Kannada Part-Of-Speech Tagging with Probabilistic Classifiers

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

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

ABSTRACT
Part-Of-Speech (POS) tagging is defined as the Natural Language Processing (NLP) task in which each word in a sentence is labeled with a tag indicating its appropriate part of speech. Of the entire supervised machine learning classification algorithms, second order Hidden Markov Model (HMM) and Conditional Random Fields (CRF) is chosen in this work for POS tagging of Kannada language. Training data includes 51,269 words and test data consists of around 2932 tokens. Both set being disjoint and taken from EMILLE corpus. Experiments show that the accuracy of the tools based on HMM and CRF is 79.9% and 84.58% respectively.

General Terms
Artificial Intelligence, Natural Language Processing, Machine learning.

Keywords
Natural Language Processing, Part of Speech Tagging, Hidden Markov Model, Conditional Random Fields.

1. INTRODUCTION
Part of Speech Tagger does the job of automatically assigning the most likely syntactic category for a particular use of a word in a sentence. Tagging results in an intermediate form of representation that is tractable and useful for various higher Natural Language Processing tasks like speech synthesis, information retrieval and machine translation. There is a long history of techniques dealing with the tagging process. It initially started with a rule based approach, where linguistic rules determined the tag of a particular word. Even with exhaustive rules it failed to handle unknown words. The probabilistic approach built a model based on the tagged training data. This language independent model predicted the probable tag for all known and unknown words. But its correctness depended on the size of the annotated corpus. Finally transformational based approach combined the advantage of both these techniques. Known words are tagged with the most probable tag, lexical and contextual rules determined the tag of out of vocabulary words.

All of these methodologies have been adopted for English and many other European languages with varying results. However there is very little work on POS tagging for Kannada. Kannada is the official language of Karnataka, a state in India. It is one of the top 30 most spoken languages of the world. The language though with a history of about 2000 years is in infancy with respect to its computational research. The obvious reasons being its agglutinative property and lack of language resources like a large text corpora, a comprehensive lexicon and a well-defined part of speech tagset. This led to the present work of tagging with HMM and CRF techniques. The experiments employ manually tagged data derived from EMILLE corpus and the tagset adopted includes 25 tags. The training and test data for the tagger includes 51,269 and 2932 word forms respectively. CRF based tagger outperformed the trigram HMM tagger with an accuracy of 84.58% while the latter approach yielded only 79.9%. Tools available for research purposes like TnT and CRF[1] have been used in this work. TnT is the implementation of second order HMM based on Viterbi algorithm. CRF tagger used is the Java based tool developed at IIT, Bombay. Its efficiency is credited to the implementation of sparse matrix operations and Quasi-Newton Optimization algorithm.

A brief survey about previous attempts is in section 2, followed by a discussion on idiosyncrasy of Kannada. Section 4 and 5 explain HMM and CRF model respectively. In the subsequent section, the proposed system is detailed. Evaluation results and comparisons of the two approaches are later explained. The paper concludes with future possible enhancements.

2. PREVIOUS WORKS
There has been an enormous body of work done in the research area of NLP for English studying everything from morphological analysis to sentiment analysis. Especially for POS tagging, accuracies of more than 95% have been obtained using various machine learning approaches. The most notable piece of work in this area are [1-4]. Recently taggers have been developed for Indian languages also. Automated POS taggers for Hindi, Bengali and Telugu were developed as part of NLPAI Machine Learning contest and SPSAL workshop in IJCAI-07 [5]. CRF was first applied to Hindi POS tagging and chunking by Ravindran et al [6] and Himanshu et al [7]. Performance of 89.69% and 90.89% were obtained by these taggers respectively. CRF when applied to Bengali POS tagging in [8] gave an accuracy of 90.3%. Navanath Saharia et.al [9] built an Assamese tagger using the HMM model with Viterbi algorithm. Using a small training corpus of 10,000 words, an accuracy of 87% was achieved by the tagger for the test inputs. Chirag Patel and Karthik Gali [10] have reported 92% accuracy for Gujarati POS tagging based on CRF. A HMM POS tagger in [11] was initially trained on a Bengali training set consisting of 20396 tokens. The tagger was tested on the Bengali development test set consisting of 5022 tokens and demonstrated 90.9% accuracy. In the paper [12], a stochastic approach to Malayalam POS tagging is presented. The statistical data of the trained corpus using the unigram and bigram probability along with a morphological analyzer are used to determine the parts of speech of the morphemes of the input text. Similarly CRF is compared with Support Vector Machine by Thoudam Doren Singh et al [13] as applied to Manipuri POS tagging

