The First Cross-Script Code-Mixed Question Answering Corpus

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

[1] Computer Science and Engineering Department, Jadavpur University, India
sb.cse.ju@gmail.com,{sudip.naskar,sbandyopadhyay}@cse.jdvu.ac.in
[2] PRHLT Reearch Center, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. In this paper, we formally introduce the problem of cross-script code-mixed question answering (QA) and we elaborate the corpus acquisition process and an evaluation strategy related to the said problem. Today social media platforms are flooded by millions of posts everyday on various topics. This paper emphasizes the use of such ever growing user generated content to serve as information collection source for the QA task on a low-resource language for the first time. A majority of these posts are multilingual in nature and many of them involve code mixing. The multilingual aspect of social media content is reflected in the use of multilingual words as well as in the writing script. For the ease of use multilingual users often pose questions in non-native script. Focusing on this current multilingual scenario, code-mixed cross-script (i.e., non-native script) data give rise to a new problem and present serious challenges to automatic QA. In the work presented in this paper, Bengali is considered as the native language while English is considered to be the non-native language. However, the dataset construction approach presented in this paper is generic in nature and could be used for any other language pair. Apart from introducing this novel problem, this paper highlights corpus development process and a suitable evaluation framework.

Keywords: Question Answering, Code Mixing, Code Switching, Cross-script, social media

1 Introduction and Related Work

Code-mixing refers to the phenomenon where lexical items and grammatical features from two languages appear in one sentence. The use of code-mixing is spreading widely in informal text communications such as newsgroups, tweets, blogs, and other social media platforms. Sometimes it is used to refer to relatively stable informal mixtures of two languages, such as Spanglish, Franponais or Portuñol. Nowadays in social media people tend to share everything under the sun. Social media users often share their travel experiences as well as seek

travel suggestions from their social networks. Similarly sports events are among the mostly discussed topics in social media. People post live updates of ongoing sports events such as Football World Cup, Champions League, T20 Series, etc. This results in potentially rich resources for languages which are less computerized.

In bilingual or multilingual countries like India, speakers often incorporate lexical items, phrases, and clauses from more than one language into their spoken or written communication act. This results in words or phrases from different languages in the same sentence or utterance. This phenomenon is referred to as code-mixing. Although this phenomenon has been studied extensively in formal and spoken context, the research community in natural language processing (NLP) has just started paying sincere attention to code-mixing due to its prevalence of use in electronic communication mainly in the social media. English is predominantly the most used language on the internet; Indians also use English extensively while surfing the internet. Even they (phonetically) use the Roman script instead of using their own native scripts. Another important reason behind the use of the English language and the Roman script may be the keyboards which are in the non-native Roman script, and Indian internet users are more comfortable using that keyboard rather than the on-screen native script keyboard or a combination of keys which generate native alphabets. Every natural language is generally written using a particular script which is referred to as the native script for that language. All other scripts which are not used in writing the language can be referred to as the non-native script with respect to that language. For example, the English language is written in the Roman script. Thus, Roman script is the native script for English, however Bengali script is a non-native script for English. We refer to the phenomenon of using a non-native script phonetically for writing native words as cross-script. For example, if a Bengali user writes Bengali words in Bengali script, that is considered as 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.

Being a classic application of NLP, QA has practical applications in various domains such as education, health care, personal assistance, etc. Presently, QA is a well addressed research problem and several QA systems are available with reasonable accuracy. A number of QA systems were developed for European languages particularly for English ([1], [2],[3],[4]), Middle Eastern languages ([5],[6],[7]) and Asian languages, e.g., Japanese ([8],[9]) Chinese ([10],[11]). In this paper, we introduce a new research problem in the context of QA research cross-script code-mixed QA.

The rest of the paper is organized as follows. Section 2 states the code-mixed cross-script QA problem. We discuss corpus acquisition in Section 3. The proposed corpus annotation process and corpus statistics are described in Section 4 and Section-5, respectively. We present the evaluation scheme in Section 6. Section 7 concludes the paper.

2 Problem Statement

Problem Statement: *Building a question answering system which takes cross-script (non-native) code-mixed questions as information request, processes a cross-script code-mixed text corpus and provides an (or a list of) exact answer(s) as information response.*

We introduce this novel research problem for the following reasons:

1. Multilingual non-native English speakers predominantly use the Roman script in social media platforms during their conversations even while the written communication takes place entirely in a native language (i.e., not English).
2. To make the written communication more fascinating, borrowing foreign words from different languages is very common in social media communication and this is a growing trend.
3. The ever increasing posts in many less-computerized languages could serve as a potential source of digital content for language research.
4. The research community need to move towards the next generation search engine that boosts the necessity of developing QA system for less-resourced languages.

