An HMM Based POS Tagger for POS Tagging of Code-Mixed Indian Social Media Text

Partha Pakray(✉), Goutam Majumder, and Amarnath Pathak

Department of Computer Science and Engineering,
National Institute of Technology Mizoram, Aizawl, India
parthapakray@gmail.com, goutam.nita@gmail.com, amar4gate@gmail.com

Abstract. Text emanated from users' posts and comments on social media constitutes important piece of information for wide ranging Natural Language Processing (NLP) applications, such as Sentiment Analysis, Sarcasm Detection, Named Entity Identification, Question Answering and Information Retrieval (IR). Part–of–Speech (POS) tagging, a prerequisite for all such applications, augments tag information to the raw text. However, an inherent tendency of social media users to include multilingual contents in their posts, called code-mixing, poses challenge to POS tagging. Besides, intricate and free style writing add to the complexity of problem. To cope with the issue, a Hidden Markov Model (HMM) based supervised algorithm has been introduced for POS tagging of code–mixed Indian social media text. Publicly available social media text of Indian Languages (ILs), particularly English, Hindi, Bengali and Telugu, have been used to train and test the proposed system. Correctness of system annotated tags has been evaluated on ground of F-measure.

Keywords: POS Tagging · HMM · NLP · Code-Mixed Data
Social media

1 Introduction

Wide scale use of social media has accelerated social and digital transformation. Social media platforms are often used by different class of people for their respective concerns, ranging from personal to professional. Personal posts are chiefly characterized by individual's view and outlook whereas promotional posts embed product promotion and end users' reviews. Comprehensive analysis of end users' reviews and comments help product manufacturers in decision making and adapting their products accordingly. Besides, analysis of sentiment, embedded in posts, help demarcate positive and negative reviews. In a nutshell, contents of social media posts serve as crucial input to wide ranging NLP applications, such as Sentiment Analysis, Sarcasm Detection, Named Entity Recognition, Question Answering and Information Retrieval. However, apparent nature of algorithms,

J. K. Mandal and D. Sinha (Eds.): CSI 2017, CCIS 836, pp. 495–504, 2018.
https://doi.org/10.1007/978-981-13-1343-1_41

used in such application domains, necessitates POS tagging of text contents a priori. POS tagging augments tag information to constituent words of sentences, thus enriching their information content.

Social media platforms offer ample flexibility to their users in writing posts, comments and reviews. For example, contents of the post needn't adhere to grammatical constructs, contents may be noisy and even worse, contents may be multilingual i.e. comprising of words from different languages. Code mixing refers to inherent tendency of multilingual social media users to embed multilingual contents in their posts. Embedding often occurs at phrase, word and morpheme level. For example, a native Bengali user is likely to adulterate his English post with Bengali words. Whereas code mixing occurs at intra-sentence level, an interchangeably used perplexing term code-switching refers to mixing of different linguistic units at inter-sentence level [1,5]. Solecistic and free style writing, noisy contents and code mixed contents pose challenge to POS tagging and distinguish contents of social media from those of conventional sources. Different linguistic backgrounds of words in code-mixed sentences necessitate revamping existing POS Tagging techniques.

An exhaustively trained HMM based supervised system, described in this paper, helps cope with the issue. System exploits publicly available training and test data of NLP tools contests at ICON 2016[1]. Dataset comprises of intermixed words from English, Hindi, Bengali and Telugu languages. HMM based tagger uses class conditional probability and makes simplifying assumptions for annotating fine-grained and coarse-grained tags to the words in test dataset. Details on task and fine-grained to coarse-grained tag mapping can be found in [8,9], respectively.

Rest of the paper is organized as follows: Sect. 2 describes related works on POS Tagging; Sect. 3 describes HMM based POS Tagger system and its underlying working idea; Sect. 4 describes experimental designs, system results and result analysis; Sect. 5 concludes the paper and points directions for future research.

2 Related Works

POS tagging is a well studied problem of NLP and Computational Linguistic domains. For languages, such as English, German, Spanish, and Chinese, several POS taggers have already acquired considerably high accuracies.

A Maximum Entropy Classifier and Bidirectional Dependency Network based POS tagger acquires per–word accuracy of 97.24% [7,14]. A Support Vector Machine (SVM) based POS tagger, discussed in [11], attains accuracy of 97.16% for English on WSJ corpus.

