ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

Part-Of-Speech Tagging for Urdu in Scarce Resource: Mix Maximum Entropy Modelling System

M.Humera Khanam[1], K.V.Madhumurthy[2], Md.A.Khudhus[3]

Associate Professor, Dept. of CSE,SVU College of Engineering, Tirupati, India [1]

Professor, Dept. of CSE,SVU College of Engineering, Tirupati, India [2]

JE, BSNL, Tirupati, India [3]

Abstract: The area of automated Part-of-speech tagging has been developed over the last few decades by involvement from several researchers. Many new models have been introduced to improve the effectiveness of the tagger and to build the POS taggers for several languages. In this paper we develop an approach for Urdu POS tagging in scarce resource. We use Maximum Entropy (ME) modelling system [1][2], Morphological analyser(MA) [3]and stemmer[4] for automatic POS Tagging. Maximum Entropy model is a very flexible method of statistical modelling which handles the data sparse problem. Under this model, a natural combination of several features can be easily incorporated. Maximum Entropy based methods can deal with various sets has common characteristics features. We mix MA with ME model, we proposed different models ME, ME+Suf, ME+MA, ME+Suf+MA. These models are tested and results were analysed.

Keywords: Maximum Entropy Model, Morphological analyser, Stemmer, Urdu Language, NLP.

I.INTRODUCTION

Part-of-speech tagging is a process of assigning a part of speech like noun, pronoun, verb, adverb etc. automatically to each and every word in a given sentence. Input to the POS tagger is a sentence and output of the tagger is a word with specific tag. Tagging natural language is not a trivial task. Due to ambiguity it is difficult to process. POS Tagger is one of the tool which is use to resolve the ambiguity and help to process natural languages.POS tagger is useful in many natural language processing(NLP) application like machine translation, information retrieval, information extraction, word sense disambiguation, speech synthesis, and speech recognition etc.

Brief overview of Urdu language

Urdu is a derivational word from Turkish and its mean "swarm". Urdu belongs to an Indo-European language of the Indo Aryan family. Urdu is a free word order language. Urdu language resembles Hindi language. It shares its phonological, morphological and syntactic structures with Hindi. Some linguists considered them as two different dialects of one language [5]. Urdu is written in Persoarabic script and takes over 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 also making use of number of vocabulary from Turkish, Portuguese and English. Many of the Arabic words have been borrowed by Urdu language through Persian language. These words vary slightly in their tone, connotations and feeling. 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.

Related work to part-of-speech tagging

POS Tagging techniques can be classified into two major categories: Rule Based approach and Statistical based approach. The rule base techniques for designing POS consist of two stage architecture. The revolutionary researcher like Harris, Kelin, Simmons, Greene, Rubin used the same architecture[6]. The first phase of this system is to apply dictionary and to assign all possible part of speech tag to every word. The second phase employs a number of hand-crafted disambiguation rules to find out most appropriate tag for each word. Stochastic approaches to POS-tagging is not a new one since during 1980s most study of Marshall, Church, Derose, Merialdo and Brants have focused on stochastic based tagging [7]. The other notable language

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

models in Part of speech tagging are Brill transform based learning algorithm, Daelemans memory based tagging algorithm[8].

The above mentioned taggers and tagging techniques have been used for English, European and some of East Asian languages. Although South Asian languages have big community all over the world but still most of language processing research has focused on other Asian languages. In this regard Urdu language processing is specifically quite far less studied and researched; therefore quite a limited work has been carried out on Urdu language processing.

II. Our Approach with ME Model

In our approach we use Maximum Entropy Modelling system, Morphological analyser (MA) and Stemmer for automatic POS tagging of Urdu text. Construction of a Maximum Entropy Modelling system is a process of trial and error. The process mainly involves identifying a set of features which reduces the system error i.e. the identification of features which has reasonably good contribution in the classification task. Figure 1 shows the architecture of our approach, it contains mainly three components, namely Language model, Disambiguator and Possible class restriction module.