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

The initial reported work for Kannada POS tagging was by Antony et al [14] based on Support Vector Machines with an accuracy of 86%. A cross language Kannada POS tagger with Telugu resources was developed by Shiva et al [15]. The set of taggers built were based on HMM model. In [16], Maximum Entropy model gave an accuracy of 81.6% while tagging random Kannada text.

3. KANNADA LANGUAGE

Development of NLP tools for Kannada language is a challenging task as the language is both agglutinative and morphologically very rich. Out of 38 basic characters, 330 conjuncts are formed due to the combination of vowels and consonants. There are more than 10,000 basic root words in the language. Also about a million morphed variants are formed due to more than 5000 distinct character variants. Here are some examples depicting the idiosyncrasy of the language:

3.1 Morphological richness

Kannada words are formed by adding suffixes to the root word in a series. When suffixes attach to the root word, several morphophonemic changes take place. The order in which suffixes attach determine the morph-syntax. For example consider the words 'ನಗಲಾರೆನು' (*nagalaarenu*) and 'ಓದಿಸಿನೋಡುತ್ತಾನೆ' (*OdisinoDuttane*) which are split into meaningful parts as:

ನಗು + ಅಲ್ + ಅರ್ + ಎನು = ನಗಲಾರೆನು

nagu + al + ar + enu = nagaalarenu

ಓದು + ಇಸಿ + ನೋಡು +ಉತ್ತ್+ ಆನೆ=ಓದಿಸಿನೋಡುತ್ತಾನೆ

Odu + isi + noDu + ntt + Ane = OdisinoDuttane

3.2 Flexible word order

The basic word order in a Kannada sentence is Subject-Object-Verb (SOV). Other orders could also be found due to stylistic variations, colloquial practice, extrapostion or for other reasons. For example the English sentence 'I bet the thief with a stick' could be translated to any of the following Kannada equivalents.

ಕಳ್ಳನನ್ನು ನಾನು ಕೋಲಿನಿಂದ ಹೊಡೆದೆನು.

ನಾನು ಕಳ್ಳನನ್ನು ಕೋಲಿನಿಂದ ಹೊಡೆದೆನು.

ನಾನು ಕೋಲಿನಿಂದ ಕಳ್ಳನನ್ನು ಹೊಡೆದೆನು.

3.3 Diglossic nature

The literary variety of the language significantly differs from the spoken variety. For example, the first person singular form of the verb 'ಓದು' (*Odu* - read) in the future tense is 'ಓದುತ್ತೇನೆ ' (*Oduttene*) in the literary variety and 'ಓತ್ತೀನಿ' (*Odttini*) in the colloquial variety.

3.4 Regional Dialects

Kannada has complex regional, stylistic and social variations. The three major regional dialects are – the 'Mysore' dialect, the 'Mangalore' dialect and the 'Dharwad' dialect. However in [18], Rajapurohit has elaborated on at least 7 dialectal regions.

4. HIDDEN MARKOV MODEL

More the context taken into account, more accurate predictions are expected. Hence TnT, a Trigram POS tagger proposed in [19] is used. Given a sequence of words $w_1,w_2\ldots..w_n$ of length n, its corresponding sequence of tags $t_1,t_2,\ldots t_n$ belonging to the tagset T is calculated according to the formula:

$$\boldsymbol{argmax_{t_1\ldots tn}}\ [\ \prod_{i=1}^{n} \boldsymbol{P(w_i|t_i)P(t_i|t_{i-1},t_{i-2})}]\ \boldsymbol{*\ P\ (t_{n+1}|t_n)}$$

To handle data sparse problem, the smoothing model is the linear interpolation of unigrams, bigrams and trigrams. Tag prediction of unknown words is through a suffix trie and successive abstraction. The conditional probability of unknown words is estimated based on the statistical data available for words that end with the same sequence of letters.

