ISSN 2007-9737

Named Entity Recognition on Code-Mixed Cross-Script Social Media Content

Somnath Banerjee[1], Sudip Kumar Naskar[1], Paolo Rosso[2], Sivaji Bandyopadhyay[1]

[1] Jadavpur University, Kolkata,
India

[2] Universitat Politècnica de València, Valencia,
Spain

sb.cse.ju@gmail.com,sudip.naskar@cse.jdvu.ac.in,
prosso@dsic.upv.es,sbandyopadhyay@cse.jdvu.ac.in

Abstract. Focusing on the current multilingual scenario in social media, this paper reports automatic extraction of named entities (NE) from code-mixed cross-script social media data. Our prime target is to extract NE for question answering. This paper also introduces a Bengali-English (Bn-En) code-mixed cross-script dataset for NE research and proposes domain specific taxonomies for NE. We used formal as well as informal language-specific features to prepare the classification models and employed four machine learning algorithms (Conditional Random Fields, Margin Infused Relaxed Algorithm, Support Vector Machine and Maximum Entropy Markov Model) for the NE recognition (NER) task. In this study, Bengali is considered as the native language while English is considered as the non-native language. However, the approach presented in this paper is generic in nature and could be used for any other code-mixed dataset. The classification models based on CRF and SVM performed well among the classifiers.

Keywords. Named entity recognition, code-mixed cross-script, Bengali-English social media content.

1 Introduction

With the rise of social media (SM), internet users are sharing information on various social media platforms (forums, Facebook, Twitter, blogs, etc.) than ever before and the information contents are mostly informal in nature. Code-mixing or language-mixing refers to the phenomenon where lexical items and grammatical features from more than one language appear in one sentence [23]. It is a common phenomenon and, in fact, widely used in multilingual communities. For majority of the multilingual speakers in Asian countries, English serves as their second language. Even some European communities such as French, German, Spanish, Italian, etc., use English and their native language alternatively as the language of classroom instruction [27].

Every natural language is generally written using a particular script (sometimes multiple, e.g., Chinese, Japanese, etc) which is referred to as the native script for that language. The phenomenon of using a non-native script (other than the native script) phonetically for writing native words is referred to as cross-script. For example, if a Bengali user writes Bengali words in Bengali script, that is considered to be using native script. However, if he writes Bengali words in Roman script or English words in Bengali script, then he is making use of cross-script. In spite of of having indigenous script(s), multilingual users often write SM content (SMC) in non-native Roman script due to various socio-cultural and technological reasons [1, 6].

NE have a distinct feature in languages, i.e., NE refer to specific things or concepts in the world and typically are not listed in lexicons. NER is the task that seeks to locate and classify NE in texts into predefined categories such as

doi: 10.13053/CyS-21-4-2850

the names of persons, organizations, locations, time expressions, quantities, etc. Automatic identification and classification of NEs benefit text processing due to their important significance in many NLP applications such as question answering, automatic summarization, information extraction, information retrieval, machine translation, etc. The NER task can be viewed as a two-phase process: (a) identification of entity boundaries and (b) classification into the correct category. For example, if *Donald Trump* appears in a sentence, it is essential to identify the beginning and end of this NE in the sentence. Following this step, the entity must be classified into the proper predefined category, which is type *Person* in this case.

In the context of code-mixed cross-script (CMCS), we have initiated to develop a CMCS question answering (QA) system for Bengali–English code-mixed data. As the answers to factoid questions are NE, we need to extract NE from the CMCS data. To the best of our knowledge, there exists no NER for CMCS SMC. Therefore, to tackle the problem of CMCS QA, we carried out a study on CMCS NER.

In addition to the typical challenges in natural language processing, different forms of user generated noise present additional challenges while processing CMCS SMC. Some of the characteristics of CMCS SMC are given below:

- A word is spelled differently by various speakers (e.g., korchi (English gloss: “am doing”) - *korchee*, *krchi*, *krchee*, *krchii*; night - *n8*, *ngt*, *ni8*).
- Words are contracted phonetically for the ease of writing. e.g. *grt* - great. *2morw* - tomorrow.
- Generally punctuations are omitted from contracted words. (Examples: can’t - *cant*, won’t - *wont*).
- Words are often intentionally misspelled (refered to as ‘wordplay’) for emphasis, e.g., *i m veryyyy happy* (“I’m very happy”).
- Asterisk (*) and numbers are used to encrypt vulgar words such as *f**k*, *b1tch*.
- Sometimes unintentional misspellings (typos) occur such as *coulf* - could, *fone* - fine, etc.
- Use of orthographically common vocabulary words create language identification problem. E.g. *take* is a valid English word as well as the transliteration of a Bengali word (English gloss is *him*).
- Usually, grammatical capitalization rules are not followed.

The rest of the paper is organized as follows. Section 2 discusses literature review. Section 3 presents the proposed NE taxonomy. The dataset is described in Section 4. The features used in this study are discussed in Section 5. Section 6 details the experiments carried out and presents the results together with some analysis. Section 7 concludes the paper and provides avenues for future work.

