TASK REPORT: TOOL CONTEST ON POS TAGGING FOR CODE-MIXED INDIAN SOCIAL MEDIA (FACEBOOK, TWITTER, AND WHATSAPP) TEXT @ ICON 2016

Anupam Jamatia, Amitava Das*
National Institute of Technology, Agartala, Tripura, India
*Indian Institute of Information Technology, Sri City, Andhra Pradesh, India
anupamjamatia@gmail.com, amitava.das@iiits.in

Abstract

This overview paper presents a comprehensive report of the datasets and performances of the submitted runs at the second shared task on POS tagging code-mixed Indian social media text collocated with thirteenth International Conference on Natural Language Processing (ICON-2016). For the task data of three widely spoken Indian languages- Hindi, Bengali, and Telugu, mixed with English was released. The data was collected from a wide variety of social media - Facebook, Twitter and WhatsApp. Altogether 15 teams participated and submitted runs on constrained and unconstrained categories.

1 Introduction

The evolution of social media texts such as blogs, micro-blogs (e.g., Twitter), WhatsApp, and chats (e.g., Facebook messages) has created many new opportunities for information access and language technology, but also many new challenges, making it one of the prime present day research areas. We have observed that monolingual English and romanized Indian languages (ILs) messages are also equally prevalent in social media. English Non-English speakers, especially Indians, do not always use Unicode to write something in social media in ILs. Instead, they use phonetic typing/ roman script/ transliteration and frequently insert English words or phrases through code-mixing and anglicisms (see the following example 1.1), and often mix multiple languages to express their thoughts.

Example 1.1 *ICON 2016 Varanasi me hold hoga! Great chance to see the pracheen nagari!*

While it is clear that English still is the principal language for social media communications, there is a growing need to develop technologies for other languages, including Indian languages. India is home to several hundred languages. Language diversity and dialect changes instigate frequent code-mixing in India. Hence, Indians are multilingual by adaptation and necessity, and frequently change and mix languages in social media contexts, which poses additional difficulties for automatic Indian social media text processing. Part-of-speech (POS) tagging is an essential prerequisite for any kind of NLP applications. This year we continued the last year's (2015) POS tagging shared-task on three widely spoken Indian languages (Hindi, Bengali, and Telugu), mixed with English.

Three different kinds of social media - Facebook, Twitter and WhatsApp data have been released for the task. We believe there are significant differences among these different social medias. Facebook messages are tend to be longer and personalized, whereas tweets are fundamentally short due to the imposed character limit and typically written for broadcasting purposes, whereas WhatsApp messages are very short in length as those are typically meant to particular person/group. Possibly this is the first time NLP related issue on WhatsApp messages is being discussed. WhatsApp messages are relatively much smaller than Facebook and Twitter messages, therefore more challenging. This forum presents a platform for academic discussion among researchers and prac-

titioners dealing and working with social media text.

Altogether 15 teams participated and submitted runs on constrained and unconstrained categories. Constrained: means the participant team is only allowed to use our corpus for the training. No external resource is allowed. Unconstrained: means the participant team can use any external resource (available POS tagger, NER, Parser, and any additional data) to train their system. Accordingly they have to mention those resources explicitly in their task-report.

The rest of the report is organized as follows: In Section 2 we discuss the background and related work on part-of-speech tagging, social media text processing and code-switching. Corpus related information with CMI calculation are described in the Section 3. Run submission and results are discussed in Section 4.

2 Background & Related Work

POS tagging on Code-Mixed Indian Social Media Text is a very incipient research problem in the field of natural language processing (NLP). Indian NLP researchers are working on various issues of Code-Mixed corpora. In this digital era, now a days inescapable social media (viz. e-mails, tweets, chat, discussion forum, comments, and blogs/microblogs etc.) are part of communication and the 'netizens' are more creative and interactive to generate content using free language forms that often are closer to spoken language and hence show phenomena previously mainly analysed in speech. It is observed that grammatical norms and use of standard lexical items (Hu et al., 2013) are more often seen in the Twitter messages compare to the chats, which are more conversational (Paolillo, 1999) and hence less formal. Because of the ease of availability of Twitter, most previous research on social media text has focused on tweets; however, the conversational nature of chats tend to increase the level of code-mixing (Cárdenas-Claros and Isharyanti, 2009).