Problems related to POS tagging of English to Spanish code–switched discourse has been reported in [13]. For this task, different heuristics based POS tag information have been combined from existing monolingual taggers. It also explores the use of different language identification methods to select POS tags

[1] http://ltrc.iiit.ac.in/icon2016/.

from the appropriate monolingual taggers. The Machine Learning approach, using features from monolingual POS tagger, attains accuracy of 93.48%. As the data has been manually transcribed from recordings, POS tagger does not incur difficulties due to code–mixing.

POS tagging for English–Hindi code mixed social media content has been reported in [15]. Efforts have been made to address issues of code–mixing, transliteration, non–standard spelling and lack of annotated data. Moreover, it is for the first time that problem of transliteration in POS tagging of code–mixed social media text has been addressed. In particular, the contributions include formalization of the problem and related challenges in processing Hindi–English code–mixed social media text, creation of annotation dataset and some initial experiments for language identification, transliteration, normalization and POS tagging of code–mixed social media text.

A language identification method for POS tagging has been developed and reported in [3]. Proposed method helps identifying language of words. Proposed method employs heuristics to form the chunks of same language. The method attains an accuracy of 79%. However, in absence of Gold language tags accuracy falls to 65%. The work reported in paper also highlights importance of language identification and transliteration in POS tagging of code–mixed social media data.

Use of distributed representation of words and log linear models for POS tagging of code–mixed Indian social media text has been reported in [12]. Furthermore, integrating pre-processing and post-processing modules with Conditional Random Field (CRF) has been found to procure reasonable accuracy of 75.22% in POS tagging of Bengali–English mixed data [4]. A supervised CRF using rich linguistic features for POS tagging of code–mixed Indian social media text finds mention in [6].

3 System Description

In this work, a supervised *bigram* Hidden Markov Model (HMM) has been implemented to identify the POS of code–mixed Indian Social Media Text. HMM based POS tagger uses two key simplifying assumptions for reducing computational complexity. Working principle of supervised algorithms, HMM based POS tagging and simplifying assumptions have been discussed and detailed in following subsections.

3.1 Working Principle of Supervised Algorithms

A supervised POS Tagging algorithm uses labeled training data of the form $(x^{(1)}, y^{(1)}) \cdots (x^{(m)}, y^{(m)})$, where $x^{(i)}$ refers to an input word and $y^{(i)}$ refers to corresponding POS label. Ultimate objective of training is to learn the optimal hypotheses, $f : \mathcal{X} \rightarrow \mathcal{Y}$, which will correctly map a previously unseen word, x, to its corresponding tag, f(x). Shorthand notations, $\mathcal{X}$ and $\mathcal{Y}$, refer to set of input words and set of corresponding POS labels, respectively.

Equation 1 signifies that given x to be a word of code-mixed sentence, objective of our learning algorithm is to find the tag, y $\in Y$, for which $P(y|x)$ is maximum.

$$f(x) = arg \max_{y \in Y} P(y|x) \tag{1}$$

Thus, the trained model outputs most probable tag y for the given word x.

3.2 HMM Based POS Tagging and Simplifying Assumptions

Joint probability distribution, $P(x, y)$, is referred to as Generative model. Equations 2 and 3 express $P(x, y)$ in term of class conditional probabilities $P(x|y)$ and $P(y|x)$, respectively.

$$P(x, y) = P(y)P(x|y) \tag{2}$$

$$P(x, y) = P(x)P(y|x) \tag{3}$$

Using Eqs. 2 and 3, $P(y|x)$ can be re-written as Eq. 4.

$$P(y|x) = [P(y)P(x|y)]/[P(x)] \tag{4}$$

However, if we are interested in finding optimal y, expressed as $\hat{y}$, static denominator of Eq. 4 can be ignored (see Eq. 5).

$$\hat{y} = arg \max_{y} P(y|x) = arg \max_{y} P(y)P(x|y) \tag{5}$$

Thus Eq. 5 expresses our objective function, given in Eq. 1, in terms of class conditional probability $P(x|y)$ and apriori probability $P(y)$.
Consider following notations:

1. w^n: Word sequence of length n.
2. t^n: Tag sequence of length n.
3. $\hat{t^n}$: Optimal tag sequence of length n.