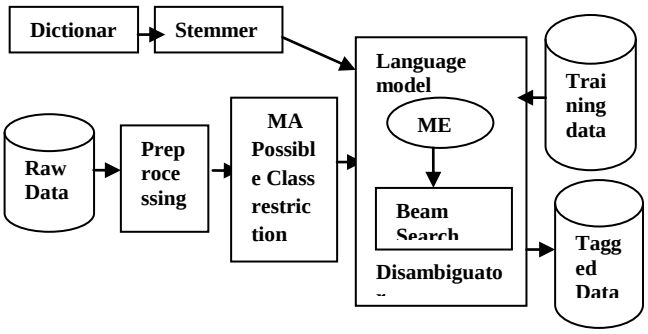

Figure 1: The ME based POS tagging architecture

Feature model

A feature based probabilistic modelling is to identify the appropriate facts about the data. We have developed a rich set of features confine lexical and morphological characteristics of the language. The feature set was arrived at after an broad 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

Primitive problem in computational linguistics is Word sense disambiguation (WSD). Majority of the cases the ambiguity can be resolved using the context of the usage. In our model, contextual features define baseline system. Baseline system is tagger with just contextual features. Consider an example Urdu statement.

آج سونے کا دام کے ہے؟
What is the price of gold today?

The word [gold] 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 encountered. To resolve such kind of ambiguities we define a feature set within a context window.

Morphological features

Another typical problem in computational linguistics is tagging of unseen words. These are set 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, a module which uses the dictionary and outputs the list of suffixes for a given word. We use the presence of suffixes as a morphological feature. An example is the suffix نا [*naa*]. Words having نا [*naa*] as suffix belong to the verb class.

For example consider the following words.

سونا sona (sleeping).
کھانا khana (eating).
گانا gaana (singing)

The features are binary valued functions which associate a tag with various elements of the context.

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

Unconditional features

Our approach extensively uses the lexical properties of words in feature functions. This is achieved by collecting absolute information from the MA. It is known that parts-of speech for a word is restricted to a limited set of tags. For example, 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. These enhance the probability of assigning a POS tag belonging to the limited category list as tag for the word. This feature is critical for unseen words where there is no explicit bias for a word in the built model and we produce 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 with very high probability.

If the above feature exists in the feature model, its corresponding parameter will contributes towards joint

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

probability. Thus in this way; the probability of a particular tag for the corresponding word will increase if there could be more than one tag for that word. Feature selection plays a crucial role in the ME framework. Experiments were carried out to identify the most suitable features for the POS tagging task. The main features for the POS tagging task have been identified based on the different possible combinations of available word and tag context. The features also include prefix and suffix for all words. The term prefix/suffix is a sequence of first/last few characters of a word, which may not necessarily be a linguistically meaningful prefix/suffix. The use of prefix and suffix information as features is found to be effective for highly inflected languages. We considered different combinations from the following set of features 'F' for identifying the best feature set for the POS tagging task.

$$\{w\,i, \mathrm{w\,i} - 1, wi - 2, wi + 1, wi + 2, ti - 1, t1 - 2, |pre| \leq 6, |suf| \leq 6\}$$

III. EXPERIMENTAL SETUP

In this section, we outline our experimental setup and discuss the effect of MA and stemmer on the system performance. We proposed total of six (ME, ME+suf, ME+MAR, ME+MAF, ME+suf+MAR, ME+suf +MAF) new models under the ME based stochastic tagging schemes. The experiments were conducted with three different sizes (3K, 5K and 7K words) of the training data to understand the relative performance of the models as we keep on increasing the size of the annotated data.

Data Used for the Experiments

Data for our experiments was taken from Department of Urdu. This data set consisted of 15,786 words of different domains, and manually tagged the data with 46 different tags. This data was spread as 3000, 5000 and 7000 words across three files. We perform three fold cross validation on this data set. Model parameters have been estimated using this data set during supervised learning. All the models have been tested on a set of randomly drawn 7000 words distinct from the training corpus.

Training the System

As revealed above, we built Mix Maximum Entropy Based Models for Part-Of-Speech Tagging in Urdu. These models were differentiated from each other by the features which were included in the model. These models use an annotated corpus. The system uses Generalized Iterative Scaling (GIS) to build the ME model, which is guaranteed to converge to a solution in this kind of problem. The procedure of training the system is summarized below.

- Describe the annotated corpus for training
- Tokenization
- Build a file of candidate features, as well as lexical features derived from the annotated corpus
- Generate an *event file* listing every feature which activates every pair <*h*,*t*> for *h* ∈*C* and *t* ∈*{T}*
- Compute the ME weightings *λi* for every *fi* using the ME toolkit with the event file as Input.

Pre-processing

System processes the data in two phases. In first phase, resources necessary for tagging the text are generated. In second phase list of suffixes for all words are generated. For every word in the corpus, dictionary stores information about the list of possible tags.