This year for the ICON-2016 POS tagging tool contest we have collected data from Twitter, Facebook posts and also from WhatApp message. Evidently, social media in itself does not constitute a particular textual domain; we use the term 'social media text' as referring to the way these texts are communicated, rather than to a specific type of text. Indeed, there is a wide spectrum of different types of texts transmitted through social media, and the common denominator of social media text is not that it is 'noisy' or informal, but that it describes language in (rapid) change (Androutsopoulos, 2011). Although social media indeed often convey more ungrammatical text than more formal writings, the relative occurrence of non-standard English syntax tends to be fairly constant across several types of social media (Baldwin et al., 2013).

However, while the first works on social media concentrated on monolingual English texts, recent years has witnessed an increased interest in the study of non-English texts and of texts in a mix of languages, as shown by the shared task on word-level language detection in code-switched text (Solorio et al., 2014) organized by the workshop on Computational Approaches to Code Switching at the 2014[1] and 2016[2] Conference on Empirical Methods in Natural Language Processing (EMNLP), and the shared tasks on information retrieval from code-mixed text held at that the 2014, 2015 and 2016 workshops of the Forum for Information Retrieval Evaluation, FIRE (Sequiera et al., 2015). Here we are in particular concerned with code-mixed social media text involving Indian languages. So though Diab and Kamboj (2011) briefly explained the process of corpus collection and suggested crowd sourcing as a good method for annotating formal (non-social media) Hindi-English code-mixed data, the first Indian code-mixing social media text corpus (Bengali-Hindi-English) was reported by Das and Gambäck (2013) in the context of language identification, while Bali et al. (2014) argued that structural and discourse linguistic analysis is required in order to fully analyse code-mixing for Indian languages. Gupta et al. (2014) discussed the phenomenon in the context of information retrieval (calling this 'mixed-script information retrieval'), applying deep learning techniques to the problem of identifying term equivalents in code-mixed text. Jamatia et al. (2015) worked on collecting and annotating code-mixed English-Hindi social media text from Twitter and Facebook messages and experimented on POS tagging.

Corpus Source	Language Pair	words	utterances (U)	switched (S)	switched (%)	C_{avg} (U)	C_{avg} (S)	P_{avg} (U)	P_{avg} (S)	δ (U)	Cc
Facebook	HI-EN	20615	772	411	53.23	16.23	30.48	1.76	3.30	17.10	60.59
	BN-EN	7462	148	148	100.00	46.72	46.72	10.91	10.91	33.78	130.06
	TE-EN	10037	744	626	84.13	45.81	54.45	3.12	3.71	26.75	115.93
Twitter	HI-EN	17311	1096	990	90.32	29.35	32.50	3.81	4.22	4.01	104.63
	BN-EN	3711	173	173	100.00	50.40	50.40	6.42	6.42	27.17	133.73
	TE-EN	12013	744	733	98.52	49.61	50.36	3.66	3.72	26.21	131.71
WhatsApp	HI-EN	3218	763	145	19.00	7.01	36.90	0.45	2.37	13.89	22.85
	BN-EN	3529	305	6	01.96	0.89	45.14	0.02	1.17	0.66	2.53
	TE-EN	7421	494	489	98.98	48.98	49.48	3.14	3.17	21.26	131.47

Table 1: CMI Statistics of Corpora

3 Data Collection

Training data for Twitter (2,013 tweets), Facebook (1,664 messages) and WhatsApp (1,562 messages) are released for all the 3 language pairs: English-Hindi, English-Bengali, and English-Telugu. Although for bi- or multilingual code-mixing is a natural practice, but what is the actual distribution of code-mixing in any social-media corpus is an important question. We have observed that monolingual English and romanized Indian languages (ILs) messages are also equally prevalent in social media. For this contest we discarded almost all the monolingual English messages, as there are other research efforts and forums, where the actual research problem with English social media has been discussed extensively. Here we will be concentrating only on code-mixed En-ILs and monolingual ILs. While two languages are blending, another important question might be raised is which language is mixing in what. To keep our data balanced we keep an equal distribution of utterances where English mixed in ILs and ILs mixed in English. Although our corpus is mostly bi-lingual mix but there are utterances with tri- quad-lingual mix. For example in the English-Bengali corpus there are significant Hindi word mix, whereas in the English-Telugu data there are significant Tamil and Hindi mix. Here we describe how we have collected data from different social media.