2 Related Work

NER for monolingual text has been studied extensively over the last two decades and state-of-the-art algorithms achieved high accuracy on formal texts in English [16] and several other languages [14, 28, 2] including Bengali [13, 11, 12, 5]. However, existing NER approaches do not perform well on informal social media text and their performance decreases significantly [10]. A few studies [24, 20, 21, 19] on NER for tweets as social media data were reported for English. In [26], the authors reported that the performance of standard NLP tools severely degrade on tweets. Based on word embeddings NER studies for informal text in English and Turkish were carried out separately [24]. In [25], the authors studied NER for Chinese SMC where they acquired Chinese SMC data from the popular Sina Weibo service and they used both unlabeled as well as labeled data for embeddings. In [20], the authors presented the HybridSeg framework which segments tweets into meaningful phrases called segments by using both global and local context. They reported that segmentation of tweets helps to preserve the semantic meaning of tweets, which subsequently benefits NER. In [21], the authors proposed a method for representing business NEs in English with a term distribution generated from web data and from social media that aligns more closely with user search query terms. In [3], the authors proposed an approach for recognizing targetable NEs, i.e., NEs in a targeted set (e.g., movies, books, TV shows, etc). In [15], the authors proposed a multilingual named entity recognition system using language independent feature templates and tested their models on Spanish and English data separately.

Text analytics on social media text have emerged as a new research area and various shared tasks have been organized on SMC in the last few years. In the context of NER, a shared task on noisy user-generated text (W-NUT[1]) was organized by Microsoft in ACL-IJCNLP'2015 where two subtasks were proposed, namely twitter text normalization and named entity recognition for English tweets. Recently, in FIRE'2015 a shared task (ESM-IL[2]) was organized to identify NEs from code mixed (Hindi, Malayalam, Tamil and English) social media data. However, to the best of our knowledge, NER study has not been conducted so far on the CMCS SMC data addressed in this paper.

3 NER Taxonomy

In formal text processing (i.e., non-code-mixed non-cross-script) for NE identification in English and other languages, the research studies mainly considered person, location, organization, temporal expression and quantity as NE classes. As discussed in Section 1, recently a few studies have been conducted on SMC to identify NEs where majority of the studies proposed the use of three basic classes – person, location and organization, and a few targetable classes such as sports team, movie name, etc. As mentioned earlier, the target domains of this study are sports and tourism. Therefore, we propose here two different NE taxonomies - one for the sports domain and another for the tourism domain. The NE classes included in the taxonomies were considered after carefully analyzing the corpus. It was observed from the sports and tourism domain corpora that in addition to identifying the well-known NE classes such as person, location, organization, temporal expression, etc., we need to identify a few domain specific classes in order to develop a properly functioning factoid QA system.

We proposed six coarse-grained classes and one domain specific class (Sports terms) for the sports domain. The detailed taxonomy is given in Table 1. In the tourism domain, besides the 6 well-known basic classes, three domain specific coarse-grained classes are proposed, namely - Transport, Tourism event and Artifact. The transport class represents objects of type vehicles such as *Calcutta Delhi Express, Volvo Bus, etc*. The tourism event class represents various cultural events such as fairs (e.g., *Rash Mela*), festivals (e.g., *Vasanta Utsav*), etc. The artifact class represents tourism spot specific valuable objects worth buying (e.g., *Baluchori Sharee, Teracota Horse, etc.*), sightseeing (e.g., *palace, sea beach, mountain, etc.*), experiencing (e.g., *opera, mountaineering,* etc.), or available entertainment activities (e.g. *skiing*, *scuba diving*, *hiking*, etc).

4 Dataset and Preprocessing

For this study, we prepared an experimental dataset from the dataset described in [6] which is the only CMCS dataset available for question answering research. The dataset contains questions, messages and answers from the sports and tourism domains in CMCS English–Bengali. The sports domain dataset contains texts on *cricket*, a popular outdoor game in the Indian subcontinent and many parts of the world. The dataset contains a total of 20 documents from the two domains. The sports domain contains 10 documents which consist of 116 informal posts and 192 questions, while the 10 documents in the tourism domain consist of 183 posts and 314 questions.

In order to extract the NEs, we had to preprocess the described corpus. The CMCS posts contained in the corpus are typically short in nature and informal. The posts are usually made up of 2/3 complete or incomplete sentences and are referred to as message segments in [6]. In most of the

[1] https://noisy-text.github.io/

[2] http://au-kbc.org/nlp/ESM-FIRE2015/

Table 1. Taxonomy and tagset for Sports and Tourism domains

Tag	Name	Sports	Tourism	Description
PER	Person	✓	✓	name of persons
LOC	Location	✓	✓	name of locations
ORG	Organization	✓	✓	name of organizations
QUAN	Quantity	✓	✓	numerical values
TEMP	Temporal	✓	✓	time related expressions
MISC	Miscellaneous	✓	✓	NEs which not fall in other classes
SPORTS	Sports terms	✓	X	sports related terms
TEVENT	Tourism Event	X	✓	recreation events in tourism spot
TRANS	Transport	X	✓	objects of type vehicles
ARFACT	Artefact	X	✓	temples, valuable objects etc.
DIST	Distance	X	✓	measurable distance related expressions
MONEY	Monitory	X	✓	money related expressions

cases the message segments contain incomplete sentences or a few words. It was observed that separating the message segments are not straightforward and requires extra effort.