This paper presents a cross-script code-mixed QA corpus for Bengali; however, this context is very common with other non-English languages, e.g. Spanish, French, etc. Despite the advances in QA research and the fact that Bengali is one of the most spoken languages, very little work ([12],[13],[14]) has been conducted in QA for Bengali so far. Language identification in the code-mixing scenario has been addressed extensively in shared tasks in EMNLP-2014[3] and FIRE-2014[4] and in few other research works [15],[16],[17],[18]. However, to the best of our knowledge, no work has been conducted so far on the novel problem addressed in this paper.

3 Corpus Acquisition

Because of the following characteristics of social media, we consider social media content for code-mixing cross-script QA corpus:

i) Substantial and ever increasing user base.

ii) A sizable volume of informal text data are added on various domains on a daily basis.

iii) Various APIs are available to access social media data.

iv) Most likely source of getting code-mixed data.

Even though acquiring a sizable volume of the code-mixed cross-script data is not a tough task, our work on developing a QA system for code-mixed cross-script data is at its initial stages. Therefore, we have collected a small set of data which could be increased in future following with a similar approach. Research

[3] http://emnlp2014.org/workshops/CodeSwitch/call.html
[4] http://fire.irsi.res.in/fire/home

in QA system primarily requires three data resources: (i) question which is asked to get a piece of information, (ii) answer to an asked question as a response, and (iii) potential sources of the answers from which a QA system can directly or indirectly infer an answer to a question. We describe the acquisition of these resources in this section. For the present study, we restricted our focus to the tourism and the sports domains which are among the most popular domains in the social media. Social media data on other domains could be acquired with a similar approach presented here. In the code-mixed cross-script QA scenario, the resource development involves two separate processes: (i) collecting social media text for the desired domains; and (ii) question acquisition and answer annotation.

3.1 Message(text) Acquisition

For the document collection we consider the social media as it is the most likely potential source of code-mixed cross-script data. We acquired all the messages from different social media platforms, e.g., twitter, blogs, forums, etc. For the sports domain, we selected social media posts on recently held 10 exciting cricket matches. Ten popular tourist spots in India were selected for tourism domain. Tweepy API and an in-house focused crawler were employed for collecting tweets, blogs, and forum posts. For collecting only code-mixed data, we set a language mix ratio (i.e., non-native:native) which is computed by employing a language identifier whose accuracy, as reported in [19], is 92.4%. Language mixing ratio (LMR) is employed for collecting only code-mixed data. The language mixing ratio has been set to 0.2 after manually verifying a small set of crawled data. Therefore, a message post is included in the corpus when at least 16.67% (i.e. 1 in 6) of the words belong to the non-native language.

Examples of valid Message:

a) Message: SA\O ja\B run\E koreche\B aj\B BD\O parbe\B ki\B ?\O

LMR = $\frac{\#non-native}{\#native} = \frac{\#English-words}{\#Bengali-words} = \frac{1}{5} = 0.2(>= 0.2)$

b) Message: Mashrafe\O well\E try\E but\E ki\B r\B kora\B jabe\B ...\O captain\E !!!\O

LMR = $\frac{\#non-native}{\#native} = \frac{\#English-words}{\#Bengali-words} = \frac{4}{4} = 1(>= 0.2)$

The language identifier, as reported in [19], does not identify named entities. Considering the fact that the answer to a factoid question is always a named entity, we filtered out the messages under human supervision which do not contain any named entity. Thus, we finalized 299 posts as messages out of the 334 messages which were initially selected by the language identifier and the LMR ratio.

3.2 Question Acquisition

The question preparation task is more challenging than the message acquisition and requires more human involvement. Our prime target was to involve as many question setters as possible to reduce bias. A cloud-based service was

built and requests were sent to the undergraduate students of the university. Two groups, namely sports-domain group (SG) and tourism-domain group (TG) with 15 students each were formed from thirty students who agreed for the question annotation task. Ten topics on sports domain were provided to each member of SG and they were asked to submit at least 10 questions on each topic. The submitted questions were stored in the web server along with the messages associated with the topic. After receiving these questions, we kept only the questions having code-mixed nature and satisfying the LMR criterion. Subsequently, the annotators were asked to find out the answer to their legitimate questions from the stored messages. An analogous procedure was followed for TG also.

