WebShodh: A Code Mixed Factoid Question Answering System for Web

Khyathi Raghavi Chandu[1(✉)], Manoj Chinnakotla[2], Alan W. Black[1], and Manish Shrivastava[3]

[1] Carnegie Mellon University, Pittsburgh, USA
{kchandu,awb}@cs.cmu.edu
[2] Microsoft India, Hyderabad, India
manojc@microsoft.com
[3] IIIT Hyderabad, Hyderabad, India
m.shrivastava@iiit.ac.in

Abstract. Code-Mixing (CM) is a natural phenomenon observed in many multilingual societies and is becoming the preferred medium of expression and communication in online and social media fora. In spite of this, current Question Answering (QA) systems do not support CM and are only designed to work with a single interaction language. This assumption makes it inconvenient for multi-lingual users to interact naturally with the QA system especially in scenarios where they do not know the right word in the target language. In this paper, we present *WebShodh* - an end-end web-based Factoid QA system for CM languages. We demonstrate our system with two CM language pairs: *Hinglish* (Matrix language: Hindi, Embedded language: English) and *Tenglish* (Matrix language: Telugu, Embedded language: English). Lack of language resources such as annotated corpora, POS taggers or parsers for CM languages poses a huge challenge for automated processing and analysis. In view of this resource scarcity, we only assume the existence of bi-lingual dictionaries from the matrix languages to English and use it for lexically translating the question into English. Later, we use this loosely translated question for our downstream analysis such as Answer Type(AType) prediction, answer retrieval and ranking. Evaluation of our system reveals that we achieve an MRR of 0.37 and 0.32 for Hinglish and Tenglish respectively. We hosted this system online and plan to leverage it for collecting more CM questions and answers data for further improvement.

1 Introduction

CM is the phenomenon of "embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language" [1]. The lexicon and syntactic formulations from both the languages are mixed to form a single coherent sentence. Some of such mixtures are known as Spanglish, Hinglish, Tenglish, Portunol and Franponaisor[1]. CM usually prevails in a multilingual

[1] Mixing of Spanish-English, Hindi-English, Telugu-English, Portugese-Spanish and French-Japanese language pairs respectively.

G.J.F. Jones et al. (Eds.): CLEF 2017, LNCS 10456, pp. 104–111, 2017.
DOI: 10.1007/978-3-319-65813-1_9

configuration with speakers having more than one common language. Moreover, anglicization of languages is also a very common phenomenon these days, which leads to the representation of native words in English letters phonetically. The study on cross script code mixing is essential mainly because of the prominent usage of English keyboards in countries like India. Studies on statistical usage of code-switching among facebookers found that there is about 33% of intra-sentential switching [2]. This work also showed that 45% of switching is due to real lexical need, which is a considerably high percentage. The increasing use of CM is also driven by the ease and speed of communication mainly facilitated by the easier choice of words and a richer set of expressions to choose from. In spite of this, current QA systems [3,4] only support interaction in a single language. This severely hampers the ability of a multi-lingual user to interact naturally with the QA system. This is especially true in scenarios involving technical and scientific terminology. For example, when a native Telugu speaker wants to know the director of the movie *Heart Attack*, he is more likely to express it as "*heart attack cinema ni direct chesindi evaru?" (Translation: who directed the movie heart attack)* where the words *heart attack*, *direct*, *cinema* are all English words. Hence, to increase the reach, impact and effectiveness of QA in multi-lingual societies [5], it is imperative to support QA in CM languages [6]. However, any automated analysis and processing of CM text poses serious challenges due to lack of normalized representations adhering to standard syntactic and phonetic rules. The problem is further compounded by the unavailability of language resources such as annotated corpora, language analysis tools such as POS taggers, parsers *etc.*

In this paper, we present *WebShodh* - an end-to-end open domain factoid Question Answering (QA) system for Web which provides a *ranked list of potential answers* to a CM question. We demonstrate our system using CM in two dominantly spoken languages in India; Hindi and Telugu[2]. In view of resource-scarcity, we only assume the existence of bi-lingual dictionaries from the source language to English. Our system performs a lexical level language identification and translation into English. We use this high-level loosely translated question to classify and infer the expected answer type. We also fire the entire loosely translated English query to Google using their Search API and retrieve the top 10 search results from the web along with their titles and snippets, which are then processed to identify potential candidates for answers based on the hints offered by AType. Finally, we rank these candidate answers based on various features to finally output a ranked list of answers. We evaluated our system on both these CM languages and share the quantitative and qualitative analysis of our results. Overall, our system achieves an MRR of 0.37 and 0.32 for Hinglish and Tenglish respectively. We hosted our system *WebShodh* online (http://128.2.208.89/webshodh/cmqa.php) and intend to use it for collecting more

[2] Hindi is one of the most spoken languages in India, with 370 million native speakers and is an official language along with English. Telugu is the most spoken Dravidian language in South India with about 70 million native speakers.