One of the challenges in segmenting the messages is to identify the dot as segment separator. Sometimes the last word of a message segment is merged with the first word of the next message segment by one or multiple dots (i.e., no spaces). Such cases having the pattern ⟨*last-word*⟩.⟨*first-word*⟩ overlap with the cases like *Mr.Singh, a.m., p.m.,* etc. The confusion occurs when the segment contains title words (e.g., *Mr., Ms.*, etc.), measurement words (e.g., *2.8 ft, 3.2 km, 32 km., 3.3 over*, etc.), temporal words (e.g., *3 a.m, 4.20 pm, 5 p.m.*, etc.), etc.

We used regular expressions and gazetteers to tackle these situations and separate the message segments. Subsequent appearances of various symbols such as '!' , '?' were noted in the corpus. A few occurrences were also noticed where repetition of symbols are merged (e.g. *!!!...*). We dealt such cases with regular expressions.

After processing, the said symbol combinations are replaced by a single dot. Thus, some of the tokens are split while some tokens are replaced by a single token. Therefore, the total number of tokens in our preprocessed corpus is different than as reported in [6].

Annotating the whole corpus (i.e., 299 messages and 506 questions) manually for NEs is a tedious and time consuming task. Moreover, the CMCS nature of the corpus introduces further complications in data annotation framework. Bilingual speakers having proficiency and sound linguistic knowledge in both the languages should be employed for the annotation task.

Although, crowd-sourcing can be an alternative approach, it is not a very reliable option for CMCS data annotation because of the risk of inaccurate annotations [17]. We annotated each domain dataset with two bilingual annotators having proficiency in both Bengali and English. The inter-annotator agreement measured in terms of Cohen's Kappa [7] were 0.71 and 0.73 for the tourism and sports domains, respectively. In case of disagreement, a bilingual linguistic expert was consulted and his decision was taken as final.

The statistics of the prepared dataset are presented in Table 2. Almost half the tokens in the sports domain are NE, while about one-third (1/3rd) of the tourism domain tokens are NE. Location is the most frequent NE class in the tourism corpus, while it is the least frequent in the sports domain.

The presence of Temporal NE is very less in Sports corpus (overall:4) compared to the tourism corpus (overall:81). The total number of NE is more in tourism corpus than that of sports corpus. After annotating, we divided the datasets belonging to each domain into approximately 7:3 ratio for training and testing, i.e., 70% and 30% data are used for training and testing, respectively.

Table 2. Corpora Statistics

		Message(M)			Question(Q)			M+Q		
Domain	Tag	Train	Test	Overall	Train	Test	Overall	Train	Test	Overall
Sports	PER	70	27	97	26	17	43	96	44	140
	LOC	1	1	2	0	0	0	1	1	2
	ORG	76	46	122	43	16	59	119	62	181
	TEMP	2	2	4	0	0	0	2	2	4
	QUAN	58	29	87	17	4	21	75	33	108
	SPORTS	133	46	179	135	54	189	268	100	368
	MISC	6	4	10	2	0	2	8	4	12
	Overall NE	**346**	**155**	**501**	**223**	**91**	**314**	**569**	**246**	**815**
	Not NEs	792	345	1137	455	173	628	1247	518	1765
Tourism	PER	10	11	21	8	2	10	18	13	31
	LOC	120	93	213	136	82	218	256	175	431
	ORG	29	10	39	12	17	29	41	27	68
	TEMP	52	24	76	5	0	5	57	24	81
	QUAN	16	5	21	0	1	1	16	6	22
	TEVNT	12	6	18	8	4	12	20	10	30
	TRANS	57	33	90	49	22	71	106	55	161
	ARFACT	25	38	63	42	27	69	67	65	132
	DIST	13	23	36	2	1	3	15	24	39
	MONEY	31	17	48	0	4	4	31	21	52
	MISC	26	13	38	12	10	22	38	23	61
	Overall NEs	**391**	**273**	**663**	**274**	**170**	**444**	**665**	**443**	**1108**
	Not NE	1290	954	2244	984	629	1613	2274	1583	3857

Overall NE: total number of NE in dataset; **Not NE**: words which are not NE

5 Feature Set

For CMCS NER classification, we employed monolingual as well as CMCS features. For the tourism domain, all of the features were used except the cricket vocabulary while for the sports domain, tourism related features were not employed.

Contextual Features: These features include context cues such as the current token (anchor word) along with the previous and next tokens. Context features are used extensively for monolingual NER. Therefore, we also consider this feature in this study.