5. CONDITIONAL RANDOM FIELDS

Conditional Random Fields was first proposed for segmenting and labeling sequential data by Lafferty et al [20]. Conditional Random Fields are discriminatively trained models for sequence segmentation and labeling. Results show that they out-perform the conventional HMM. Fie Sha and Fernando Periera [21] define CRF as follows. Let us assume the random variable sequences X and Y have the same length and use $x=x_1,x_2...x_n$ and $y=y_1,y_2..y_n$ for the generic input sequence and label sequence, respectively. A CRF on (X, Y) is specified by a vector f of local features and a corresponding weight vector λ. The CRF's global feature vector for input sequence *x* and label sequence *y* is given by

$$\boldsymbol{F}\ (\boldsymbol{y},\boldsymbol{x}) = \sum_{\boldsymbol{i}} \boldsymbol{f}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{i})$$

where *i* ranges over input positions. The conditional probability distribution defined by the CRF is then

$$p\lambda(Y|X) = \frac{\exp\lambda.\ F(Y,X)}{Z\lambda\ (X)}$$

where the normalization factor

$$Z\lambda\ (x) = \sum_{y} exp\lambda.F(y,x)$$

6. THE PROPOSED SYSTEM

There are important issues influencing the performance of a POS tagger. They are selecting the tagset, collecting and annotating corpora, size of corpora and corpus ambiguity. These issues are discussed in this section.

6.1 POS Tagset

Fortunately the communities of research scholars working on Indian languages have helped in designing the tagset. We will use the wealth of experience generated by them to finalize the tagset. Our tagset is an adoption of the work proposed by Bharati et al [22] as part of Indian Language Machine Translation (ILMT) project. The tagset listed in Table 1 includes 25 tags covering the different parts of speech of the language. It is designed to take advantage of the machine learning process and also facilitate further NLP processing tasks.

Table 1: List of Kannada Tagset used in the corpus

Sl. No	TAG	Description	Example
1	NN	Noun	ಭಾಷೆ/*language*
2	NNC	Compound Noun	ಆಲದ ಮರ/*banyan tree*
3	NNP	Proper Noun	ಕರ್ನಾಟಕ/*Karnataka*
4	NNPC	Compound Proper Noun	ಮಹಾತ್ಮ ಗಾಂಧಿ/ Mahatma Gandhi
5	PRP	Pronoun	ನಾನು/*I/me*
6	DEM	Demonstrative	ಆ/*That*
7	VM	Verb Finite	ಬರೆದನು/ *wrote*
8	VAUX	Auxiliary Verb	ಬರೆಯುತ್ತಾ/ *having written*
9	JJ	Adjective	ಸುಂದರವಾದ/ *beautiful*
10	RB	Adverb(only manner adverb)	ವೇಗವಾಗಿ/ *fast*
11	PSP	Postposition	ಜೋತೆ/ *along*
12	CC	Conjuncts	ಮತ್ತು/ *and*
13	WQ	Question Words	ಯಾರು / *who*
14	QF	Quantifiers	ಬಹಳ/ *more*
15	QC	Cardinal	ಒಂದು/ *one*
16	QO	Ordinal	ಒಂದನೆ / *first*
17	INTF	Intensifier	ತುಂಬಾ/ *very*
18	INJ	Interjection	ಅಯ್ಯೊ / *alas*
19	NEG	Negative verbs	ಬಂದಿಲ್ಲ/ *not come*
20	SYM	Symbol	. , ( )
21	RDP	Reduplication	ಬೇಗ ಬೇಗ /*quick quick*
22	UT	Quotative	ಎಂದು
23	NUM	Numbers	೪೫/45
24	ECH	Echo words	ಅಕ್ಕಪಕ್ಕ/ *neighbouring*
25	UNK	Unknown	Hello

