A Maximum Entropy Approach to Kannada Part Of Speech Tagging

Shambhavi.B. R
Department of CSE,
R V College of Engineering,
Bangalore

Ramakanth Kumar P
Department of ISE,
R V College of Engineering,
Bangalore

Revanth G
Department of CSE,
R V College of Engineering,
Bangalore

ABSTRACT
Part Of Speech (POS) tagging is the most important pre-processing step in almost all Natural Language Processing (NLP) applications. It is defined as the process of classifying each word in a text with its appropriate part of speech. In this paper, the probabilistic classifier technique of Maximum Entropy model is experimented for the tagging of Kannada sentences. Kannada language is agglutinative, morphologically very rich but resource poor. Hence 51267 words from EMILLE corpus were manually tagged and used as training data. The tagset included 25 tags as defined for Indian languages. The best suited feature set for the language was finalised after rigorous experiments. Data size of 2892 word forms was downloaded from Kannada websites for testing. Accuracy of 81.6% was obtained in the experiments which prove that Maximum Entropy is well suited for Kannada language.

General Terms
Artificial Intelligence, Natural Language Processing

Keywords
Natural Language Processing, Part Of Speech tagging, Maximum Entropy

1. INTRODUCTION
Part Of Speech (POS) tagging in the strategy of labeling each word in an unseen text with its appropriate syntactic category also called Part of Speech. This task forms a very crucial preprocessing component of almost all Natural Language Processing (NLP) applications including speech synthesis, speech recognition, information retrieval and machine translation. POS taggers are developed either with expert selected features, statistically extracted features or a combination of both. Features selected by experts result in rule based taggers which require extensive use of human domain knowledge, while statistical feature extractions result in a stochastic model. Stochastic models are represented as Hidden Markov models [1], Maximum Entropy model [2], Support Vector Machines [3], Conditional Random Fields [4] and neural networks [5].

Kannada is one of the major Dravidian Languages of India spoken by around 45 million speakers. But research in Computational linguistics of Kannada is in its infancy and availability of Kannada language tools is still limited. The most probable reasons being lack of annotated text corpora and absence of a well defined tagset that contains all the information. In this paper Maximum entropy approach as applied to POS tagging of Kannada sentences is presented. For all evaluations, data collected randomly from EMILLE corpus have been used. Manually tagged 51,269 tokens are used and the tagset includes 25 different POS tags. The POS tagger produced an accuracy of 81.69% for a test set of 2892 word forms.

The paper is organized as follows. Section 2 discusses previous similar attempts in the NLP field. Section 3 briefs the Maximum Entropy Model (MEM). Next, development of training corpus is discussed. Section 5 and 6 explain the feature set design and the architecture of the system respectively. Later the experimental results are given. The final section concludes the paper along with future work.

2 PREVIOUS ATTEMPTS
Machine learning based taggers give more than 95% accuracy for English and many European Languages. The main reason for this achievement is the availability of large annotated corpus. But for Indian languages, the scenario is not the same. Limited availability of annotated corpora and morphological richness of Indian languages hinder the usage of machine learning techniques. MEM or Conditional Random Field based model only can deal with such limited labeled data and can incorporate diverse features.

MEM was first developed to an Indian language POS tagging by Aniket Dalal et al [6]. In this system, the main POS tagging features used were context based features, dictionary features, word features, and corpus-based features. In 2007, as part of the SPSAL workshop in IJCAI-07, IIIT, Hyderabad conducted a competition on POS tagging and chunking for South Asian languages of Hindi, Bengali and Telugu. All eight participants tried wide range of machine learning techniques. The best POS tagging accuracy for Hindi [7], Bengali [8] and Telugu [7] were 78.66 %, 77.61 % and 77.37 % respectively. Sandipan [8] experimented MEM for the three languages obtaining an average accuracy of 75.92%. ME approach to Bengali POS tagging with an enhanced feature set is reported in [9].

A POS tagger for Kannada based on Support Vector machines is reported by Antony et al [10]. An accuracy of 86% is obtained in this work. Very recently Shiva et al [11] have developed a set of cross language POS taggers for Kannada with Telugu resources. The various models built are based on HMM model and show results comparable to existing mono-lingual POS taggers. A Kannada POS tagger is proposed in [12] by Vijaylaxmi F. Patil and Shahid Mushtaq Bhat. With this tool, 10,000 Kannada words from Aesthetics domain are tagged according to BIS standards. The tagset adopted includes 11 categories and 35 sub categories.