Facebook: We have manually identified open discussion pages like campus bill boards, technology related pages and pages related to movies. The only selection criteria was there is enough code-mixing present in the page for the particular language pair.

Twitter: Twitter API supports search, therefore we collected languages specific high frequent words from previous year corpus. Using such words we searched Twitter and then run a language identifier (Das and Gambäck, 2013) on the obtained tweets. Due to the mixed nature of the corpora we have calculated the Code-Mixing Index(CMI) of Gambäck and Das, first introduced in (Das and Gambäck, 2014; Gambäck and Das, 2014), but extended and detailed in (Gambäck and Das, 2016). Tweets with CMI 0 have been discarded. This process continues until we obtain desired number of tweets for the task. **WhatsApp:** This data is mainly selectional. Student groups have been chosen. Then language identifier and CMI based sorting has been applied here as well as described in the Twitter corpus section.

In the Table 1 shows statistics and switching complexity (Gambäck and Das, 2016) for the corpora used in this year ICON tool contest. Here we have used both at utterance level (C_u) and overall corpus level switching (C_c). Here U and S show how many utterances there are in each corpus and how many of those that contain code-switching. P gives the average number of switching points and δ inter-utterances switching.

4 Submissions and Results

In our 2016 ICON tool contest total 15 team has submitted the result for all the language pair. Among the 15 teams, 7 teams has submitted for both Constrained unconstrained.

AMRITA_CEN team (Kumar and Soman, 2016) proposed a method based on SVM classi-

[1] http://emnlp2014.org/workshops/CodeSwitch/call.html

[2] http://care4lang1.seas.gwu.edu/cs2/call.html

Team	Institute	Category (Const./ Unconst.)	BN					
			FB_FG (f1)	FB_CG (f1)	TWT_FG (f1)	TWT_CG (f1)	WA_FG (f1)	WA_CG (f1)
Amrita_CEN	Amrita Vishwa Vidyapeetham	C	74.30	80.90	66.10	74.55	76.85	81.30
		UC	81.63	**86.77**	**77.90**	71.38	**87.52**	**82.30**
BITS_PILANI_TEAM 1	BITS Pilani	C	68.40	69.90	64.30	69.61	71.97	73.60
		UC	67.30	69.10	64.70	70.70	83.18	75.70
BITS_PILANI_TEAM 2	BITS Pilani	C	74.58	77.94	72.28	71.51	73.37	76.23
		UC	74.58	77.94	72.18	71.67	75.10	76.80
BITS_PILANI_TEAM 3	BITS Pilani	C	64.51	61.10	68.70	55.26	76.05	82.21
		UC	64.51	60.60	68.70	54.91	76.05	80.63
MISIM-UB	University of Buffalo	C	71.94	79.11	71.35	57.47	60.59	48.10
Prakash Pimpale	CDACM	C	58.32	66.50	61.78	65.73	66.57	**85.02**
		UC	64.60	73.70	63.40	69.95	83.71	77.40
PreCogTexts	IIIT Delhi	C	68.80	76.60	66.70	73.75	84.68	77.40
IITP_MNIT	IIT Patna & MNIT Jaipur	C	**79.77**	75.00	63.20	70.21	84.85	77.50
		UC	**83.14**	78.80	65.00	72.19	72.73	76.90
		UC	79.77	75.00	63.20	70.21	84.85	77.50
Divya	IIIT-Hyderabad	C	–	–	–	–	–	–
IIIT-Hyderabad	IIIT-Hyderabad	C	72.63	79.95	68.52	76.07	78.03	81.75
		UC	71.66	79.21	67.74	**75.12**	77.11	80.93
KS_JU	Jadavpur University	C	71.6	79.6	**77.27**	69.95	**88.14**	82.20
NLP-NITMZ	NIT Mizoram	C	76.54	**82.25**	72.37	75.90	81.73	84.35
TeamSurukam	Surukam & IIIT-H	C	70.60	77.60	66.00	73.28	75.11	78.30
Anuj	YMCA	C	–	–	–	–	–	–
Learner	IITK	C	73.79	80.70	68.18	**76.14**	78.11	81.58