4 Annotation

For document management and storing, EXtensible Markup Language (XML) was chosen because of its popularity and ease of understanding. The QA annotation framework which was adopted in this work is depicted in Fig. 1. The tagset defined in Table 1 was used for three purposes: document information, message annotation and QA annotation. We will format the corpus in Text Encoding Initiative[5](TEI) in future.

Table 1. Corpus tagset

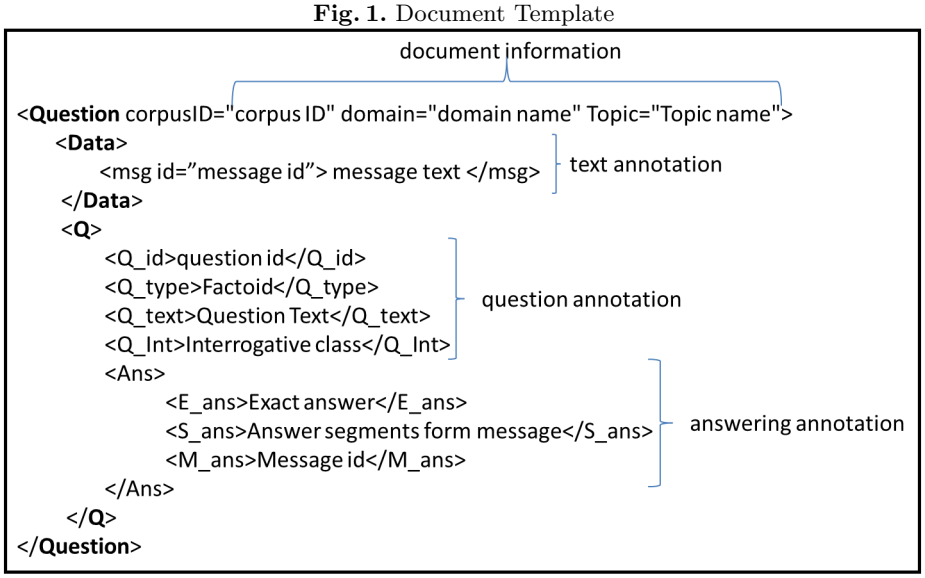
Tag	**Definition**	**Tag**	**Definition**
Question	Document body	CorpusID	Corpus id number
Domain	Domain name	Topic	Topic name
Data	Data section	Q	Question
Q_id	Question unique number	Q_type	Question type, e.g., Factoid, Procedural
Q_text	Code-mixed NL question	Q_Int	Interrogative class
Ans	Answer	E_ans	Exact answer
S_ans	Segment answer	M_ans	Message Id of a message that contains answer
Msg	Public posts as messages		

A document in the corpus comprises of data section and question section. The data section contains the public posts collected from social media. Each public post is referred to as a message and described in the $< msg >$ tag. Each message is assigned a unique number, i.e., msg_Id. The factoid questions follow the data section. Each question is marked by the Q tag, (i.e., $< Q >$ and $< /Q >$). Like each message, every question is also assigned a unique question identifier. The question type (*Q_type*) denotes the type of a question such as factoid, procedural, etc. The code-mixed cross script question is enclosed by the *q_text* tag.

Interrogative types of questions are very much useful for answer extraction and validation. On the basis of syntactic structure, Bengali interrogatives are

[5] http://www.tei-c.org/index.xml

Fig. 1. Document Template



classified into three categories - single interrogative (SI), dual interrogative (DI) and compound interrogative (CI) [12]. The interrogative type (i.e., SI, DI, and CI) of a question gives a clue about the number information of the candidate answer.

The answer to a question is annotated by the Ans tag. The exact answer is given in *E_ans* tag. The Segment answer (*S_ans*) tag refers to the portion or segment of the message text which provides the answer. The message id from which the exact answer can be found is given in the message answer (*M_ans*) tag. The segment answer tag and message tag could be thought of as supporting information for the exact answer.

5 Corpus Statistics

The statistics of the messages, i.e., public posts and questions in the corpus for the two different domains, namely Sports and Tourism, are given in Table 2. Altogether 299 code-mixed cross-script messages were collected of which 183 and 116 messages are from the tourism and sports domains respectively. 506 code-mixed cross-script questions were acquired of which 314 questions are from the tourism domain and 192 questions belong to the sports domain. Average number of messages per document (Avg. M/D in Table 2)is higher for the tourism domain than for the sports domain. Average number of questions generated per document (Avg. Q/D in Table 2) is higher for the tourism domain than for the sports domain accordingly.