Capitalization: : In English, if a word starts with a capital letter or the entire word is in uppercase then it is highly likely to be an NE. Although capitalization is a key orthographic feature for NEs, this feature may be misleading for SMC. In SMC, non-NEs are often capitalized for emphasis while NEs are often written in small. We further classified this feature into four specialized cases: entire word is in uppercase, first letter of the word is in capital, any intermediate letter is in capital, and other than the aforementioned three cases.

Alphanumeric: In social media content, users often express legitimate vocabulary words in alphanumeric form for saving typing effort, to shorten message lenth, or to express their style. Examples include abbreviated words like 'gr8' ('great'), 'b4' ('before'), etc. We observed by analyzing the corpus that alphanumeric words generally are not NEs. Therefore, this feature serves as a good indicator to recognize negative examples.

Length of the word: This feature is widely used for the regular NER task under the assumption that NEs typically have short word length.

Language of the token: In the CMCS content scenario, language of the word is an important feature for identifying NE since out of vocabulary words have a high chance of being NE. In the present work, the language of the token automatically identified by the language identifier is considered as a feature. We employed the language identifier described in [4] which reported an accuracy of 92.88%.

Presence in English dictionary: In English, very few dictionary words are NEs. Therefore, we checked whether a token identified as an English word is a valid dictionary word from an *Engilsh word list*[3] having 355 thousands words.

Gazetteer list: In the NER task, gazetteer lists are very helpful for identifying specific classes. For example, names of the weekdays, months are very helpful to recognize temporal NEs. Similarly, abbreviations like *Mr., Mrs., Dr.,* etc. are often used before person names. Thus, the list of honorifics could be used as a potential clue for identifying person NEs. We manually prepared 3 gazetteer lists, namely - honorifics, names of weekdays, names of months.

Cricket word list: This feature is used as a binary feature. We collected the cricket vocabulary list from Wikipedia and compiled it under human supervision. This feature plays a crucial role to recognize cricket specific NEs.

Temporal cues: These features play a crucial role to recognize temporal NEs. We employed four cues separately. All the cues are used as binary features.
Cue-1: Often *@* or the preposition *at* sits before time expressions (e.g., *arrvd murshidabad at 11:31 ...*). If the previous word is *at* or *@*, then the value is set to 1, otherwise set to 0.
Cue-2: People generally use ':', '.' or '-' in time expressions, e.g., *4:30 am, 4.20 pm, 4-30* etc. Therefore, if ':', '.' or '-' is present in the token then there is a high chance that the token being considered is a temporal expression.
Cue-3: If the next word of the token belongs to the set {*am, pm, hrs, min, sec*}, then the current token is most likely a temporal NE.
Cue-4: Generally year is expressed in four digits in text. Therefore, if the current token is a four digit word then it can represent a year and hence the value is set to 1 and 0 otherwise.

Distance cues: Distance between two places or height of a mountain or area of a construction is generally expressed as $\langle value\rangle\langle unit\rangle$ or $\langle value\rangle\langle space\rangle\langle unit\rangle$, e.g., *22km, 3000 ft, 234 hector*, etc. We built a gazetteer that contains the units of distance. For both the expressions, $\langle unit\rangle$ token is checked in the distance gazetteer. This is also a binary feature.

Transport cues: This is another binary feature. The last word of the modes of transport may be *expr*, *express*, *bus*, *cabs*, etc., e.g., *Volvo bus*, *Puri Express*. *Kanchanjangha expr*, *Ola cabs*, etc. Therefore, if the following word of the current token belongs to the transport gazetteer (which contains transport related CMCS spelling variations along with the original words) then this feature value is set to 1, and 0 otherwise.

Event cues: It is a binary valued feature. NE of type tourism event may end with one of the words in the set E = {*mela*, *utsab*, *utsav*, *jayanti*, *puja*, *pujo*, *festival*, *fair*}. The set E contains event related words from both languages. Therefore, the event flag value is set to 1 if the next word of the current token is in set E.

Monetary cues: Monetary entities may start with currency words/symbols (e.g., *Rs. 5*, *npr 10*, \$20, €19, etc.) or may end with them (e.g., *5 taka*, 20 €, etc). If the token under consideration represents a number and the previous or the next token represents a currency, the value is set to 1 and 0 otherwise.

Quantity Feature: Quantity NEs are expressed using ordinal and cardinal numbers. In social media, ordinal numbers are often expressed in alphanumeric fashion such as *1st, 2nd, 3rd, 4th,* etc. Therefore, if the current token is alphanumeric and starts with a number and followed by an alphabetic string belonging to the set S ={*st, nd, rd, th*}, the value is set to 1. Also, if the current token is solely alphabetic and belongs to set $N_{ordinal}$ ={*first, second, third*, ...}, then the value is set to 1. If a token is a cardinal number, i.e., it belongs to the set $N_{cardinal}$ = {*one, two, three,* ..., 1, 2, 3, ...}, then the value is also set to 1.

