might be used as a feature.

Word suffix: Word suffix information is helpful to identify the POS information of a word. This feature can be used in two different ways. The first and the naïve one is to use a fixed length word suffix of the current and/or the surrounding word(s) as features. More helpful approach is to modify the feature as binary valued. Variable length suffixes of a word can be matched with pre-defined lists of useful suffixes for different classes. The different inflections that may occur with the noun, verb and adjective words have been considered.

Word prefix: Prefix information of a word is also helpful. A fixed length of the current and/or the surrounding word(s) might be treated as features.

Part-of-speech information: The part-of-speech tag of the previous word can be used as a feature. This is the only dynamic feature in the experiment and denoted by the bigram template feature of CRF.

We have developed a POS tagger using the statistical CRF framework that gave accuracy of 95% with the context window consisting of the current word (w_i), previous two words (w_{i-1} and w_{i-2}) and previous two tags (t_{i-1} and t_{i-2}), prefix and suffix of length up to three. The accuracy of this system has been improved significantly by augmenting several resources such as word suffix, word prefix, POS information and context features for handling the unknown words.

V. CONCLUSIONS

To summarize, we presented in this paper a few efficient models that we have developed for automatic tagging of Urdu text even when the amount of available annotated text is small. These models are shown to achieve higher accuracy than the existing models. Improvement in performance was achieved by augmenting morphological features of a word and word suffixes to the existing models. We developed manually annotated Urdu corpus of 1500 sentences, to use in our experiments. Finally, we did a comparative study of the performance of different part-of-speech tagging methods, in combination with different machine learning techniques. The best performance was achieved for CRF model combined with morphological features.

REFERENCES

[1] Klein, S and Simmons, RF (1963) A computational approach to grammatical coding of English words. In: Journal of the Association for Computing Machinery, 10: 334-347.

[2] Church 1988 and Kempe 1993 A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the second conference on Applied NLP.

[3] Brill, E Corpus-based rules. In: van Halteren (1999a).

[4] Karlsson, F, Voutilainen, A, Heikkilä, J and Anttila, A (eds.) (1995) Constraint Grammar: a language-independent system for parsing unrestricted text. Berlin: Mouton de Gruyter van Halteren, H and Voutilainen, A (1999) Automatic taggers: an introduction. In: van Halteren (1999a).

[5] Bahl, LR and Mercer, RL (1976) Part-of-speech assignment by a statistical decision algorithm. In: IEEE International Sym Part-of-Speechium on Information Theory.

[6] Garside, R (1987) The CLAWS word-tagging system.

[7] Jelinek (1985) and Cutting et al. (1992) Markov source modeling of text generation. In: Skwirzinski, JK (ed.) (1983) Impact of Processing Techniques on Communication: Proceedings of the NATO Advanced Study Institute 1983.

[8] Smriti et al. 2006 Schmid H. 1994. Probabilistic Part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. 44-49.

[9] Dalal, K. Nagaraj, U. Swant, S. Shelke and P. Bhattacharyya. 2007. Building Feature Rich part-of-speech Tagger for Morphologically Rich Languages: Experience in Hindi. In Proceedings of ICON, India.

[10] Shrivastav M., Melz R., Singh S., Gupta K. and Bhattacharyya P., 2006. Conditional Random Field Based PART-OF-SPEECH Tagger for Hindi. Proceedings MSPIL,

[11] Sanchez Leon and Nieto Serrano (1997: 163 -164), AF (1997) Retargeting a tagger. In: Garside, Leech and McEnery (1997).

[12] Dandapat, Sudeshna Sarkar, Anupam BasuDepartment of Computer Science and Engineering, Indian Institute of Technology, Kharagpur India 721302 "Automatic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario"

[13] Hardie, A., 2003.Developing a tagset for automated Part-of-speech tagging in Urdu Corpus. Proceedings of the Linguistics Conference. Department of Linguistics, Lancaster University.

[14] Leech, G and Wilson, A (1999) Standards for tagsets. In: van Halteren (1999a). (Edited version of EAGLES Recommendations for the Morphosyntactic Annotation of Corpora (1996): available on the internet at http://www.ilc.cnr.it/EAGLES96

[15] Waqas Anwar, Xuan Wang, Lu Li and Xiao-long Wang, 2007. "A Statistical Based Part-of-Speech tagger for Urdu Language", International Conference on Machine Learning and Cybernetics, Hong Kong, China.

[16] Adam L. Berger, Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, [17]. Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. of the18th ICML'01, 282-289

[17] Sha, F. and Pereira, F. 2003. Shallow Parsing with Conditional Random fields. In Proc. of NAACL-HLT, Canada, 134-141.

[18] An introduction to Condition Random Field by Charles Sutton and Andrew Mccalum