3. MAXIMUM ENTROPY MODEL

One of the popularly used probabilistic methods for POS tagging task in Maximum Entropy model [2]. One of its advantages as compared to other models is that it potentially tags words which have never been seen in the training data. The framework estimates probabilities based on the constraints derived from the training data and making least number of assumptions possible. The ME model computes conditional probability $P(x|y)$ of a tag x Є T given some history y Є H, where T is the set of allowable tags and H is the history of observation and tag sequences. The feature function $f_j(x,y)$ is binary valued which associates a POS tag with various elements of the context. The model is given by

$$P(x|y) = \frac{\prod_{j=1}^{k} \alpha_j^{f_j(x,y)}}{\mathrm{Z(y)}}$$

Where the normalizing factor Z(y) is given by

$$\mathrm{Z(y)} = \sum_x \prod_{j=1}^{k} \alpha_j^{f_j(x,y)}$$

So given the training data and the set of features, the model is developed in which every feature f_j is associated with a parameter α_j. Of all the various efficient algorithms existing for estimating the parameter α_j, generalized iterative scaling algorithm is widely used. ACOPOST[1] toolkit used in this work adopts this algorithm for estimating the parameter α_j and hence a non-negative improvement at each iteration is guaranteed. The efficient implementation of the MEM in the toolkit gave us more time on finding out which set of features was best suited for Kannada.

4. ANNOTATION OF KANNADA TRAINING CORPUS

Language models play a fundamental role in statistical POS taggers. Probability distribution estimation by such a tool relies on the available training data. To ensure statistical significance, large training data is needed. But even with huge data, Out-Of-Vocabulary (OOV) word problem exists. This is more applicable to a morphologically rich language like Kannada. This section first details the linguistic challenges of Kannada language. Later the task of designing the tagset and the choice of corpus is discussed.

4.1 Kannada Characteristics

Kannada is a major literary language of the Dravidian family. It has a literary tradition of about 12 centuries. It is diglossic since distinctions exist between the spoken and written forms. It also has a wide range of social, regional and stylistic variation. It is highly agglutinative with a very complex system of morphology. The different parts of speech in the language are noun, pronoun, verb, adjectives, adverbs, postpositions, conjunctions and numerals.

The various affixes which get added to noun root words are gender markers, number markers, case markers and post positional elements. Pronouns are classified as personal, demonstrative, interrogative or reflexive. Verb morphology is very complex in Kannada. Verb root is followed by causative suffix, tense suffix and PNG markers. Verb may be either finite or non finite, transitive or intransitive. Adjectives are qualifiers occurring before the noun. They are classified as inherent or derived. Two main types of numerals are cardinals and ordinals. Postpositions are similar to English prepositions. They may be associative, comparative or locative in nature. Details of Kannada grammar can be studied in [13].

In addition to this huge system of morphology and classification which makes POS tagging difficult, the context also plays a vital role in determining the part of speech of a word. It is handled by designing an efficient feature set.

4.2 Designing a tagset

Table 1. Proposed Tagset

Sl. No	TAG	Description	Example
1	NN	Noun	ಭಾಷೆ
2	NNC	Compound Noun	ಆಲದ ಮರ
3	NNP	Proper Noun	ಕರ್ನಾಟಕ
4	NNPC	Compound Proper Noun	ಮಹಾತ್ಮ ಗಾಂಧಿ
5	PRP	Pronoun	ನಾನು
6	DEM	Demonstrative	ಆ
7	VM	Verb Finite	ಬರೆದನು
8	VAUX	Auxiliary Verb	ಬರೆಯುತ್ತಾ
9	JJ	Adjective	ಸುಂದರವಾದ
10	RB	Adverb(only manner adverb)	ವೇಗವಾಗಿ
11	PSP	Postposition	ಜೋತೆ
12	CC	Conjuncts	ಮತ್ತು
13	WQ	Question Words	ಯಾರು
14	QF	Quantifiers	ಬಹಳ
15	QC	Cardinal	ಒಂದು
16	QO	Ordinal	ಒಂದನೆ
17	INTF	Intensifier	ತುಂಬಾ
18	INJ	Interjection	ಅಯ್ಯೊ
19	NEG	Negation	ಬಂದಿಲ್ಲ
20	SYM	Symbol	. , ( )
21	RDP	Reduplication	ಬೇಗ ಬೇಗ
22	UT	Quotative	ಎಂದು
23	NUM	Numbers	೪೫
24	ECH	Echo words	ಅಪ್ಪಿತಪ್ಪಿ
25	UNK	Unknown	Hello

[1] http://acopost.sourceforge.net/

Our tagset consists of 25 tags as listed in Table 1. In 2006, as part of Indian Language Machine Translation (ILMT) project a common POS tagset for Indian languages was proposed by Bharati et al[14]. Our tagset is an adoption of this work. Lesser the size of tagset better is the efficiency of machine learning. But such a coarse tagset might miss out the significant information required for further processing. Hence a balance is being tried to strike between the fineness and coarseness of the tagset. Each tag emphasizes the lexical category of the word and syntactic features to be obtained from a morphological analyzer. All tags except UNK handle Kannada part-of-speech. UNK tag is given to all words of non-Kannada origin.