Table 2: f1 Score of BE-EN Corpus for Team participated in the tool contest

Team	Institute	Category (Const./ Unconst.)	HI					
			FB_FG (f1)	FB_CG (f1)	TWT_FG (f1)	TWT_CG (f1)	WA_FG (f1)	WA_CG (f1)
Amrita_CEN	Amrita Vishwa Vidyapeetham	C	**83.67**	78.48	90.17	85.75	68.57	73.81
		UC	71.07	62.83	**88.21**	**91.19**	**84.02**	**78.30**
BITS_PILANI_TEAM 1	BITS Pilani	C	82.21	70.86	83.16	77.66	56.98	61.34
		UC	81.60	72.16	83.51	78.21	72.90	64.58
BITS_PILANI_TEAM 2	BITS Pilani	C	65.08	64.92	78.74	77.92	69.80	61.84
		UC	65.46	65.17	78.57	77.92	68.63	60.84
BITS_PILANI_TEAM 3	BITS Pilani	C	68.84	63.51	70.88	57.77	64.18	70.33
		UC	68.84	63.55	70.22	56.89	64.18	70.33
MISIM-UB	University of Buffalo	C	64.84	63.55	74.94	74.66	70.96	59.30
Prakash Pimpale	CDACM	C	57.73	60.43	71.57	77.10	65.88	**82.74**
		UC	70.72	62.37	85.57	81.49	80.12	74.68
PreCogTexts	IIIT Delhi	C	78.41	70.36	85.97	89.44	82.57	76.30
IITP_MNIT	IIT Patna & MNIT Jaipur	C	72.09	62.69	87.24	**90.39**	79.85	77.30
		UC	80.72	75.11	86.92	89.89	74.88	69.20
		UC	72.09	62.69	87.24	90.39	79.85	77.30
Divya	IIIT-Hyderabad	C	80.39	75.62	**90.50**	85.33	64.09	69.70
IIIT-Hyderabad	IIIT-Hyderabad	C	83.07	79.40	79.04	84.55	67.83	74.43
		UC	**84.97**	**81.34**	78.07	83.44	65.08	71.94
KS_JU	Jadavpur University	C	72.58	63.23	89.60	85.43	**84.77**	79.17
NLP-NITMZ	NIT Mizoram	C	68.80	76.01	81.04	85.63	66.11	76.04
TeamSurukam	Surukam & IIIT-H	C	80.13	74.74	89.55	84.04	80.44	70.91
Anuj	YMCA	C	79.56	73.54	86.09	80.38	81.36	75.56
Learner	IITK	C	82.60	**79.45**	80.20	84.36	68.04	74.53

Table 3: f1 Score of HI-EN Corpus for Team participated in the tool contest

fication with character embeddings feature is used for tagging the words with its corresponding part-of-speech information in the unconstrained run and rich features followed by SVM classification in an constrained run. The important features for the POS tagging task have been identified based on the different possible combinations of available word and tag contexts. Other than the token features, binary features, punctuation features and the other features like length and position of the token are used. In the unconstrained run, this team achieved the highest numbers of f1 score in all the