[3] https://github.com/dwyl/english-words

6 Experiments and Results

In this section, we discuss the various experiments carried out and present the evaluation results. We carried out evaluation employing the standard evaluation metrics - accuracy, precision (P), recall (R) and F-1 score using the *conll evaluation script*[4]. Two experiments were performed separately for the sports and tourism domain. In both the cases, four classifiers were employed separately, namely CRF, MIRA, MEMM and SVM. Three models were built for each classifier to evaluate the efficiency on – messages, questions and combining messages and questions.

M-model: This model was prepared using the CMCS messages/posts as training data.

Q-model: This model was developed using the CMCS questions as training data.

MQ-model: This model was built on the combined CMCS messages and questions as training data.

6.1 Classifiers Employed

For this NER study on CMCS content, we experimented with four machine learning based sequence labeling algorithms: Conditional Random Fields (CRF) [18], Margin Infused Relaxed Algorithm (MIRA) [9], Support Vector Machine (SVM) [8] and Maximum Entropy Markov Model (MEMM) [22]. We employed open source tools *CRF++*[5], *miralium*[6], *YamCha tool*[7] and *Wapiti toolkit*[8] for implementing CRF, MIRA, SVM and MEMM, respectively.

6.2 Baseline

In absence of any NER system for CMCS Bn–En data, we employed off-the-shelf three models (3-class, 4-class and 7-class) of the Stanford NER [16] as Baseline. The 3-class model contains Person, Organization and Location. Since our NE taxonomy contains classes additional to these 3 classes, it is not possible to map all our NE classes to the 3-class taxonomy. The 4-class model contains Person, Organization, Location and MISC. Therefore, other than the 3 basic classes (i.e. Person, Organization and Location), the rest of the classes in our NE taxonomy were mapped to the MISC class. The 7-class model contains 7 classes: Person, Location, Organization, Date, Money, Percent, Time. The 'Percent' class was mapped to 'QUAN'; and the 'Date' and 'Time' classes are analogous to the 'Temporal' class. We applied the three models on the tourism and sports datasets and the obtained results are reported in Table 3.

6.3 Tourism Domain Experiments

While preparing the models for the tourism domain, all the features (cf. Section 5) were used except the cricket word list. As reported in Table 2, the tourism corpus has the following datasets: message trainset (M$_{train}$), message testset (M$_{test}$), question trainset (Q$_{train}$), question testset (Q$_{test}$), message-question trainset (MQ$_{train}$) and message-question testset (MQ$_{test}$). Detailed results obtained on the tourism domain dataset are presented in Table 4 and Table 5. In aforesaid tables, the model-specific best results are shown in bold and the cross-dataset best results are shown in italics.

The four M-models built on the M$_{train}$ data employing the four classifiers were tested on the M$_{test}$ (NE: 273). Overall, the SVM based model performed best (identified 221 NE of which 126 were correct). The CRF, MIRA and MEMM classifiers identified 229, 244 and 182 NE respectively of which 123, 107 and 62 NE were correct respectively. Similarly, for the Q-Models built on Q$_{train}$ and tested on Q$_{test}$, MIRA achieved the best performance (identified:168; correct:95) in terms of F-1 and accuracy. MEMM performed well behind the others classifiers (F-1:43.42%). Analogously, for the MQ-models trained on MQ$_{train}$ and tested on MQ$_{train}$, the SVM classifier outperformed the other three classifiers with an F-1 score of 64.93% and 86.96% accuracy (identified:361; correct:261).

Additionally, four cross-dataset experiments were carried out - (i) M-models tested on the Q$_{test}$;

[4] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt
[5] https://taku910.github.io/crfpp/
[6] https://code.google.com/p/miralium/
[7] http://chasen.org/ taku/software/yamcha/
[8] https://github.com/Jekub/Wapiti

Table 3. Baseline Results

		Question				Post			
		Acc	P	R	F-1	Acc	P	R	F-1
Tourism	3-class	76.06	33.66	15.32	21.05	71.83	25.88	9.92	14.35
	4-class	76.51	35.15	15.99	21.98	71.98	26.27	10.11	14.60
	7-class	73.65	43.59	7.66	13.03	69.12	22.64	07.87	11.68
Sports	3-class	68.76	49.06	14.44	22.32	71.90	51.79	19.69	28.54
	4-class	68.86	47.47	14.97	22.76	72.02	51.01	20.08	28.82
	7-class	64.11	09.76	01.27	02.25	67.32	27.62	05.77	09.54

(ii) Q-models tested on the M$_{test}$; (iii) MQ-models tested on the Q$_{test}$; and (iv) MQ-models tested on the M$_{test}$. It is to be noted that M$_{test}$ and Q$_{test}$ together make up MQ$_{test}$. When the M-models were applied on Q$_{test}$, it was noted that accuracies of the models either increased or dropped slightly, whereas there were notable decrease in performance when Q-models were tested on M$_{test}$. This can be attributed to the fact that training set size of the M-models (i.e., M$_{train}$) is far bigger than that of the Q-models (i.e., Q$_{train}$). When the MQ-models were applied on Q$_{test}$ and M$_{test}$, as expected, the NER performance for all the classifiers improved drastically for both testsets. The SVM based model correctly identified 154 NE in M$_{test}$ and 107 NE in Q$_{test}$. The MQ-model based on SVM outperformed the other classifiers with a good margin on both the testsets, i.e., M$_{test}$ and Q$_{test}$.