Given w^n, our objective is to find the optimal tag sequence $\hat{t^n}$. Using Eq. 5, $\hat{t^n}$ can be written as:

$$\hat{t^n} = arg \max_{t^n} P(t^n|w^n) = arg \max_{t^n} P(t^n)P(w^n|t^n) \tag{6}$$

Probability values $P(t^n)$ and $P(t^n|w^n)$ are referred to as prior probability and likelihood probability, respectively (See Eq. 7).

$$\hat{t^n} = arg \max_{t^n} \underbrace{P(t^n)}_{prior} \underbrace{P(w^n|t^n)}_{likelihood} \tag{7}$$

Let tags $t_1, t_2, ..., t_n$ (denoted by shorthand notation t_{1-n}) constitute tag sequence t^n and words $w_1, w_2, ..., w_n$ (denoted by shorthand notation w_{1-n})

constitute word sequence w^n. Using these notations, $P(t^n)$ and $P(w^n|t^n)$ can be re-written as Eqs. 8 and 9 respectively.

$$P(t^n) = P(t_1)P(t_2|t_1)P(t_3|t_{1-2})P(t_4|t_{1-3})...P(t_n|t_{1-\{n-1\}}) \tag{8}$$

$$P(w^n|t^n) = P(w_1|t_1)P(w_2|w_1,t_{1-2})P(w_3|w_{1-2},t_{1-3})..P(w_n|w_{1-\{n-1\}},t_{1-n}) \tag{9}$$

HMM based POS taggers make following two assumptions to simplify Eqs. 8 and 9:

1. The probability of a tag appearing is dependent only on the previous tag and independent of other tags in tag sequence also known as bi–gram assumption.
2. The probability of word appearing depends only on its own POS tag and independent of other POS tags and words.

Using first assumption, Eq. 8 can be simplified and re-written as Eq. 10.

$$P(t^n) = P(t_1)P(t_2|t_1)P(t_3|t_2)P(t_4|t_3)...P(t_n|t_{n-1}) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}) \tag{10}$$

Using second assumption, Eq. 9 can be simplified and re-written as Eq. 11.

$$P(w^n|t^n) = P(w_1|t_1)P(w_2|t_2)P(w_3|t_3)...P(w_n|t_n) = \prod_{i=1}^{n} P(w_i|t_i) \tag{11}$$

Equation 12, which is used by HMM based POS Tagger to estimate the most probable tag sequence, is obtained by plugging simplified Eqs. 10 and 11 into Eq. 6.

$$\hat{t^n} = arg\max_{t^n}\ P(t^n|w^n) \approx arg\max_{t^n} \prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i) \tag{12}$$

Probability values $P(t_i|t_{i-1})$ and $P(w_i|t_i)$ in Eq. 12, referred to as tag transition probability and word emission probability, are computed from the labeled training corpus.

For example, tag transition probability $P(t_i|t_{i-1})$, for the two tags t_i and t_{i-1}, can be computed by dividing count of occurrences of t_i after t_{i-1} by count of t_{i-1} (see Eq. 13).

$$P\left(t_i|t_{i-1}\right) = \frac{Count\left(t_{i-1},t_i\right)}{Count\left(t_{i-1}\right)} \tag{13}$$

Furthermore, word emission probability $P(w_i|t_i)$, for word w_i and tag t_i, is computed by dividing count of number of times word w_i has been assigned tag t_i by count of number of times tag t_i appears in the dataset (see Eq. 14).

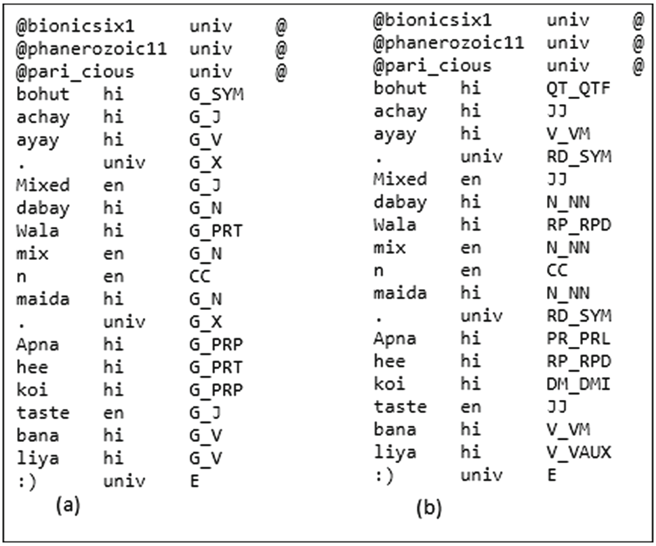
$$P(w_i|t_i) = \frac{Count(t_i,w_i)}{Count(t_i)} \tag{14}$$

4 Experiment Design and Results

4.1 Dataset Description

To train and test HMM based POS Tagger implementation, publicly available train and test data of NLP tools contest at ICON 2016 have been used. Broadly, the dataset comprises of three sets/language pairs (Bengali–Hindi (BN–EN), Hindi–English (HI–EN) and Telugu–English (TE–EN)) of code–mixed social media text of Indian Languages, collected from Facebook, Twitter and WhatsApp. For each language pair and for each source, the dataset has been further bifurcated into fine–grained and coarse–grained code–mixed data.