4.3 EMILLE Corpus

EMILLE (Enabling Minority Language Engineering) was a collaborative project between Lancaster University, UK and CIIL, Mysore [15]. It resulted in creation of 97 million electronic words for fourteen South Asian languages. The monolingual corpora consist of approximately 2 million words encoded in Unicode for Kannada language. The multiple domains cover art, science, commerce, literature, leisure and social science. In this work, our training data is restricted to the domain of only novels and stories.

5. DESIGNING FEATURE FUNCTIONS

An ideal POS tagger should accommodate bigger set of feature patterns. This directly determines the precision of the model developed. But choosing from a large possible feature set was a difficult task. Redundant features were to be avoided. Unreliable and rarely occurring features were not to be considered. Hence a number of experiments were conducted to identify best-suited feature set for Kannada and finally the following features were incorporated in the tagger.

F= { w_{i-3},w_{i-2},w_{i-1},w_i,w_{i+1},w_{i+2},w_{i+3}, dynamic POS tags of previous 3 words, Digit Feature, Symbol Feature}

5.1 Static features

For a particular word, its previous three words and successive three words are considered. This is because the surrounding context primarily determines the POS tag of the current word.

5.2 Dynamic features

POS information of three preceding words is used as the dynamic features.

5.3 Additional features

Words consisting of digits and symbols are also included in the feature set. If the current token contains digit or symbol, then its respective feature is set to 1 otherwise, it is set to 0.

6. ARCHITECTURE OF THE SYSTEM

Figure 1 shows the architecture of the POS tagging system. Manually tagged data is given to the Maximum Entropy Trainer to produce the model file and the dictionary. This dictionary file lists the words along with their probable tags and corresponding number of occurrences in the training data. The untagged Kannada input text is downloaded from commercial websites. It is then tokenized or formatted to be input to the tagger. Tokenization includes removal of unwanted characters. The tagger in consultation with the model file and dictionary predicts the tags for the input tokens. The output tagged text is then evaluated for accuracy.

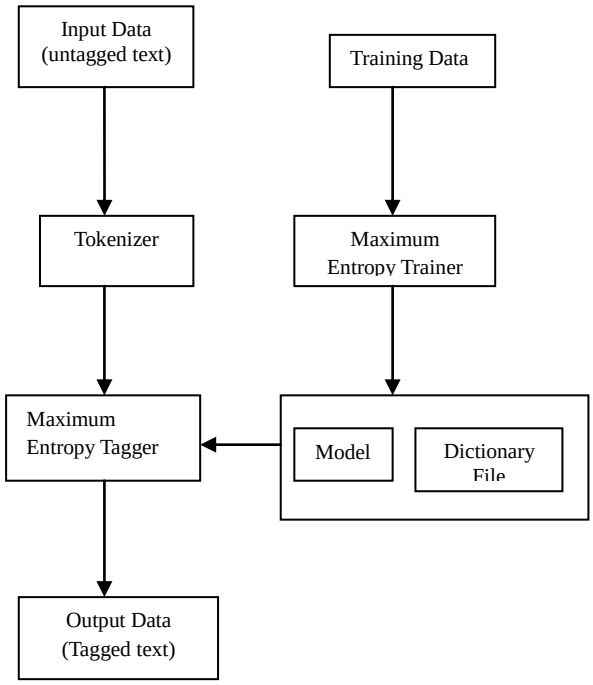

Figure 1: Architecture of POS tagging system

7. EXPERIMENTAL RESULTS

The POS tagger was trained on a corpus of 51,269 word forms. The training data were run for 100 iterations. The test data size was 2892 distinct words. Evaluation of the tagger in comparison with the baseline model is shown in Table 2. The baseline model is defined as the one where the tag probabilities depend only on the current word i.e., each word in the test data is assigned the tag which occurred most frequently for that word in the training data.

$$P(t_1, t_2 \ldots, t_n \mid w_1, w_2 \ldots, w_n) = \prod_{i=1}^{n} P(\mathrm{t_i}|w_i)$$

Table 2. MEM results

Model	Accuracy (%)
Baseline Model	53.12
MEM Model	81.69

The training data size versus accuracy is shown in Table 3. As the data size was increased the performance of the system also increased from 73.73% to 81.69%. Results with and without the dictionary were also noted. The use of dictionary initially gave improvement in performance but the increase in training data size made this improvement insignificant. But the dictionary when in effect speeded up the tagger. The values in Table 3 are shown graphically in Figure 2.