6.2 Corpus Used

The main reason for tagging performance being too good for English and other European languages is the availability of large quantity of tagged data. Progress on POS tagging is very hard (if not impossible) without tagged corpora. Further absence of a good POS tagger hinders the development of other NLP tools. Hence it was decided to manually annotate a standard Kannada corpus and then build the tagger. The choice was in favor of the EMILLE (Enabling Minority Language Engineering) corpus. The corpus was the result of collaborative work of researchers at Lancaster University, UK and CIIL, Mysore [23]. It consists of around 2 million Kannada words, from different domains like science, art, leisure, literature and commerce. In our work, only novels and stories are taken for the training and testing phase.

6.3 System Architecture

The architecture of the system is as given in Figure 1. The annotator, considering the lexical and semantic rules of the language would tag the Kannada text. This annotated corpus would form the training data and is input to the trainer. The trainer builds the model and the iteration statistics file, sometimes referred to as the internal dictionary. The test data is given to the statistical tagger for automated tagging. The output of this module is compared with the gold standard by the tester module to analyze the efficiency of the system. Automatically tagged data is also displayed on the GUI, where a language expert or linguist can verify the results and correct the wrongly tagged words. The verified data can then be merged with the training data. This feature of the developed system helps to incrementally increase the training data size with reduced time and effort. The efficiency of manual tagging and speed of automatic tagging is achieved in the process.

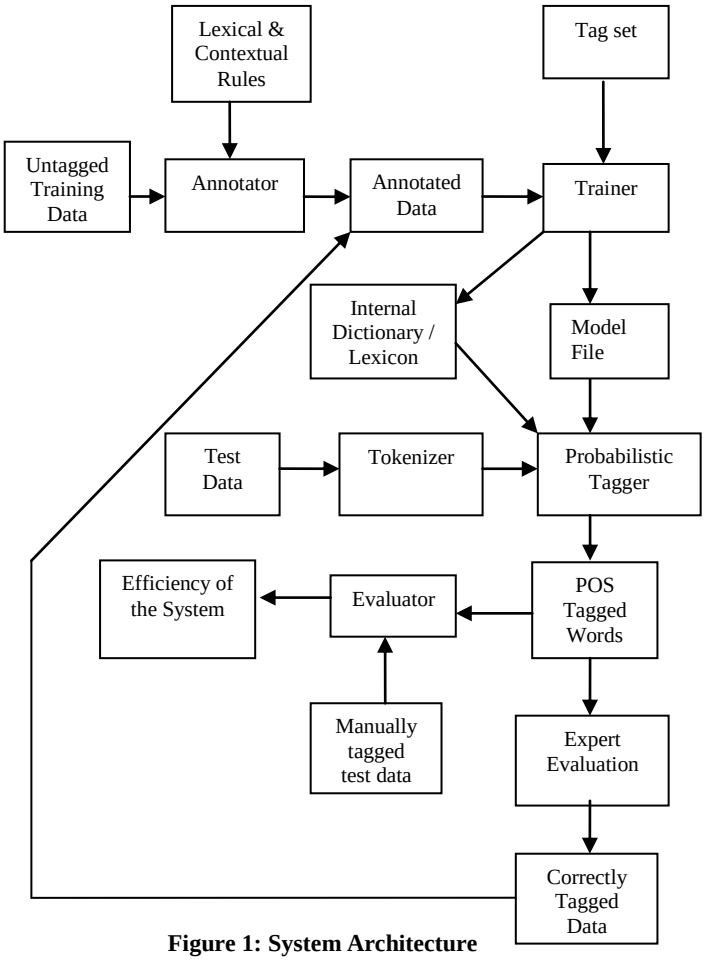

Figure 1: System Architecture

7. SYSTEM EVALUATION

The taggers' performance was evaluated with various parameters. The corpus was partitioned into about 95% training data (51269 tokens) and around 5% test data (2932 tokens). This was to ensure that test data included unseen tokens. Accuracies for known tokens and unknown tokens along with the overall accuracy were calculated. Comparisons of the taggers with respect to these factors are listed in Table 2. Experiments showed that the CRF handled unknown words better than the HMM tagger. CRF tagger gave an accuracy of

61.61% for unknown test data, while it was only 54.48% for the HMM tagger. Tagger with a better unknown word handling process is preferred in the current poor resource scenario.