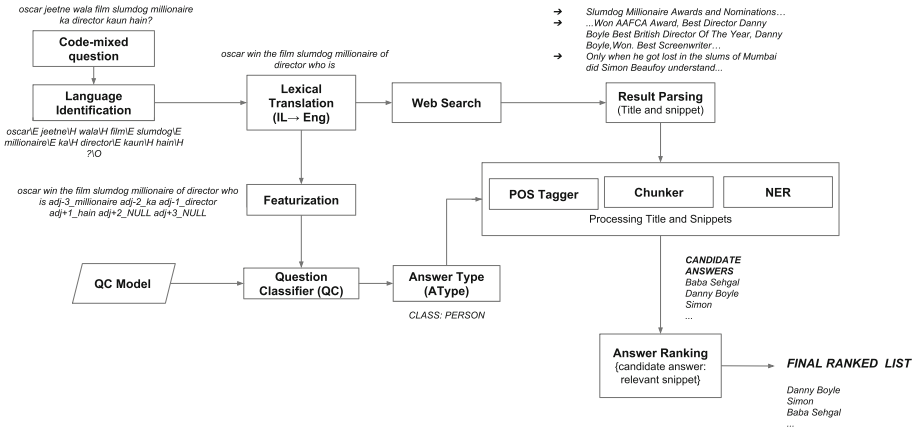

Fig. 1. Architecture of *WebShodh*: a web based factoid QA system for code-mixed languages

QA data for CM languages - an important step forward if we want to try out more data-intensive techniques such as deep learning.

The paper is organized as follows: in Sect. 2, we discuss the related previous work in this area. Section 3 describes the overall system architecture and delves into each of the steps in the pipeline. In Sect. 4, we present the experimental setup including data creation, experimental results and qualitative error analysis. Section 5 discusses the conclusions and future scope of the work.

2 Related Work

Linguistic and conversational motives for CM have been studied in [7–9]. [10] describes the grammatical contexts in which CM has taken place in student interactions. The recent years have shown rapid upsurge in understanding and analyzing these languages as they are among the most prominently used languages on social media. The intuitive first step towards tackling this domain is lexical language identification, which has been addressed in EMNLP[3] and FIRE[4] in 2014. The challenges of this non-trivial task have been presented by [11]. [12] have studied POS tagging in code-mixed social content and have concluded that the tasks of language identification and transliteration still stand as major challenges. Question Classification (QC) and Question Answering (QA) systems have been well studied for monolingual settings previously by [13–16]. [17] have introduced the space of QC in code-mixed languages. This work used an SVM based QC technique and presented results for coarse and fine grained categorizations, based on ontology of question hierarchy described by [16]. While this work was mainly done for Hindi-English pair, it was later studied for Bengali-English by

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

[18]. [19] have presented an approach to mine the ever growing content on social media for generating a CM QA corpus in Bengali-English which contains both CM questions and answers and also proposed an evaluation strategy using the corpus.

To the best of our knowledge, our system is the first end-end factoid QA system designed specifically for CM questions.

3 Web Based Code-Mixed QA System

In this section, we describe the details of our system - *WebShodh*. This system is hosted at http://128.2.208.89/webshodh/cmqa.php and is currently supporting Hinglish and Tenglish. A video demonstration of the working of *WebShodh* is available at https://www.youtube.com/watch?v=aVsZVfere5w[5]. Figure 1 presents the architecture of *WebShodh* along with an illustrated example Hinglish CM question *"Oscar jeetne wala Slumdog Millionaire film ka director kaun hain?" (Translation: who is the director of the oscar award winning film "Slumdog Millionaire"?*). Given a natural language question expressed in CM, it was passed through a language identification module from [20]. The principal idea is to lexically translate this question into English so that - (a) we can leverage monolingual resources in English, which is a resource rich language for subsequent processing (b) quality of web search in English is better compared to that in the matrix languages.