6.4 Sports Domain Experiments

Similar to the tourism domain, the sports corpus (cf. Table 2) has the following datasets: message trainset (M$_{train}$), message testset (M$_{test}$), question trainset (Q$_{train}$), question testset (Q$_{test}$), message-question trainset (MQ$_{train}$) and message-question testset (MQ$_{test}$). Like the tourism domain, three NER models, namely M-model, Q-model and MQ-model were built for each classifier. Contextual features (target word and surrounding words), capitalization, alphanumeric, language, presence in English dictionary, gazetteer list and cricket word list were employed as features among which cricket word list feature is the only domain specific feature. Table 6 and Table 7 show the experimental results for the sports domain. In aforesaid tables, the model-specific best results are shown in bold and the cross-dataset best results are shown in italics.

The performance of the CRF classifier was the best with respect to accuracy except for Q-models. Compared to the tourism domain, we obtained very high accuracies and F-1 scores with the Q-models for the sports domain among which the SVM classifier resulted in the highest accuracy and F-1 score. With the MQ-models, CRF classifier performed the best and provided slight gains over other classifiers.

For the cross-dataset experiments, the M-models performed better than the Q-models on Q_{test} in terms of accuracy (except SVM) and F-1 score. Since the questions are based on messages/posts, NE present in Q$_{test}$ are also present in M$_{train}$ and the size of M$_{train}$ is much larger than Q$_{train}$ dataset. As expected, the performance of all the Q-model classifiers degraded by a large margin when tested on M$_{test}$. The smaller size of Q$_{train}$ causes this relatively low performance. Application of the MQ-models on both M$_{test}$ and Q$_{test}$ enhanced the NER performance.

6.5 Observations

Baseline results confirmed that NER model used for formal dataset can not handle the informal CMCS NE efficiently. For both the domains, the performance obtained using baseline models were very poor and 7-class model achieved less than 10% of F-1 score for sports domain.

The tourism corpus contains 11 classes whereas sports corpus 7 classes. The performance of all the models were better for sports domain in comparison to tourism domain. One of the reasons

Table 4. Results:Tourism Domain

		CRF			MIRA			MEMM			SVM		
Model	Testset	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
M-Model	M	53.71	45.05	49.00	43.85	39.19	41.39	34.07	22.71	27.25	57.01	46.15	**51.01**
	Q	65.32	47.65	55.10	57.89	45.29	50.83	60.66	43.53	50.68	71.65	53.53	***61.28***
Q-Model	M	50.25	36.63	***42.37***	34.71	39.93	37.14	39.29	32.23	35.41	49.46	33.7	40.09
	Q	63.28	47.65	54.36	56.55	55.88	**56.21**	49.25	38.82	43.42	63.48	42.94	51.23
MQ-Model	M	60.17	52.01	55.80	45.55	48.72	47.08	50.25	36.63	42.37	67.54	56.41	***61.48***
	Q	69.85	55.88	62.09	61.35	58.82	60.06	61.24	46.47	52.84	80.45	62.94	***70.63***
	MQ	63.71	53.50	58.16	51.21	52.60	51.89	54.57	40.41	46.43	72.30	58.92	**64.93**

Table 5. Results:Tourism Domain (accuracy)

Model	Testset	CRF	MIRA	MEMM	SVM
M-Model	M	**83.32**%	81.63%	78.84%	82.88%
	Q	85.63%	84.23%	83.64%	***85.98***%
Q-Model	M	***79.21***%	77.88%	76.41%	77.22%
	Q	**86.57**%	87.15%	83.53%	84.70%
MQ-Model	M	85.53%	83.69%	82.07%	***85.67***%
	Q	88.79%	88.67%	85.16%	***89.02***%
	MQ	86.78%	85.61%	83.27%	***86.96***%

Table 6. Results:Sports Domain (accuracy)

Model	Testset	CRF	MIRA	MEMM	SVM
M-Model	M	**86.56**%	85.80%	85.03%	85.41%
	Q	***95.04***%	93.62%	93.97%	93.97%
Q-Model	M	72.74%	74.47%	***75.24***%	73.13%
	Q	93.26%	93.62%	92.91%	**95.74**%
MQ-Model	M	***88.87***%	87.52%	87.33%	85.60%
	Q	***97.16***%	96.81%	95.74%	95.39%
	MQ	**91.78**%	90.78%	90.29%	89.04%

may be the presence of NEs (almost 50%) in the sports traing dataset is much higher than the tourism training dataset.

It was observed from the experimental results that artifacts were often misclassified as locations since the location class shares overlapping features with the artifact class.

