Language Models for POS Tagging of Code Mixed Social Media Text : UB submission at ICON'16

Conference Paper · December 2016

CITATIONS
0

READS
58

2 authors, including:

Nikhil Londhe
University at Buffalo, The State University of New York
6 PUBLICATIONS **18** CITATIONS

SEE PROFILE

Language Models for POS Tagging of Code Mixed Social Media Text : UB submission at ICON'16

Nikhil Londhe
SUNY Buffalo
nikhillo@buffalo.edu

Rohini K Srihari
SUNY Buffalo
rohini@buffalo.edu

Abstract

Code switching, i.e., usage of two or more languages within the same text, is a fairly common phenomenon observed on social media within the Indian subcontinent. Thus, POS tagging of such text, specifically to facilitate automatic downstream processing is an important and challenging task given the nature of the medium and mixing. In this work, we propose two models for fine and coarse grained POS tagging for code switched social media text for three Indian languages (Bengali, Hindi and Telugu with English) on three different media (Facebook, Twitter and WhatsApp). We show how a unified model that uses language partitioning combined with simplistic language models can produce competitive results across languages and media.

1 Introduction

Multilingual speakers have a proclivity of switching between languages (Cárdenas-Claros and Isharyanti, 2009; Shafie and Nayan, 2013) during informal communication. For several language dense regions around the world, like the Indian subcontinent, this phenomenon is fairly common in online communication (like WhatsApp) and social media (Twitter and Facebook). By large, the two main variants include code switching - when users employ one writing scheme (via transliteration) but switch between languages and code mixing wherein the native writing script is retained. Some illustrative examples are listed in Table 1

Automatic processing of such data requires at the very least, the ability to perform POS tagging. However, this presents a variety of challenges like:

- **Lack of grammar and structural issues:** Typical to the nature of social media itself, posts may be extremely short in length, abundant in spelling mistakes, grammatical errors and acronyms etc. This not only drastically reduces the available context for tagging a post but often requires considerable normalization for accurate parsing.
- **Multiple languages:** Depending on the user, region and context, a given post may contain anywhere from one to four languages. Whilst most Indian languages follow the same word ordering (S-O-V), the shared vocabulary and heavy inflection makes it harder (Singh et al., 2006) to assign POS tags
- **Different switching types:** Finally, POS tagging of code switched data is fundamentally a challenging problem owing to the different types of switching types. As can be seen from Table 1, even though the attached language to a given token may be one, the sentence or the phrase follows grammatical structure of another. Thus, typical POS taggers that often use sequential tagging may be confused by such differing structures.

Having thus introduced the basic problem and the underlying challenges, the rest of this paper is organized as follows. In Section 2, we describe the actual task and details about the datasets provided. Then in Section 3, we describe our initial experiments with the training data and observations thereof. We then pro-

Sno	Post	Description
1	Aaj *match* hai kya	English noun embedded in a Hindi clause
2	*We Lost The Match But* Qandeel Balooch Se Guzarish Hai	An English phrase and a Hindi phrase joined
3	*I think* Ye Humein apna *Competitor* samajh rahi hai	Switch within a clause

Table 1: Examples of code switched social media posts and messages

vide implementation details for our submitted run in Section 4 and discuss the results in Section 5. We conclude in Section 6, by listing the scope of future work and possible system improvements.

2 Task and Dataset

The task involves POS tagging given code switched / code mixed datasets. We were provided training data split six ways : three languages (English-Hindi, English-Bengali and English-Telugu) and two levels of granularity in tags (coarse grained and fine grained). The Google Universal tagset (Petrov et al., 2011) was used to define the coarse grained tags whereas the fine grained tags were as defined in (Jamatia et al., 2015). The latter also defines a one-to-one mapping between a fine grained tag to a coarse grained tag. Further, for each token within a post, the language of the token was also provided. We further summarize the different datasets in terms of their language counts, document characteristics and tag distributions in Table 2. Further details regarding how the dataset was prepared etc. can be found in (Jamatia and Das, 2016)

In total we were allowed to submit four runs : one for each tagset (coarse vs fine) and in two modes (constrained vs unconstrained). The constrained run prohibited the use of any external resources for training except the provided data whereas the unconstrained run had no such restrictions. For our submission, we only entered a constrained run.

We now turn our attention to our initial experiments with the given training data that provided us the building blocks for our final system run.

3 Experiments

In this section, we describe our experiments and different methodologies used. Given the multitude of training data, we began by asking the following questions and present related experiments in order:

1. How do we tackle coarse grained labels - convert guessed fine grained labels to coarse labels or train separate models?
2. Does the source matter? Can models trained on one source type be applied to another? Does a combination scheme yield better results?

3.1 Tagset granularity and conversion

For the first task to measure the applicability of fine-grained models to coarse-grained data, we train two sets of models. We use the ICON 2015 training dataset[1] as a blind test set for this task. The said dataset contains fine grain annotated Twitter data in two of the three (Bengali and Hindi) languages in consideration. We assume that there are no significant differences in terms of structure, length, etc between the two Twitter datasets (this year and previous year). We thus, trained two models on the given Twitter datasets - a fine grained model and a coarse grained model. We present the test results in Table 3. We observe that the coarse grained models work better than fine grained ones and distinctly outperform the converted models.