Table 3. Results with and without Dictionary

Training Data Size	Accuracy (%)	
	with dictionary	without dictionary
11761	73.73	72.66
25115	78.19	77.41
35182	79.55	78.91
51269	81.69	81.69

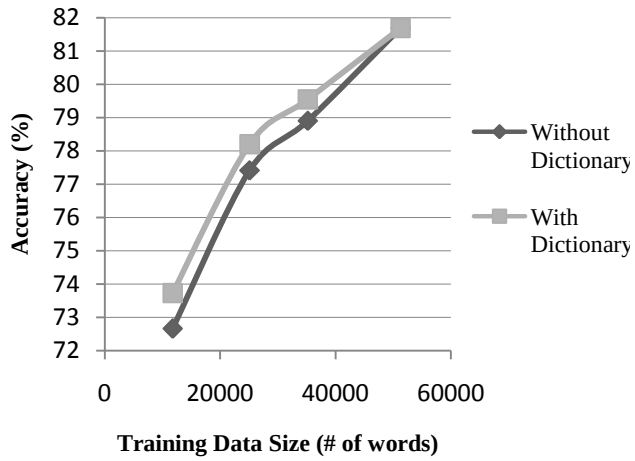

Figure 2: Performance Analysis of POS tagger

8. CONCLUSION

This is a pioneering attempt of POS tagging for Kannada language using Maximum Entropy model. Well defined tagset and quality of the corpus adopted in the language model have resulted in achieving the state of the art accuracy of 81.69%. In future it is planned to attempt other novel approaches in order to improve performance and investigate of which tagging technique is most appropriate for Kannada language.

9. ACKNOWLEDGMENTS

Our special thanks to Amogh Avadhani and Aditya M D who have contributed towards development of the tool.

10. REFERENCES

[1] D. Cutting, J. Kupiec, J. Pederson and P. Sibun. A practical part-of-speech tagger. In Proceedings of the 3rd Conference on Applied NLP, pp. 133-140. 1992.

[2] Adwait Ratnaparkhi. Maximum Entropy models for Natural Language Ambiguity Resolution, PhD thesis, University of Pennsylvania. 1998.

[3] Gimenez, J. and L. Marquez, Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. In Proceedings of the Fourth RANLP. 2003.

[4] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In the proceedings of International Conference on Machine Learning (ICML), pp. 282-289, 2001.

[5] H Schmid, Part of Speech Tagging with Neural Network. In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94) pp 172-176. 1994

[6] Aniket Dalal, Kumar Nagaraj, Uma Sawant, Sandeep Shelke. Hindi Part-of-Speech Tagging and Chunking : A Maximum Entropy Approach. In Proceedings of NLPAI Machine Learning Workshop on Artificial Intelligence. 2006.

[7] Avinesh, P., Karthik, G. Part Of Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning. In: Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages, India. pp 21–24. 2007.

[8] Sandipan Dandapat. Part Of Speech Tagging and Chunking with Maximum Entropy Model. In Proceedings of IJCAI Workshop on "Shallow Parsing for South Asian Languages", Hyderabad, India. pp 29–32. 2007.

[9] Antony P.J , Soman K.P. Kernel based Part of Speech Tagger for Kannada. In Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, , pp 2139 – 2144. 11-14 July 2010

[10] A. Ekbal, R. Haque and S. Bandyopadhyay. Maximum Entropy Based Bengali Part of Speech Tagging, Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal, Vol. (33), pp. 67-78. 2008.

[11] Siva Reddy, Serge Sharoff. Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources. In Proceedings of IJCNLP workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies. Thailand, 2011

[12] Vijayalaxmi .F. Patil and Shahid Mushtaq Bhat, Part-of-Speech Tagging for Kannada, National Seminar on POS Annotation for Indian Languages: Issues and Perspectives. Organized by Linguistic Data Consortium for Indian Languages (LDC-IL), Government of India, 12-13 Dec 2011.

[13] S.N. Sridhar, KANNADA, A Kannada grammar book, Series Editor, Bernard Comrie.

[14] A. Bharati, R. Sangal, D. M. Sharma, and L. Bai. Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages. In Technical Report (TR-LTRC-31), LTRC, IIIT-Hyderabad. 2006.

[15] Baker, P, Hardie, A, McEnery, A, Xiao, R, Bontcheva, K, Cunningham, H, Gaizauskas, R, Hamza, O, Maynard, D, Tablan, V, Ursu, C, Jayaram, BD and Leisher, M. 2004 Corpus linguistics and South Asian languages: corpus creation and tool development. Literary and Linguistic Computing 19(4):509-524.