Table 2. Corpus statistics

Domain	Documents(D)	Messages(M)	Questions(Q)	Avg. M/D	Avg. Q/D
Tourism	10	183	314	18.3	31.4
Sports	10	116	192	19.2	19.2
Overall	20	299	506	14.95	25.3

6 Proposed Evaluation

Along with the corpus development, we also propose an evaluation scheme to evaluate the code-mixed QA performance which is suitable to our corpus annotation. In the annotated corpus an answer is basically structured as [*Answer String* (AS), *Message Segment* (MS), *Message ID* (MId)] triplet, where-

– AS is the one of the exact answers (EA) and must be an NE in this case,
– MS is the supported text segment for the extracted answer, and
– MId is the unique identifier of the message that justifies the answer.

The evaluation methodology was designed taking into consideration the following issues:

i) The QA system has the provision of not answering, i.e., no answer option (NAO).

ii) The answer returned should be the exact answer to the question.

iii) The exact answer must be a Named Entity.

iv) The system has to return a single exact answer. In case there exists more than one correct answer to a question, the system needs to provide only one of the correct answers.

While designing the evaluation strategy, our primary focus was on "responsiveness" and "usefulness" of each answer. Each answer has to be manually judged by native speaking assessors. Each answer [AS, MS, MId] triplet is assigned a score in a five-valued (range 0.0-1.0) scale which is weighted correctness measure using hard-coded weights and marked with exactly one of the following judgments depicted in Table 3:

– **Incorrect:** The AS does not contain EA (i.e., responsive but not useful)
– **Unsupported:** The AS contains correct EA, but MS and MId do not support the EA (i.e., missing usefulness)
– **Partial-supported:** The AS contains the correct EA with correct MId, but MS does not support EA
– **Correct:** The AS provides the correct EA with correctly supporting MS and MId (i.e., "responsive" as well as "useful").
– **Inexact:** The supporting MS and MId are correct, but the AS is wrong.

The QA evaluation forums such as TREC[6], CLEF[7], etc. proposed accuracy, c@1[20], and Mean reciprocal rank (MRR) [21] as evaluation metrics for the

[6] http://trec.nist.gov/
[7] http://www.clef-initiative.eu/

Table 3. Judgment Scale

Judgment	**AS**	**MS**	**MId**	**Score**
Incorrect (W)	X	X	X	0.00
Inexact(I)	X	✓	✓	0.25
Unsupported (U)	✓	X	X	0.50
Partial-supported (P)	✓	X	✓	0.75
Correct (C)	✓	✓	✓	1.00

monolingual and cross-lingual QA. In order to maintain the consistency with the state-of-the-art QA evaluation metrics, we also suggest the use of accuracy and c@1 for the code-mixed cross-script QA task. As the prepared corpus contains only one correct answer (as opposed to a list of exact answers) for every question, MRR is not useful for evaluation on the said dataset. Just as in the past ResPubliQA[8] campaigns, systems have the option of withholding the answer to a question because they are not sufficiently confident that it is correct (i.e., NAO). As per ResPubliQA, the inclusion of NAO improves the system performance by reducing the number of incorrect answers.

Now, C@1 = $\frac{1}{N}(N_r + N_u \cdot \frac{N_r}{N_u})$

Accuracy = $\frac{N_r}{N}$

C@1 = Accuracy; if $N_u = 0$

Where, N_r = number of right answers.

N_u = number of unanswered questions

N = total questions

Correct, Partially-supported and Unsupported answers provide the exact answers only.

Therefore, $N_r = (\#C + \#U + \#P)$

Considering the importance of supporting segment, we introduce a new metric "answer-support performance" (ASP) which measures the answer correctness and which is defined as follows:

$ASP = \frac{1}{N}(c \times 1.0 + p \times 0.75 + i \times 0.25)$

where, c, p and i denote total number of correct, partially-supported and inexact answers respectively.

7 Conclusions

In this paper we presented a novel research problem - cross-script code-mixed QA. Our major contributions include (i) proposing an annotation scheme, ii) creating a dataset which is the first resource of its kind, and (iii) proposing an evaluation strategy that is suitable to our corpus annotation. Bearing in mind the small dataset, the proposed evaluation methodology and created dataset will be helpful for the QA research and development community, particularly those who want to address code-mixed cross-script QA.