Four digits words (e.g., 1890) were misclassified a few times since four digit words can represent year, distance, quantity or money. A few examples from the corpus are given below for which the system resulted in false positives.

Example-1: *poisa thakle onek expensive hotel ache-2300*\B-MONEY *per night.*

Example-2: *2050*\B-DIST *m*\I-DIST *height e Manali...*

The system made mistakes in correctly tagging the word '*temple*'. This is because of the fact that the word is tagged inconsistently in the training set; it is tagged as both NE (Miscellaneous and Artifact) and not an NE, and the distribution of all these cases are almost equal in the training set. Some of the occurrences of '*temple*' in the trainset are given below:

Example-3: *sudui temples*\O *dekhlam aj.*

Example-4: *Amra Hadimba*\B-ARFACT *Temple*\I-ARFACT ...

Similar to the case of the word '*temple*', there are a few words (e.g., '*river*') which are not NE themselves and belong to the English dictionary. The trained models sometimes misclassify such words when they turn into NEs. In the example given below, '*river*' is used to construct the Artifact NE.

Example-5: *Vyaas*\B-ARFACT *river*\I-ARFACT *r pashe Himalay...*

In majority of the training examples, the word '*taxi*' was tagged as Transport. However, in the tourism testset there are cases where the word *taxi* is tagged as Person (e.g. *taxi driver*) or Location (e.g. *taxi stand*). The trained model could not correctly classify the word *taxi* in some of those cases. A few examples from the testset are given below:

Example-6:*ghorar jonno taxi*\B-TRANS *available .*

Example-7:*Taxi*\B-PER *driver*\I-PER *r lunch pay korte hoyechilo*

Example-8: *Airport theke taxi*\B-LOC *stand*\I-LOC ...

In the sports domain, many sports related words such as *out*, *run*, *match*, *all*, etc., belong to the English dictionary. These words mislead the training models since sometimes these words are (part of) NE and sometimes not. This affects the results of the sports domain.

Table 7. Results:Sports Domain

		CRF			MIRA			MEMM			SVM		
Model	Testset	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
M-Model	M	68.39	68.39	68.39	69.87	70.32	**70.10**	78.05	61.94	69.06	68.99	70.32	69.65
	Q	91.76	85.71	88.64	90.36	82.42	86.21	95.00	83.52	***88.89***	88.51	84.62	86.52
Q-Model	M	50.26	62.58	55.75	49.46	58.71	53.69	55.68	63.23	59.21	51.9	70.32	***59.73***
	Q	82.42	82.42	82.42	84.62	84.62	84.62	83.33	82.42	82.87	88.17	90.11	**89.13**
MQ-Model	M	73.01	76.77	***74.84***	72.29	77.42	74.77	76.98	69.03	72.79	68.42	75.48	71.78
	Q	92.31	92.31	***92.31***	92.22	91.21	91.71	95.29	89.01	92.05	86.96	87.91	87.43
	MQ	79.92	82.52	**81.20**	79.30	82.52	80.88	83.93	76.42	80.00	74.90	80.08	77.41

7 Conclusion

This paper presents NER on CMCS data, with the goal of developing a QA system. The experimental dataset contains CMCS Bn–En data where Bengali words appear in English (Roman) script. The paper first proposes domain specific taxonomies for the sports and tourism domains, and then presents experiments and results on NER using four different classifiers. Since the NER work reported here is specifically targeted towards development of QA system, we proposed NE taxonomies that are suitable for question classification and answering. The proposed NE taxonomies are comprised of generic basic classes along with domain specific classes. Four classifiers (i.e. CRF, MIRA, MEMM and SVM) were employed to build each of the three models (i.e. M-model, Q-model and MQ-model) for both domains. As expected, the combined models (i.e. MQ-models) outperformed the individual models with notable margins. The CRF based combined NER model performed best for all the sports domain testsets, whereas, SVM classifier was the best performer for all testsets in the tourism domain. The approach presented here is generic and could be used for any CMCS dataset. Our contributions in this paper include the following points:

— Introducing and addressing for CMCS SM data as a research problem.

— Creation and annotation of a Bn–En CMCS dataset for the NER task.

— Proposal of single-layer taxonomies for the sports and tourism domains, which contain basic as well as domain specific classes.

— Proposal of suitable features targeted towards the said task.

— Developing machine learning models from CMCS QA dataset to identify NE for CMCS QA.

As future work, we would like to investigate the use of ensemble techniques and the state-of-the-art deep learning architectures to enhance the performance of NER on CMCS data.

Acknowledgements

Somnath Banerjee and Sudip Kumar Naskar are supported by Media Lab Asia, MeitY, Government of India, under the Visvesvaraya PhD Scheme for Electronics & IT.