The learning curve for the two taggers is depicted in Figure 2 and 3. The size of the training data was incremented in steps of around 5000 words, with the test data size also increasing gradually. Increase in accuracies with varying training size was evident in both the taggers.

Better performance of CRF tagger in comparison with the HMM tagger may be credited to its technique of using feature function unlike the latter which uses only local features. The power of CRF lies in its diverse and overlapping set of features.

Table 2: Accuracy Results for the taggers

		HMM Tagger		CRF Tagger	
		Token Count	Percentage	Token Count	Percentage
Known Tokens	Correctly Tagged	2106/2497	84.34%	2212/2497	88.59%
	Incorrectly Tagged	391/2497	15.66%	285/2497	11.41%
Unknown Tokens	Correctly Tagged	237/435	54.48%	268/435	61.61%
	Incorrectly Tagged	198/435	45.52%	167/435	38.39%
Overall Count	Correctly Tagged	2343/2932	79.91%	3480/2932	84.58%
	Incorrectly Tagged	589/2932	20.09%	452/2932	15.42%

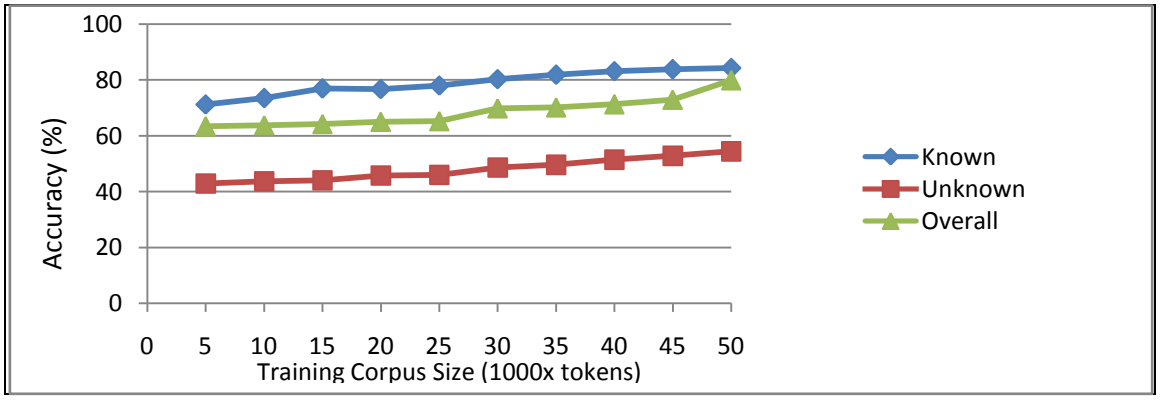

Figure 2: POS Learning curve for HMM tagger

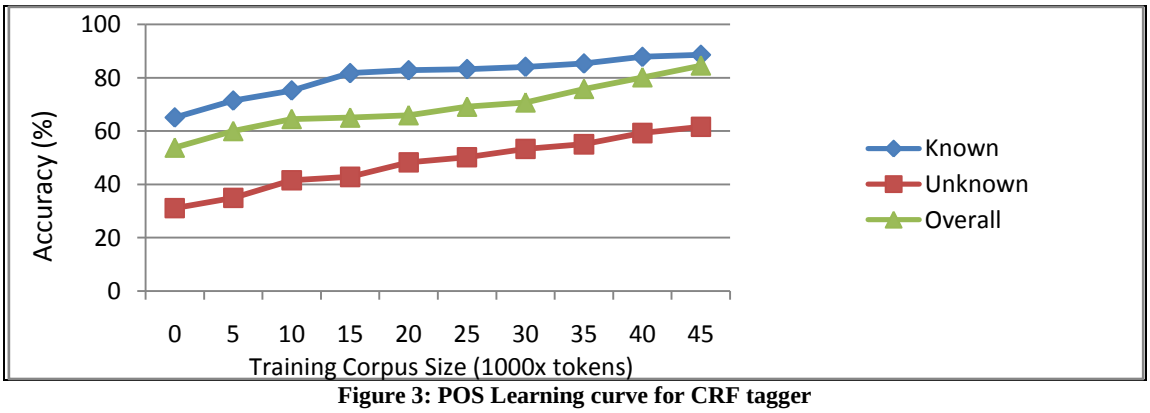

Figure 3: POS Learning curve for CRF tagger

8. CONCLUSIONS

NLP research work for a resource poor language like Kannada is very little. In this work an effort has been made to apply two supervised machine learning techniques namely, HMM and CRF for POS disambiguation task. Results are encouraging even for a annotated data size of about 54k words. CRF model outperforms the HMM model by 4.69%. Future work includes automatic POS tagging along with a morphological analyzer considering the morphological richness of the language.

9. REFERENCES

[1]Brill E. 1992 A Simple Rule-Based Part of Speech Tagger. In Proceedings of the Third Conference on Applied Computational Linguistics (ACL), Trento, Italy.

[2] Ratnaparkhi, A. 1996 A Maximum Entropy Model for Part-of Speech Tagging. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 133–142.

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

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

[5] Proceedings of IJCAI- 2007, Workshop on Shallow Parsing for South Asian Languages (SPSAL-2007), Hyderabad, India

[6] Pranjal Awasthi, Delip Rao, Balaraman Ravindran 2006 Part Of Speech Tagging and Chunking with HMM and CRF. In Proceedings of the NLPAI ML contest workshop, National Workshop on Artificial Intelligence.