3.2 Source and Model Generalization

In the second set of experiments, we try and validate the generality of each of the models. We thus, use one source as the training set and test it on the other two datasets. We tabulate the results in Table 4.

In order to explain the above results, we propose a simple *data similarity* model as explained below. In principle, it is similar to the CMI approach by (Gambäck and Das, 2016), but uses a simpler but coarser approach. If we represent each dataset to consist of k languages with each language i contributing to τ^i percent of the total language tags, then a dataset d can be represented as a vector as

[1]Note that this does not violate the constrained restriction as the data is not used to train but used as a blind test set

Key	# docs	Length	Avg len	en	hi	bn	te	ta	ne	univ	acro	mixed	undef	total
fb-bn	146	7609	51.12	29.70	0.54	48.65	0	0	2.89	17.03	1.17	0.01	0.01	7462
tw-bn	172	3884	21.58	26.73	0.27	48.64	0.00	0.00	3.07	19.83	0.67	0.11	0.67	3711
wa-bn	303	3833	11.65	0.20	0.00	19.72	0.00	0.00	0.40	40.04	2.15	0.00	37.49	3529
fb-hi	770	21386	26.77	64.10	13.86	0.00	0.00	0.00	3.18	17.60	1.22	0.03	0.01	20615
tw-hi	1095	18407	15.81	21.56	56.49	0.00	0.00	0.00	2.39	19.37	0.18	0.01	0.00	17311
wa-hi	762	3981	4.22	11.28	78.90	0.00	0.00	0.00	1.09	8.73	0.00	0.00	0.00	3218
fb-te	742	10780	13.53	37.21	0.00	0.00	26.36	0.00	3.91	32.09	0.39	0.01	0.03	10037
tw-te	742	12756	16.19	26.65	0.00	0.00	33.73	0.00	2.13	37.25	0.20	0.02	0.02	12013
wa-ta	492	7914	15.09	25.51	0.00	0.00	28.50	0.00	1.31	44.58	0.11	0.00	0.00	7421

Table 2: Dataset summary

Language	Fine-Fine	Fine-Coarse	Coarse-Coarse	Full / Fine (CV)
Bengali	48.66	6.89	55.30	69.47
Hindi	49.72	9.42	55.04	71.22
Telugu (CV)	59.97	xx	64.12	64.60

Table 3: Measuring tag conversion accuracy

Language	Granularity	Source : Twitter		Source : Facebook		Source: WhatsApp	
		FB	WA	TW	WA	TW	FB
bn	Fine	47.17	50.66	47.96	**50.74**	46.28	45.40
	Coarse	**65.52**	57.97	55.65	57.60	52.34	53.18
hi	Fine	60.06	**62.87**	45.43	32.94	51.05	43.15
	Coarse	66.22	**70.85**	56.10	44.47	59.31	50.12
te	Fine	58.90	**71.10**	48.18	49.49	64.64	57.36
	Coarse	54.67	**73.94**	53.77	56.16	68.41	64.56

Table 4: Source Type Dependence

$\hat{d} = \tau_d^1 \hat{l_1} + \tau_d^2 \hat{l_2} + ... + \tau_d^k \hat{l_k}$. Thus, the similarity between two datasets, d_1 and d_2 can thus be expressed as $sim(d_1, d_2) = \hat{d_1} \cdot \hat{d_2}$. We found that for Bengali, Twitter & Facebook are most similar (99.76%) whereas for Hindi & Telugu, it is Twitter & WhatsApp datasets (95.63% and 98.81% resp.) which coincides with our results in Table 4.

4 System Description

Despite our earlier conclusions on source-model dependence in Section 3.2, we ended up using a unified model for all languages given the lack of time. As we discuss further in Section 6, one of our first proposed changes is to train source specific models and use them to tag instead. We now present details of our system submission.

4.1 Language partitioning

Instead of trying to train the taggers to understand the language splits, we instead decided to train language specific models. This is in line with the work presented in (Vyas et al., 2014), where combining language specific POS taggers has proven to be very successful. We can devise a simple partitioning strategy as follows:

1. The following tags are language independent and do not participate in the language specific tagging : Punctuation, Emoticons, Mentions, URLs & HashTags.
2. They can act as rudimentary language partitioners
3. Most of the above tags can be tagged using regular expressions or fixed lists

Thus, we implemented simple regex based taggers that could accept a regular expression and on a match on a token, emit the corresponding mapped tag. Further, on encountering any one of these "special" tags, we partition the token stream into smaller sub-streams. For any given stream, we determine its underlying language by a majority voting scheme , i.e., language with the most tokens in the stream. We opine that although this may help better process longer sentences, it could be overfit on smaller ones (<= 3 tokens). We processed the given training datasets and created language specific splits (bn, en, hi and te) which we used to train models as described in the following section.