Figure 1 shows samples of Coarse–Grained and Fine–Grained training dataset. Details of tag sets used in training data is available in [10].



@bionicsix1	univ	@
@phanerozoic11	univ	@
@pari_cious	univ	@
bohut	hi	G_SYM
achay	hi	G_J
ayay	hi	G_V
.	univ	G_X
Mixed	en	G_J
dabay	hi	G_N
Wala	hi	G_PRT
mix	en	G_N
n	en	CC
maida	hi	G_N
.	univ	G_X
Apna	hi	G_PRP
hee	hi	G_PRT
koi	hi	G_PRP
taste	en	G_J
bana	hi	G_V
liya	hi	G_V
:)	univ	E

(a)

@bionicsix1	univ	@
@phanerozoic11	univ	@
@pari_cious	univ	@
bohut	hi	QT_QTF
achay	hi	JJ
ayay	hi	V_VM
.	univ	RD_SYM
Mixed	en	JJ
dabay	hi	N_NN
Wala	hi	RP_RPD
mix	en	N_NN
n	en	CC
maida	hi	N_NN
.	univ	RD_SYM
Apna	hi	PR_PRL
hee	hi	RP_RPD
koi	hi	DM_DMI
taste	en	JJ
bana	hi	V_VM
liya	hi	V_VAUX
:)	univ	E

(b)

Fig. 1. Sample of training dataset: (a) Coarse_Grained and (b) Fine_Grained

4.2 Code–Mixed Index (CMI) of the Dataset

For inter-corpus comparisons, level of code-mixing needs to be measured for each dataset comprising of words from different languages. Code–Mixed Index (CMI) compares non-frequent words in the dataset against total number of language dependent words [2]. CMI is computed by subtracting count of words belonging to most frequent language in the dataset (n) from total number of language dependent words (N) and dividing the result by total number of language dependent words (see Eq. 15).

$$CMI = \frac{N - n}{N} \tag{15}$$

CMI statistics of training dataset is shown in Table 1.

Table 1. CMI statistics of training dataset (BN–Bengali, HI–Hindi, TE–Telugu, EN–English, FB–Facebook, TWT–Twitter and WA–WhatsApp)

Code–Mixed language	Dataset type	FB	TWT	WA
BN–EN	Fine-Grained	0.486	0.486	0.197
	Coarse-Grained	0.230	0.267	0.002
HI–EN	Fine-Grained	0.139	0.565	0.789
	Coarse-Grained	0.641	0.216	0.113
TE–EN	Fine-Grained	0.265	0.338	0.285
	Coarse-Grained	0.372	0.265	0.255

4.3 Results

F-measure of HMM based POS Tagger for coarse-grained and fine-grained tag sets are listed in Tables 2 and 3, respectively. As seen from the two tables, for each language pair, *F-measure* of predicted coarse-grained tags are better as compared to *F-measure* of predicted fine-grained tags.

Table 2. F-measure of coarse–grained tag sets

Code–Mixed language	FB	TWT	WA
BE–EN	82.25	75.90	84.35
HI–EN	76.02	85.64	76.04
TE–EN	79.89	75.08	78.26

Table 3. F–measure of fine–grained tag sets

Code–Mixed language	FB	TWT	WA
BE–EN	76.55	72.37	81.74
HI–EN	68.81	81.05	66.11
TE–EN	72.96	72.88	72.46

4.4 Performance Comparison with Other Systems

In ICON 2016 Tool Contest on POS Tagging for Code-Mixed Indian Social Media (Facebook, Twitter and, WhatsApp) Text, a total of 13 system results were submitted for evaluation. Performances of systems were evaluated on grounds of F-measure. Ranks of our NLP-NITMZ team for each language pair and for each of the datasets have been tabulated in Table 4. Our team, using HMM based POS Tagger, ranked first in coarse–grained POS Tagging of Facebook code–mixed data of Bengali–English language pair. Team ranked second in coarse–grained POS Tagging of Twitter and WhatsApp code–mixed data of Bengali–English language pair. Team also ranked second in fine–grained POS Tagging of Facebook and Twitter code–mixed data of Bengali–English language pair.