Question Classification: The complexities of identifying the Answer Type (AType) for CM questions are discussed in [17] and they also propose a technique for SVM based AType classification. Given the translated CM questions, they use a featurizer to create a bag of features consisting of lexical level features along with the adjacent words of 'Wh-' word, for representing the query. This is passed through an SVM based Question Classifier (QC) which classifies the CM question into one of the given types such as - HUMAN, LOCATION, ENTITY, ABBREVIATION, DESCRIPTION and NUMERIC, the type hierarchy defined by [16]. In this work, we just consider the coarse-grained categories for AType classification since the training data is too sparse in fine-grained category.

In this work, we extended the work done by [17] by including additional class of features to the SVM model. To improve the generalization capability, POS tags features of the words from the respective languages that are identified lexically are used. In addition, pre-trained embeddings from Google news vectors for each of the lexically translated words are used. We considered 10 representative samples from each AType. Later, we compute the centroids for each AType in this 300-dimensional space. For each of the adjacent words on both sides of 'Wh-' word, we get their word2vec embedding, calculate distance with the AType centroids and find out the closest AType to include that as a feature. Besides this, we performed a five-fold cross validation with a grid search for tuning the kernel

[5] This video is recorded in real time frame to demonstrate the speed of the system for practical purposes.

and C parameters in SVM. We used an RBF kernel with *gamma* value set to the inverse of feature vector size for better performance. Due to the above changes, we improved the overall accuracy of the QC system across the 6 categories from 63% to 71.96%.

Retrieval of Web Results: We submit the loosely translated English question as a search query to Google using Search API and retrieve the top 10 relevant documents along with their titles, URLs and snippets. "Snippet-tolerant property" [21] is leveraged to arrive at the answer by exploiting the information present in the relevant snippets, as processing the entire document is computationally expensive and time consuming.

Candidate Answer Generation: We run POS tagging, chunking and Named Entity Recognizer tools on the retrieved snippets and titles. The categories of NER are mapped to QC categories, based on which the relevant candidate answers are retrieved. We filter only the words and phrases whose NER tags map to the given AType and pass them to the next phase as candidate answers.

Answer Ranking: For each candidate answer, its relevance score is computed by adding the cosine similarity between the translated CM question and all congregated titles and snippets where the candidate answer occurs. The final list of answers is displayed in a ranked order according to the above relevance score. Redundancy of the correct answer, which occurs in multiple relevant documents, potentially improves its ranking score. But NER on huge text introduces latency in the pipeline. Hence we need to decide on an appropriate trade-off between them.

4 Evaluation Dataset and Results

We used *WebShodh* - our end-end open domain CM QA system for also collecting the evaluation data. We took the help of 10 native speaker volunteers each for Hindi and Telugu languages. All of them were bi-lingual speakers who were also fluent in English. We gave them access to the web interface of our system and requested them to try out at least 10 factoid questions of their choice.

A maximum of 10 ranked answers for each of the CM factoid question are displayed. As a part of the feedback process, the user was asked to select the correct answer and submit it to the system. Through this, we are collecting the data corresponding to a question, its answer along with the answer rank. In this way we have collected 100 questions for each language pair. The details of this evaluation dataset is given in Table 1. The user feedback on question category was purposefully omitted from the interface as there is certain domain knowledge involved in annotating question types, which the users may not be aware of. The data obtained through this platform offers a huge potential to improve CM QA further and hence the system is hosted online. Language Mix Ratio (LMR) is the ratio of the number of words from Embedded language to the total number of words in the sentence. From Table 1, we can observe that on an average, LMR is 0.3937 and 0.3973 respectively for Hinglish and Tenglish CM questions.