Team	Institute	Category (Const./ Unconst.)	TE					
			FB_FG (f1)	FB_CG (f1)	TWT_FG (f1)	TWT_CG (f1)	WA_FG (f1)	WA_CG (f1)
Amrita_CEN	Amrita Vishwa Vidyapeetham	C	78.57	70.38	78.49	82.83	80.70	84.88
		UC	81.36	84.51	**82.58**	**86.09**	**85.69**	**88.19**
BITS_PILANI_TEAM 1	BITS Pilani	C	78.05	80.22	76.27	79.01	75.90	80.95
		UC	77.44	82.62	75.97	80.31	75.09	70.12
BITS_PILANI_TEAM 2	BITS Pilani	C	76.12	77.05	75.81	71.93	77.70	75.38
		UC	74.09	77.34	75.79	71.01	77.60	74.29
BITS_PILANI_TEAM 3	BITS Pilani	C	71.45	72.22	74.00	75.28	76.87	73.75
		UC	71.45	72.22	74.00	75.28	76.87	73.75
MISIM-UB	University of Buffalo	C	71.19	74.88	73.68	68.46	78.22	75.82
Prakash Pimpale	CDACM	C	67.08	79.30	63.88	74.28	73.30	83.69
		UC	**81.83**	**85.04**	76.11	66.90	84.95	87.83
PreCogTexts	IIIT Delhi	C	81.03	84.63	79.22	81.83	83.85	86.37
IITP_MNIT	IIT Patna & MNIT Jaipur	C	80.73	84.60	**81.89**	**84.81**	84.36	87.92
		UC	79.13	82.75	77.65	23.00	78.80	61.46
		UC	80.73	84.60	81.89	84.78	84.36	87.92
Divya	IIIT-Hyderabad	C	65.37	71.33	78.12	69.10	56.52	49.91
IIIT-Hyderabad	IIIT-Hyderabad	C	80.26	84.57	77.52	81.56	79.99	73.27
		UC	79.98	84.38	77.59	82.12	79.11	72.98
KS_JU	Jadavpur University	C	**82.13**	**85.60**	81.81	73.70	**84.90**	**88.10**
NLP-NITMZ	NIT Mizoram	C	72.95	79.88	72.87	75.01	72.45	78.26
TeamSurukam	Surukam & IIIT-H	C	65.25	82.94	77.16	81.13	78.05	70.81
Anuj	YMCA	C	–	–	–	–	–	–
Learner	IITK	C	80.11	83.94	79.86	82.15	81.84	74.79

Table 4: f1 Score of TE-EN Corpus for Team participated in the tool contest

dataset provided by the organizer.

BITS_PILANI_TEAM1 (Bhargava et al., 2016a) proposed approach uses machine learning approaches for POS tagging. Best set of features is extracted by applying a grid search over all the manually engineered features such as prefix, suffix, previous POS tag, current word language, starts With @, etc. Different Machine Learning techniques are then used to build the classification model such as Random Forest, Naive Bayes and Logistic Regression and evaluated using 5 fold cross validation. Finally the proposed methodology for constrained system uses Random Forest technique whereas proposed technique for unconstrained task, incorporates a small dictionary of the hundred most common English words along with their POS tags in conjunction to model built for constrained task.

BITS_PILANI_TEAM2 (Bhargava et al., 2016c) proposed approach uses various derived features and their combinations for training the classifier. Best set of features and classifier is then selected by splitting training data into ratio of 70:30 for judging the accuracy. After performing various experiments, Random forest is used for building constrained system. Unconstrained System was built using the same approach with Extra Tree classifier in conjunction of a dictionary, which contained POS Tags for frequently occurring words.

BITS_PILANI_TEAM3 (Bhargava et al., 2016b) used bi-directional Long Short Term Memory (LSTM) network to tag words in the corpora. In this system the best results obtained through the proposed algorithm of constrained as well as unconstrained system is 82%.

CDAC Mumbai Team (Patel et al., 2016) proposed an approach to POS tag code-mixed social media text using Recurrent Neural Network Language Model (RNN-LM) architectures such as Simple RNN, Long Short-Term Memory (LSTM), Deep LSTM Gated Recurrent Unit (GRU). They used the modified version of RNN-LM architecture, which predicts the POS tag of the word in the slot instead of predicting the word. The word sequence is the input to the system, similar to the standard RNN-LM. The approach is language independent and requires only plain text for POS tagging. The RNN-LM model uses the context words embedding as the input features. Out of the all experiments, GRU models are outperforms other models. The best f1 score obtained among all the language pair using GRU model in coarse grain (hi-en) and in fine grain (hi-en) is 78.29 and 69.32 respectively.