Table 4. Rank list of NLP–NITMZ team for fine–grained (FG) and coarse–grained (CG) tag sets

Dataset	Code–Mixed language	Rank (FG)	Rank (CG)
FB	BN–EN	2	1
TWT		2	2
WA		4	2
FB	HI–EN	13	5
TWT		9	4
WA		11	5
FB	TE–EN	8	7
TWT		12	8
WA		12	7

4.5 Result Analysis

Decrease in F-measure for fine-grained dataset owes to ambiguity in tag annotation to the words in training dataset. In fine-grained training datasets, same word has been annotated differently for its different occurrences and this holds true for majority of words. For example, Hindi word "kya" has been 18 times annotated as G_PRP and 1 time annotated as PSP, out of its 19 occurrences in Hindi–English Facebook course-grained training data. In contrast, the same word has been 5 times annotated as PR_PRQ, 13 times annotated as DM_DMQ and 1 time annotated as PSP, out of its 19 occurrences in Hindi–English Facebook fine-grained training data. Ambiguity in tag annotation often reduces word–emission probability which eventually degrades F-measure.

5 Conclusion

Multilingual social media users have flooded social media platforms with code–mixed and noisy contents. Code–mixed data needs to be POS tagged for its productive utilization in NLP application domains. To cope with the challenge of POS tagging heterogeneous and noisy code–mixed data, an HMM based POS Tagger has been implemented and evaluated using code–mixed social media text of Indian Languages. Obtained system results and values of F-measure, for different language pairs and social media categories, prove worthiness of HMM based POS Tagger, particularly for coarse–grained POS tagging.

Using heuristics for reducing search space of probable tag sets, using Neural Network approach for training and testing and increasing number of instances in the training dataset are some of the notable future modifications which are likely to improve current evaluation scores.

Acknowledgement. Authors would like to acknowledge Third phase of Technical Education Quality Improvement Programme (TEQIP-III) for providing financial support. Authors are also thankful to Department of Computer Science & Engineering, National Institute of Technology Mizoram for providing infrastructural facilities and support.

References

1. Bokamba, E.G.: Code-mixing, language variation, and linguistic theory: evidence from Bantu languages. Lingua **76**(1), 21–62 (1988)
2. Gambäck, B., Das, A.: On measuring the complexity of code-mixing. In: Proceedings of the 11th International Conference on Natural Language Processing, Goa, India, pp. 1–7 (2014)
3. Gella, S., Sharma, J., Bali, K.: Query word labeling and back transliteration for Indian languages: shared task system description. FIRE Working Notes 3 (2013)
4. Ghosh, S., Ghosh, S., Das, D.: Part-of-speech tagging of code-mixed social media text. In: EMNLP 2016, p. 90 (2016)
5. Gumperz, J.J.: Discourse Strategies, vol. 1. Cambridge University Press, Cambridge (1982)
6. Gupta, D., Tripathi, S., Ekbal, A., Bhattacharyya, P.: SMPOST: parts of speech tagger for code-mixed indic social media text. arXiv preprint arXiv:1702.00167 (2017)
7. Heckerman, D., Chickering, D.M., Meek, C., Rounthwaite, R., Kadie, C.: Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res. **1**(Oct), 49–75 (2000)
8. Jamatia, A., Das, A.: Part-of-speech tagging system for Indian social media text on twitter. In: Social-India 2014, First Workshop on Language Technologies for Indian Social Media Text, at the Eleventh International Conference on Natural Language Processing (ICON-2014), pp. 21–28 (2014)
9. Jamatia, A., Das, A.: Task report: tool contest on POS tagging for codemixed Indian social media (Facebook, Twitter, and Whatsapp) Text@ icon 2016. In: The Proceeding of ICON 2016 (2016)

10. Jamatia, A., Gambäck, B., Das, A.: Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In: Association for Computational Linguistics (2015)
11. Màrquez, L., Giménez, J.: A general POS tagger generator based on support vector machines. J. Mach. Learn. Res. (2004)
12. Pimpale, P.B., Patel, R.N.: Experiments with POS tagging code-mixed Indian social media text. arXiv preprint arXiv:1610.09799 (2016)
13. Solorio, T., Liu, Y.: Part-of-speech tagging for English-Spanish code-switched text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1051–1060. Association for Computational Linguistics (2008)
14. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 173–180. Association for Computational Linguistics (2003)
15. Vyas, Y., Gella, S., Sharma, J., Bali, K., Choudhury, M.: POS tagging of English-Hindi code-mixed social media content. In: EMNLP, vol. 14, pp. 974–979 (2014)