QUESTION (HINDI-ENGLISH CM)	QC LABEL	ANSWER	ANALYSIS
Question: oscar jeetne wala film slumdog millionaire ka director kaun hain? **Gloss Translation:** oscar won of film slumdog millionaire of director who is ? **Meaning:** Who is the director of the oscar winning film Slumdog Millionaire?	**Predicted:** Human **Correct:** Human	**Given answer:** Danny Boyle **Correct answer:** Danny Boyle	'Danny Boyle' was provided redundantly in the candidate sentences and was tagged appropriately as PERSON in NER and hence is the highest scored answer. Entity Normalization could further increase this score.
Question: world war 1 kis saal mein shuru hua hain? **Gloss Translation:** world war 1 which year in begin happen is ? **Meaning:** In which year did World War 1 begin?	**Predicted:** Entity **Correct:** Numeric	**Given answer:** 1914 **Correct answer:** 1914	Misclassification as entity leads to candidate answers as all noun phrases and the year '1914' had a POS tag of NP and hence retrieved as highest ranked answer..
Question: acetyl salicylic acid ka doosra naam kya hain? **Gloss Translation:** acetyl salicylic acid of second name what is ? **Meaning:** What is another name of acetyl salicylic acid ?	**Predicted:** Human **Correct:** Entity	**Given answer:** Not found **Correct answer:** aspirin	Most of the examples in training data annotated with 'HUMAN' has adjacent word as 'name'. The misclassification lead to not identifying an answer.
Question: cheap thrills gana kis album se hain? **Gloss Translation:** cheap thrills song which album from is? **Meaning:** Which album does cheap thrills song belong to?	**Predicted**: Entity **Correct:** Entity	**Given answer:** Sia **Correct answer:** This is Acting	Correctly classified as entity. Goes through all noun phrases and explicit mention of the word 'album' is not present in the candidate sentences that increases the score of 'This is Acting'.

QUESTION (TELUGU-ENGLISH CM)	QC LABEL	ANSWER	ANALYSIS
Question: Dan Brown rasina modati pustakam lo protagonist evaru? **Gloss Translation:** Dan Brown written first book in protagonist who? **Meaning:** Who is the protagonist in Dan Brown's first book?	**Predicted:** Human **Correct:** Human	**Given answer:** Robert Langdon **Correct answer:** Robert Langdon	The question is correctly classified and the redundancy of the exact answer 'Robert Langdon' increased its score.
Question: Amnesty International yokka headquarters ekkada undi ? **Gloss Translation:** Amnesty International of headquarters where is? **Meaning:** Where is the headquarters of Amnesty International?	**Predicted:** Location **Correct:** Location	**Given answer:** Uganda **Correct answer:** London	'yokka' has been incorrectly classified as an English word. As a result of this it was not lexically translated to English. Hence accurate set of documents were not retrieved by the query.
Question: ee 19th century painter Marquesas Islands lo chanipoyaru? **Gloss Translation:** What 19th century painter Marquesas Islands in died? **Meaning:** Which 19th century painter died in Marquesas Islands?	**Predicted:** Entity **Correct:** Human	**Given answer:** Paul Gauguin **Correct answer:** Paul Gauguin	The question is misclassified as entity. Candidate answers include all Noun Phrases in this scenario and Paul Gauguin is an NP which is ranked highest.
Question: Japan lo highest point ekkada undi? **Gloss Translation:** Japan in highest point where is? **Meaning:** Where is the highest point in Japan?	**Predicted:** Location **Correct:** Location	**Given answer:** Shizuoka **Correct answer:** Mount Fuji	The correct answer is in the third position in the ranked list. Mount Fuji is in the border of Shizuoka. The extent of granularity of answer varies according to questions.

Fig. 2. Qualitative analysis of results with representative positive and negative examples

Table 1. CM QA Evaluation Dataset Details

Distribution parameters	Hinglish	Tenglish
Number of questions	100	100
Total number of words	833	667
Percentage of English words	39.37%	39.73%
Percentage of native words	60.62%	60.26%
Avg. CM words per question	5	4
Avg. length of questions	8	6

Table 2. Results of end to end WebShodh QA system

Metric	Hinglish	Tenglish
Precision at 1	0.37	0.32
Precision at 3	0.58	0.55
Precision at 5	0.67	0.65
Precision at 10	0.73	0.71
MRR	0.37	0.32

This section presents the quantitative and qualitative analysis of end-to end CM QA system. We use standard evaluation metrics such as precision at various ranks and Mean Reciprocal Rank (MRR) for measuring the effectiveness of our QA system. Table 2 shows the precision at 1, 3, 5 and 10 for both the language pairs along with their corresponding MRR. Figure 2 provides a qualitative analysis of our results for both the language pairs. This analysis is based on the following categories: (a) Both QC label and answer are correct and correct

answer is present at rank 1 (b) QC label is incorrect but the predicted answer is correct (c) QC label and answer are incorrect (d) QC label is correct but the answer predicted is incorrect.