Jadavpur University Team KS_JU (Sarkar, 2016) developed POS tagging system uses a variety of contextual and orthographic features for

POS tagging. In constrained mode a Conditional Random Field (CRF) based model is developed. The overall average performance of the JU system over all three language pairs is F1 score of 79.991 which is the highest average F1 score among all 15 participating systems.

IIIT Delhi team named PreCogTexts (Singh and Sen, 2016), trained a Coarse-Grained tagger and a Fine-Grained tagger for each of the 3 language pairs. They used CRF based classifiers with hand-crafted features, supplemented with a list of rules to detect on-line social network specific tokens such as hashtags and mentions. The team observe that with 5-fold cross validation on the data-set, their coarse-grained classifier performs the best on the En-Hi data-set (0.816 F1 score), followed by En-Be (0.81 F1 score) and te-en (0.73 F1 score) where as fine-grained classifier performed best on be-en (0.752 F1 score), followed by hi-en (0.74 F1 score) and En-Te (0.682 F1 score).

IIT, Patna and MNIT, Jaipur jointly formed a team IITP_MNIT (Gupta et al., 2016) has claimed that their proposed system is language and platform independent which performed well on all three language pair and social media data into finer and coarser POS tag level. In their experiments, training has done in two different manner while the classification system remains the same for the both cases. First, their system was trained by augmenting the respective language pair training data from all three social media platform. Secondly, their system was trained individually by only respective language pair training data from different social media platform. CRF based classifier used in this case.

MISIM-UB team (Londhe and Srihari, 2016) from SUNY Buffalo, NY, USA proposed two models for fine and coarse grained POS tagging for code switched social media text for the given corpus and shows how a unified model that uses language partitioning combined with simplistic language models can produce competitive results across languages and media.

NLP-NITMZ team (Pakray and Majumder, 2016) from NIT, Mizoram proposed supervised learning approach to build the model. Further, conditional model is implemented POS tagging and then Bayesian classification based generative model is used for simplification. After that, this model is simplified based on two key assumption of Hidden Markov Model (HMM).

IIIT-Hyderabad Team (Pandey and Mishra, 2016) used for lexical and sub-lexical features to design constrained and unconstrained systems for icon corpus. For the unconstrained track of the competition, the team transliterated the romanized data-set into their respective scripts for unconstrained systems. This method helped improve accuracy. The team proposed 5-fold cross validation in the constrained setting for the released training data-sets. The final task of tagging of the code-mixed test corpus was achieved by implementing conditional random fields.

Team Surukam (Ramesh and Kumar, 2016) presented a CRF based POS tagger using a library called *sklearn-crfsuite* to achieve an overall f1 score of 76.45%.

from the Table 2, 3 and 4 shows the evaluation results of the submitted runs on constrained and unconstrained. Total 8 runs on unconstrained and 15 runs on constrained are submitted by the teams. Team IITP_MNIT of Gupta et al. (2016) submitted 2 runs on unconstrained. Among all the team, AMRITA_CEN of Amrita Vishwa Vidyapeetham achieved total highest number of f1 score in unconstrained run and team KS_JU from Jadavpur University (Sarkar, 2016) achieved total highest number of f1 score in constrained run.

Acknowledgements

Thanks to Björn Gambäck, Professor of Language Technology in the Department of Computer and Information Science from Norwegian University of Science and Technology,Trondheim, Norway for his kind support and guidance.

References

Jannis Androutsopoulos. 2011. Language change and digital media: a review of conceptions and evidence. In Tore Kristiansen and Nikolas Coupland, editors, *Standard Languages and Language Standards in a Changing Europe*, pages 145–159. Novus, Oslo, Norway,

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. AFNLP.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing *ya* mixing?": An analysis of English-Hindi code mixing

in Facebook. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 116–126, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.