Implementation

Maxent7 package [9] for maximum entropy model has been used to implement this tagger. This package makes use of generalized iterative scaling (GIS) algorithm to estimate the model parameters. The number of iterations for GIS is configurable and we ran the algorithm for 50 iterations. During the tagging phase, beam search algorithm is employed to find the most promising tag sequence with a beam width of 5. Typical execution times on an Intel Pentium 4 machine with Linux are approximately 15 seconds for training and 3 seconds for tagging.

IV.RESULTS AND SYSTEM PERFORMANCE

We conducted experiments by taking different combinations of feature from set „F" to recognize the best suited feature set for the POS tagging task with the mix ME model.

We use three measures to evaluate the accuracy of the system, namely, Overall word tagging accuracy, Known word tagging accuracy and unknown word tagging accuracy. The Overall accuracy, known word accuracy and unknown word accuracies are shown in table 1 and figure 2.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

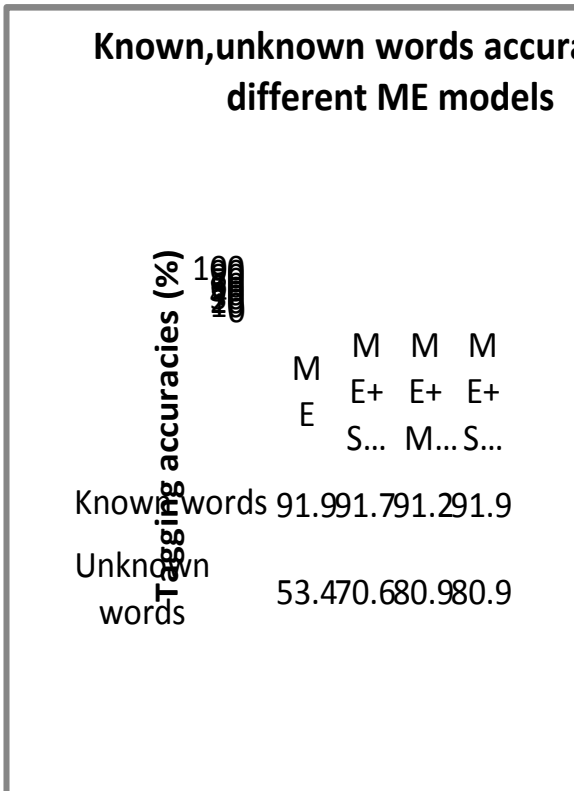

Figure2: The known and unknown word accuracy under different ME based model

It is interesting to note that the known word accuracy under the above three model are almost same when a reasonable amount of annotated data is available. But, it is clear from the figure 2 that the unknown word error rate is much lower when a morphological analyser is used to restrict the probable set of tags for a given word.

Nevertheless, the unknown word accuracy gives an improvement of 17%, 24% and 27% in case of ME+suf, ME+MAR and ME+suf+MAR models respectively over the simple ME model.

Table1summarizes the final accuracies achieved by different ME based POS tagging models with the varying size of the training data (3K,5K and 7K). Note that the baseline model (i.e., the tag probabilities depends only on the current word) has an accuracy of 78.12%.

Table1: tagging accuracies (%) of different models with 3k,5k and 7k training data. The accuracies are represented in the form of overall accuray (known word accuracy, unknown word accuracy)

Training Corpus (words)	Methods and accuracies in %			
	ME	ME+Suf	ME+M A_R	ME +Suf+M A_R
3000	76.39 (87.98,53.40)	79.40 (88.50,67.90)	84.53 (89.23,77.01)	86.75 (89.97,78.09)
5000	81.52 (90.02,54.38)	84.65 (91.02,64.68)	86.99 (91.00,79.14)	89.12 (91.65,79.02)
7000	86.58 (91.97,53.49)	89.80 (91.74,70.68)	89.40 (91.27,80.91)	90.43 (91.97,80.97)

In order to estimate the effect of using MA as feature in the ME based POS tagging model along with the features, two experiments has been conducted ME+MAF and ME+suf+MAF. The results of the experiments are shown in Table 2 using the 7K annotated training data.