The work of the third author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

1. **Ahmed, U. Z., Bali, K., & Choudhury, M. (2011).** Challenges in designing input method editors for indian languages: The role of word-origin and context. *Proceedings of the Workshop on Advances in Text Input Methods (WTIM), IJCNLP*.
2. **Arévalo, M., Carreras, X., Màrquez, L., Martí, M. A., Padró, L., & Simón, M. J. (2002).** A proposal for wide-coverage spanish named entity recognition. *Procesamiento del lenguaje natural*, Vol. 28, pp. 63–80.
3. **Ashwini, S. & Choi, J. D. (2014).** Targetable named entity recognition in social media. *arXiv preprint arXiv:1408.0782*.

4. **Banerjee, S., Kuila, A., Roy, A., Naskar, S. K., Rosso, P., & Bandyopadhyay, S. (2014).** A hybrid approach for transliterated word-level language identification: Crf with post-processing heuristics. *Proceedings of the Forum for Information Retrieval Evaluation (FIRE)*, ACM, pp. 54–59.

5. **Banerjee, S., Naskar, S. K., & Bandyopadhyay, S. (2014).** Bengali named entity recognition using margin infused relaxed algorithm. *17th International Conference on Text, Speech and Dialogue (TSD)*, volume 8655, Springer, pp. 125.

6. **Banerjee, S., Naskar, S. K., Rosso, P., & Bandyopadhyay, S. (2016).** The first cross-script code-mixed question answering corpus. *Proceedings of the workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016), co-located with ECIR*.

7. **Cohen, J. (1960).** A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46.

8. **Cortes, C. & Vapnik, V. (1995).** Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297.

9. **Crammer, K. & Singer, Y. (2003).** Ultraconservative online algorithms for multiclass problems. *The Journal of Machine Learning Research*, Vol. 3, pp. 951–991.

10. **Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2015).** Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, Vol. 51, No. 2, pp. 32–49.

11. **Ekbal, A. & Bandyopadhyay, S. (2007).** Maximum entropy approach for named entity recognition in bengali. *Symposium on Natural Language Processing (SNLP)*.

12. **Ekbal, A. & Bandyopadhyay, S. (2008).** A web-based bengali news corpus for named entity recognition. *Language Resources and Evaluation*, Vol. 42, No. 2, pp. 173–182.

13. **Ekbal, A., Naskar, S. K., & Bandyopadhyay, S. (2007).** Named entity recognition and transliteration in bengali. *Lingvisticae Investigationes*, Vol. 30, No. 1, pp. 95–114.

14. **El Bazi, I. & Laachfoubi, N. (2015).** Rena: A named entity recognition system for arabic. *18th International Conference on Text, Speech and Dialogue (TSD)*, Springer International Publishing, pp. 396–404.

15. **Etter, D., Ferraro, F., Cotterell, R., Buzek, O., & Van Durme, B. (2013).** Nerit: Named entity recognition for informal text. *Human Language Technology Center of Excellence, Johns Hopkins, vol. Technical Report*, Vol. 11.

16. **Finkel, J. R., Grenager, T., & Manning, C. (2005).** Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 363–370.

17. **Jamatia, A. & Das, A. (2014).** Part-of-speech tagging system for indian social media text on twitter. *Proceedings of workshop on Language Technology of Indian Social Media Text (Social-India 2014)*, ICON.

18. **Lafferty, J., McCallum, A., & Pereira, F. C. (2001).** Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *18th International Conference on Machine Learning*.

19. **Li, C. & Sun, A. (2014).** Fine-grained location extraction from tweets with temporal awareness. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, pp. 43–52.

20. **Li, C., Sun, A., Weng, J., & He, Q. (2015).** Tweet segmentation and its application to named entity recognition. *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 27, No. 2, pp. 558–570.

21. **Mansour, R., Refaei, N., & Murdock, V. (2014).** Augmenting business entities with salient terms from twitter. *25th International Conference on Computational Linguistics (COLING)*, pp. 121–129.

22. **McCallum, A., Freitag, D., & Pereira, F. C. (2000).** Maximum entropy markov models for information extraction and segmentation. *International Conference on Machine Learning (ICML)*, volume 17, pp. 591–598.

23. **Muysken, P. (2000).** *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

24. **Onal, K. D. & Karagoz, P. (2015).** Named entity recognition from scratch on social media. *Proceedings of 6th International Workshop on Mining Ubiquitous and Social Environments (MUSE), co-located with the ECML PKDD*.

25. **Peng, N. & Dredze, M. (2015).** Named entity recognition for chinese social media with jointly trained embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 548–554.

26. **Ritter, A., Clark, S., Etzioni, O., et al. (2011).** Named entity recognition in tweets: an experimental study. *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1524–1534.

27. **Skiba, R. (2005).** Code switching as a countenance of language interference. *The Internet TESL Journal*.

28. **Wu, Y., Zhao, J., & Xu, B. (2003).** Chinese named entity recognition combining a statistical model with human knowledge. *Proceedings of workshop on Multilingual and mixed-language named entity recognition*, ACL, pp. 65–72.

Article received on 19/12/2016; accepted on 28/02/2017.
Corresponding author is Somnath Banerjee.

doi: 10.13053/CyS-21-4-2850