4.2 Coarse tagging : Language models

Further to the language splits obtained above, we performed two operations : training coarse taggers and language models. As discussed before, we trained simple MEMM models for the

coarse taggers. For the language models, we used a trigram model that utilized only the coarse tags supplemented with additional start and end tags. Thus, for a given sentence, the coarse model tagging performed three steps : (a) partition by language (b) enumerate all possible sequences and probabilities using the corresponding tagger(s) (c) pick the sequence with the highest probability using LMs.

4.3 Fine tagging : CRF models

Finally, for tagging fine grained tags, we used the coarse grained outputs along with a variety of features as listed below to train a CRF tagger:

1. **Word itself:** The word itself lowercased and used with a window size of 2
2. **Language:** The assigned language, used with a window size of 2 and as bigrams and trigrams
3. **Capitalization:** Binary feature indicating whether the original word began with a capital letter
4. **Suffix:** Extracted suffix from the word from the last occurring vowel, e.g., *ing* from *running*
5. **Coarse tag:** Assigned coarse grain tag, used with a window size of 2, as bigrams, trigrams and bigrams with the assigned language

Having presented our system details, we now present the results on the test dataset in the following section.

5 Results

We present a summary of the test dataset in Table 5 and the performance of our system in Table 6. It must be noted that there were several discrepancies between the training and test data. Apart from the variations in overall structure and composition, we also found incorrect tags (labeled "undef" by us). Also the training data marked named entities as "ne"s but the test data did not follow this consistently. Overall, the system shows high recall except for the Bengali WhatsApp dataset that has a high percentage of "undefined" tags. The fine models perform in line with the corresponding coarse models as expected. Beyond that, we can not ascertain any other patterns and may require tag level analysis.

6 Conclusions

Having thus presented our system description and results, we now discuss further research avenues specifically to improve system performance. Firstly, we would train source specific models to better account for source level variations. Secondly, we may get some improvements by data cleanup and standardizing the annotations on the test set. We believe our overall approach of data partitioning and language models is encouraging. As against simple symbol and special token based partitioning, we may be able to improve it by using low level tags (CCs for example) and also take into account the mixed tags as well as dominant / embedded language considerations while evaluating partitioning structures.

References

Mónica Stella Cárdenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Betweenyes, ya,andsi-a case study. *The Jalt Call Journal*, 5(3):67–78.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1850–1855.

Anupam Jamatia and Amitava Das. 2016. Task report: Tool contest on pos tagging for code-mixed indian social media (facebook, twitter, and whatsapp) text @ icon 2016.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. *RECENT ADVANCES OF NATURAL LANGUAGE PROCESSING*, page 239.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Latisha Asmaak Shafie and Surina Nayan. 2013. Languages, code-switching practice and primary functions of facebook among university students. *Studies in English Language Teaching*, 1(1):187.

Smriti Singh, Kuhoo Gupta, Manish Shrivastava, and Pushpak Bhattacharyya. 2006. Morphological richness offsets resource demand-experiences in constructing a pos tagger for

Key	# docs	Length	Avg len	en	hi	bn	te	ta	ne	univ	acro	mixed	undef	total
fb-bn	616	17478	28.37	16.36	0.45	63.69	0.00	0.00	4.63	13.27	1.37	0.23	0.00	16863
tw-bn	413	9295	22.51	23.38	0.16	55.47	0.00	0.00	1.89	17.62	0.59	0.06	0.85	8859
wa-bn	749	8900	11.89	0.00	0.08	21.47	0.00	0.00	0.32	36.81	1.94	0.00	39.40	8152
fb-hi	112	2277	4.92	22.25	50.23	0.00	0.00	0.00	4.71	21.42	1.20	0.09	0.09	2166
tw-hi	111	2272	4.89	20.72	58.14	0.00	0.00	0.00	3.52	17.62	0.00	0.00	0.00	2162
wa-hi	219	1020	21.47	28.43	56.11	0.00	0.00	0.00	1.50	13.59	0.00	0.00	0.37	802
fb-te	247	2360	10.47	46.90	0.57	0.00	23.64	0.00	3.12	25.58	0.05	0.00	0.14	2115
tw-te	251	4070	6.17	31.88	0.00	0.00	30.21	0.00	1.05	36.54	0.00	0.00	0.31	3820
wa-ta	198	3913	5.06	33.80	0.00	0.00	29.95	0.00	0.75	35.12	0.03	0.00	0.35	3716

Table 5: Test dataset summary

Dataset		Coarse			Fine		
		Precision	Recall	F-measure	Precision	Recall	F-measure
bn	fb	0.657	0.994	79.11	0.569	0.978	71.94
	tw	0.575	0.575	57.48	0.555	0.999	71.36
	wa	0.481	0.481	48.10	0.493	0.785	60.60
hi	fb	0.466	0.999	63.56	0.502	0.914	64.84
	tw	0.596	0.998	74.67	0.601	0.996	74.95
	wa	0.593	0.593	59.30	0.554	0.988	70.97
te	fb	0.600	0.994	74.88	0.559	0.981	71.19
	tw	0.523	0.991	68.47	0.588	0.988	73.68
	wa	0.616	0.985	75.82	0.654	0.972	78.23

Table 6: Test results

hindi. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 779–786. Association for Computational Linguistics.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *EMNLP*, volume 14, pages 974–979.