Rupal Bhargava, Raghav Bhartia, Indrajeet Mishra, and Yashvardhan Sharma. 2016a. BITS PILANI TEAM1 @ POS Tagging for Code-Mixed Indian Social Media. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Rupal Bhargava, Gargi Sharma, and Yashvardhan Sharma. 2016b. BITS PILANI TEAM3 @ POS Tagging for Code-Mixed Indian Social Media. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Rupal Bhargava, Bapiraju Vamsi Tadikonda, and Yashvardhan Sharma. 2016c. BITS PILANI TEAM2 @ POS Tagging for Code-Mixed Indian Social Media. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Mónica Stella Cárdenas-Claros and Neny Isharyanti. 2009. Code switching and code mixing in internet chatting: between 'yes', 'ya', and 'si' a case study. *Journal of Computer-Mediated Communication*, 5(3):67–78.

Amitava Das and Björn Gambäck. 2013. Code-mixing in social media text: The last language identification frontier? *Traitement Automatique des Langues*, 54(3):41–64.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 169–178, Goa, India, December.

Mona Diab and Ankit Kamboj. 2011. Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for Hindi English code switched data: A pilot annotation. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 36–40, Chiang Mai, Thailand, November. AFNLP. 9th Workshop on Asian Language Resources.

Björn Gambäck and Amitava Das. 2014. On measuring the complexity of code-mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1–7, Goa, India, December. 1st Workshop on Language Technologies for Indian Social Media.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 677–686, New York, NY, USA. ACM.

Deepak Gupta, Shubham Tripathi, Asif Ekbal, and Pushpak Bhattacharyya. 2016. SMPOST: Parts of Speech Tagger for Code-Mixed Indic Social Media Text. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Yuhen Hu, Kartik Talamadupula, and Subbarao Kambhampati. 2013. *Dude, srsly?*: The surprisingly formal nature of Twitter's language. In *Proceedings of the 7th International Conference on Weblogs and Social Media*, Boston, Massachusetts, July. AAAI.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proceedings of 10th International Conference on Recent Advances in Natural Language Processing*, pages 239–248, in Hissar, Bulgaria, 7-9 September, September.

Anand M Kumar and K P Soman. 2016. AMRITA_CEN: Character Embeddings for Code-Mixed Part-of-Speech Tagging in ICON2016 NLP Tools Contest. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Nikhil Londhe and Rohini K Srihari. 2016. Language Models for POS Tagging of Code–Mixed India Social Media Text : UB submission at ICON'16. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Partha Pakray and Goutam Majumder. 2016. NLP–NITMZ @ Part–of–Speech (POS) Tagging for Code–Mixed India Social Media Text using HMM. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Ayushi Pandey and Pruthwik Mishra. 2016. POS Tagging for Code-Mixed Indian Social Media Text: Systems Submitted for Tools Contest in ICON-2016. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

John Paolillo. 1999. The virtual speech community: Social network and language variation on IRC. *Journal of Computer-Mediated Communication*, 4(4), June.

Raj Nath Patel, Prakash B. Pimpale, and M. Sasikumar. 2016. Recurrent Neural Network based Part-of-Speech Tagger for Code-Mixed Social Media Text. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Sree Harsha Ramesh and Raveena R Kumar. 2016. A POS Tagger for Code-Mixed Indian Social Media Textt — ICON-2016 NLP Tools Contest Entry from Surukam. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Kamal Sarkar. 2016. A CRF based POS tagger for Code-mixed Indian Social Media Text. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, and Kunal Chakma. 2015. Overview of FIRE-2015 shared task on mixed script information retrieval. In *FIRE Workshops*, volume 1587 of *CEUR Workshop Proceedings*, pages 19–25. CEUR-WS.org.

Kushagra Singh and Indira Sen. 2016. POS Tagging in Code-Mixed Online Social Media Text. In *Working notes of ICON Tool Contest 2016, 13th International Conference on Natural Language Processing*, IIT (BHU), Varanasi, December.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Doha, Qatar, October. ACL. 1st Workshop on Computational Approaches to Code Switching.