Table 2: Tagging Accuracy with morphology as a feature in ME based POS tagging model

Method	Accuracy(%)
ME+MA_F	87.90
ME+Suf+MA_F	90.12

Observations

The above experiments lead us to the following observations. The use of suffix information plays a vital role, especially when the amount of training data is less. It is interesting to note that the ME+suf model gives an improvement of around 5%, 5% and 3% over the simple ME model for 3K, 5K and7K training data respectively. Another significant observation is that the use of morphological restriction (ME+MAR) gives an improvement of 10%, 7% and 5% respectively over the ME in case of 3K,5K and 7K training data. This essentially signifies that the use of morphological restriction works well in the case of small training data. As the improvement due to MA decreases with

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 9, September 2013

increasing data, it might be concluded that the use of morphological restriction may not improve the accuracy when a large amount of training data is available. The above two observations motivated us to use both suffix and MA together for all the models. From our empirical observations we found that both suffix and morphological restriction gives an improvement of 10%, 8% and 4% over the ME model respectively for the three different sizes of training data. In order to compare the ME models with the Hidden Markov Models, it has been observed that the ME models perform significantly better when the size of the training data is less and suffix information is not considered. However, the ME models achieve comparable accuracy with HMM models when suffix information and/or morphological restriction is used. Furthermore, in order to estimate the relative performance of the models, experiments were carried out using MA (ME+MAF and ME+suf+MAF) as feature in the ME model. The respective accuracies achieved by the above models are 87.90% and 90.12% for 7K word training data. The accuracy of the model is quite comparable with the accuracy achieved by the ME model when morphology is used as restriction on the choice of the possible POS tags.

Assessment of Error Types

Due to part-of-speech ambiguity, errors are produced by ME model. Ambiguity mainly affects the assignment of correct part-of-speech to every word in a sentence. For example, the word کھانا/*khana*)" can be either a *noun* or *verb*; the word سہنا/*sona*)" can be either a *finite-verb* or a noun. It has been observed from the corpora that the word ,,شریف/shareef/)" is more likely to be a noun compare to an adjective. Similarly, the word ے/*hi*)" is more likely to be a verb compares to post-position. The above observation probably fails to classify all occurrences شریف /shareef)" as an adjective and ,,ے /*hi*)/)" as post-position. Table 3 shows the top 8 confusion classes of the ME+Suf+MAR model. First column gives the actual class with their frequency of occurrence in the test data, second column gives the predicted class corresponds to the actual class, third column gives the percentage of total error and fourth column gives the percentage of error of for the particular class.

Table 3: Eight most common type of errors.

Actual Class (frequency)	Predicted Class	% of total errors	% of class errors
NN	ADJ	2.75	0.65
NN	PN	3.16	5.12
NN	ADV	2.67	0.01
NN	VB	0.93	0.53
VB	TA	0.31	0.29
KER	P	0.31	0.13
ADV	ADJ	0.75	1.02
PD	PP	0.15	0.34

VI.CONCLUSION

We have described a Mix Maximum Entropy based approach for automatic POS tagging of Urdu text in limited resource scenario. The models described here are very effective for automatic tagging even when the amount of available annotated data is small. The best performance is achieved for the ME model along with suffix information and morphological restriction on the possible grammatical categories of a word. Although simple ME based tagger performs reasonably better compare to the simple Hidden Markov Model.

REFERENCES

[1]. Adwait Ratnaparakhi 1999: A Maximum Entropy Markov Model for POS-Tagging

[2].Adwait Ratnaparakhi. 1997. A simple introduction to maximum entropy models for natural language processing. Technical report 97-08, Institute for Research in Cognitive Science, University of Pennsylvania.

[3].M.Humayoun, H.Hammarström, and A.Ranta. Urdu Morphology, Orthography and Lexicon Extraction. CAASL-2: The Second Workshop on Computational Approaches to Arabic Script-based Languages, July 21-22, 2007, LSA 2007 Linguistic Institute, Stanford University. 2007. (Design decision of version 1.2 are based on this paper) (pdf) (presentation)

[4]. Urdu stemmer Assas-band - Center for Language Engineering www.cle.org.pk/software/langproc/UrduStemmer.htm -

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

[6] . 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.

[7] .Bahl, LR and Mercer, RL (1976) Part of speech assignment by a statistical decision algorithm. In: IEEE International Symposium on Information Theory, 88- 89. Ronneby

[8].Daelemans, W (1999) Machine learning approaches. In: van Halteren (1999a).

[9]. Maxent package for Maximum Entropy Markov models at http://maxent.sourceforge.net/