[7] Himanshu Agrawal, Anirudh Mani 2006 Part Of Speech Tagging and Chunking Using Conditional Random Fields. In Proceedings of the NLPAI ML contest workshop, National Workshop on Artificial Intelligence.

[8] A. Ekbal, R. Haque and S. Bandyopadhyay 2007 Bengali Part of Speech Tagging using Conditional Random Field. In Proceedings of the 7th International Symposium on Natural Language Processing (SNLP-07), Thailand. 131-136.

[9] Navanath Saharia, Dhrubajyoti Das, Utpal Sharma, Jugal Kalita 2009 Part of Speech Tagger for Assamese Text. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore. 33–36.

[10] Chirag Patel, Karthik Gali 2008 Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages,Hyderabad, India. 117-122

[11] Ekbal, Asif, Mondal, S., and S. Bandyopadhyay 2007 POS Tagging using HMM and Rule-based Chunking. In Proceedings of SPSAL-2007, IJCAI-07, 25-28.

[12] Manju K, Soumya S, Sumam Mary Idicula 2009 Development of A Pos Tagger for Malayalam-An Experience. In Proceedings of 2009 International Conference on Advances in Recent Technologies in Communication and Computing, IEEE

[13] Thoudam Doren Singh, Sivaji Bandyopadhyay 2008 Manipuri POS Tagging using CRF and SVM: A Language Independent Approach. In Proceedings of ICON-2008: 6th International Conference on Natural Language Processing.

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

[15] Siva Reddy, Serge Sharoff. 2011 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.

[16] Shambhavi B R, RamakanthKumar P, Revanth G 2012 A Maximum Entropy Approach to Kannada Part Of Speech Tagging. International Journal of Computer Applications (IJCA), Volume 41 –No.13,9-12.

[18] Rajapurohit B B, 1982. Accoustic Characteristics of Kannada, Central Institute of Indian Languages, Mysore.

[19] T. Brants. 2000 TnT – A statistical part-of-speech tagger. In Proceedings of the 6th Applied NLP Conference, 224-231.

[20] J. Lafferty, A. McCallum, and F. Pereira 2001 Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning (ICML-2001), Williams, MA.

[21] F. Sha and F. Pereira. 2003 Shallow parsing with conditional random fields. In Proceedings of HLT-NAACL.

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

[23] 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.