[8] http://nlp.uned.es/clef-qa/repository/resPubliQA.php

Acknowledgements

We acknowledge the support of the Department of Electronics and Information Technology (DeitY), Government of India, through the project "CLIA System Phase II". The work of the third author was in the framework of the SomEMBED MINECO TIN2015-71147-C2-1-P research project.

References

1. Buscaldi, D., Rosso, P., Gómez, J.M., Sanchis, E.: Answering Questions with an n-gram based Passage Retrieval Engine. In: Journal of Intelligent Information Systems, 34:113-134 (2010)
2. Brill, E., Dumais, S., Banko, M.:. An analysis of the AskMSR question-answering system. In: Empirical methods in natural language processing-Volume 10, pp. 257-264, Association for Computational Linguistics (2002)
3. Zheng, Z.: AnswerBus question answering system. In: International conference on Human Language Technology Research, pp. 399-404, Morgan Kaufmann Publishers Inc. (2002)
4. Ittycheriah, A., Franz, M., Zhu, W. J., Ratnaparkhi, A., Mammone, R. J.: IBM's Statistical Question Answering System. In: TREC (2000)
5. Mohammed, F. A., Nasser, K., Harb, H. M.: A knowledge based Arabic question answering system (AQAS). In: ACM SIGART Bulletin, 4(4), 21-30 (1993)
6. Kanaan, G., Hammouri, A., Al-Shalabi, R., Swalha, M.: A new question answering system for the Arabic language. In: American Journal of Applied Sciences, 6(4), 797 (2009)
7. Hammo, B., Abu-Salem, H., Lytinen, S.: QARAB: A question answering system to support the Arabic language. In: ACL-02 workshop on Computational approaches to semitic languages, pp. 1-11, Association for Computational Linguistics (2002)
8. Sakai, T., Saito, Y., Ichimura, Y., Koyama, M., Kokubu, T., Manabe, T.: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis. In: RIAO, pp. 215-231 (2004)
9. Isozaki, H., Sudoh, K., Tsukada, H.: NTTs japanese-english cross-language question answering system. In: NTCIR Workshop 5 Meeting, pp. 186-193 (2005)
10. Yongkui, Z. H. A. N. G., Zheqian, Z. H. A. O., Lijun, B. A. I., Xinqing, C. H. E. N.: Internet-based Chinese Question-Answering System. In: Computer Engineering, 15 (2003)
11. Sun, A., Jiang, M., He, Y., Chen, L., Yuan, B.: Chinese question answering based on syntax analysis and answer classification. In: Acta Electronica Sinica, 36(5) (2008)
12. Banerjee, S., Bandyopadhyay, S.: Bengali Question Classification: Towards Developing QA System. In: SANLP-COLING, IIT,Mumbai,India (2012)
13. Banerjee, S., Lohar, P., Naskar, S. K., Bandyopadhyay, S.: The First Resource for Bengali Question Answering Research. In: PolTAL-2014. Poland. In Advances in Natural Language Processing, pp. 290-297. Springer International Publishing (2014)
14. Banerjee, S., Naskar, S. K., Bandyopadhyay, S.: BFQA: A Bengali Factoid Question Answering System. In: Text, Speech and Dialogue (TSD), pp. 217-224. Springer International Publishing, Czech Republic (2014)
15. Gupta, P., Bali, K., Banchs, R., Choudhury, M., Rosso, P.: Query Expansion for Mixed-script Information Retrieval. In: The 37th Annual ACM SIGIR Conference, SIGIR-2014, Gold Coast, Australia, June 6-11, pp. 677-686 (2014)

16. King, B., Abney, S.: Labeling the languages of words in mixed-language documents using weakly supervised methods. In: NAACL-HLT, pages 1110–1119 (2013)
17. Barman, U., Wagner, J., Chrupala, G., Foster, J.: Identification of languages and encodings in a multilingual document. In: EMNLP (2014)
18. Choudhury, M., Chittaranjan, G., Gupta, P., Das, A.: Overview of FIRE 2014 Track on Transliterated Search. In: FIRE (2014)
19. Banerjee, S., Kuila, A., Roy, A., Naskar, S. K., Bandyopadhyay, S., Rosso, P.: A Hybrid Approach for Transliterated Word-Level Language Identification: CRF with Post Processing Heuristics. In: Forum for Information Retrieval Evaluation, pp. 54-59, ACM Digital Publication (2014)
20. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011). Portland, Oregon, USA (2011)
21. Voorhees, E.M.: The TREC-8 question answering track report. In: 8th Text Retrieval Conference (TREC), Gaithersburg, Maryland, USA, pp. 77-82 (1999)