5 Discussion and Conclusion

An accurate one to one mapping of alphabet does not exist across most languages that belong to different language families. This raises the issues of spelling variations while romanizing. This problem is commonly observed in the case of 'th' and 't'. Romanized Hindi and Telugu do not have specific environmental conditions or rules to use these letters and are often used interchangeably for wx notations of 't', 'T', 'w' and 'W'. The same problem is observed in the case of other hard and soft consonants. Similarly inconsistencies in representing long and short vowels usually cause errors in transliteration and thus the error is sent downstream to the task of translation. Consider the code-mixed question *'phata poster nikla hero movie lo protogonist evaru?'* (meaning: who is the protagonist in phata poster nikla hero movie?). Though the question itself is in Tenglish, *'phata poster nikla hero'* itself is a Hinglish code-mixed entity, which is the name of a movie. So the non-English words within the entity should be not be lexically translated to get the correct answer. Such entities need to be identified to avoid lexical translation.

Telugu is an agglutinative language which combines multiple morphemes to form a single word. *Sandhi* is the phenomenon of interplay of sounds at the boundaries of adjacent words leading to fusion and alteration of sounds, commonly observed in this language. For example, consider the word *'perenti'* in Telugu (meaning: what is the name) which is a frequently occurring word in Tenglish question dataset. It is a combination of two words *'peru'* (meaning: name) and *'enti'* (meaning: what) based on certain phonetic *sandhi* rules. It depends on the idiolect of the person on choosing to type *'peru enti'* or *'perenti'*. Hence the problem of dealing with code-mixing is compounded with noisy text. We plan to work on these issues further so that we can maximize the benefit of reaping bilingual dictionaries. We also plan to extend the system to Spanglish (code-mixing of Spanish and English) by building a cross script bilingual dictionary and language identification system. Unlike a pidgin, Spanglish could be the primary language of some people, mostly in the areas of Puerto Rico.

In conclusion, today's linguistically pluralistic societies need tools which support interaction in CM languages. In this paper, we presented *WebShodh* - an end-end web-based Factoid QA system for CM languages. We demonstrated our system with two pairs of CM languages - Hinglish and Tenglish. In view of resource scarcity, our system used very few resources such as bi-lingual dictionaries for these languages. We use Google Search API for retrieving the web results along with their snippets and titles. Evaluation of our system reveals that we achieve an MRR of 0.37 and 0.32 for Hinglish and Tenglish respectively. We hosted the system *WebShodh* online to collect more questions in order to understand the intricate variations of these newly formed languages in real world and

leverage it for collecting more CM question/answer data which is critical for future research and further improvement of the system.

References

1. Myers-Scotton, C., Linguistics, C.: Bilingual Encounters and Grammatical Outcomes. Oxford University Press, Oxford (2002)
2. Hidayat, T.: An Analysis of Code Switching used by Facebookers (2008)
3. Brill, E., Dumais, S., Banko, M.: An analysis of the AskMSR question-answering system. In: EMNLP-Volume 10 (2002)
4. Zhang, D., Lee, W.S.: A web-based question answering system (2003)
5. Magnini, B., et al.: Overview of the CLEF 2004 multilingual question answering track. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 371–391. Springer, Heidelberg (2005). doi:10.1007/11519645_38
6. Tay, M.W.J.: Code switching and code mixing as a communicative strategy in multilingual discourse. World Englishes **8**(3), 407–417 (1989)
7. Lesley, M., Pieter, M.: One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching. Cambridge University Press, Cambridge (1995)
8. Beatrice, A.: Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing. Dissertation, University of Edinburgh (2007)
9. Auer, P.: Code-Switching in Conversation: Language, Interaction and Identity (2013)
10. Dey, A., Fung, P.: A hindi-english code-switching corpus. In: LREC, pp. 2410–2413 (2014)
11. Barman, U., Das, A., Wagner, J., Foster, J.: Code mixing: a challenge for language identification in the language of social media. In: EMNLP (2014)
12. Vyas, Y., et al.: POS tagging of english-hindi code-mixed social media content. In: EMNLP, vol. 14, pp. 974–979 (2014)
13. Ferrucci, D., et al.: Building watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
14. Moschitti, A., et al.: Using syntactic and semantic structural kernels for classifying definition questions in Jeopardy! In: EMNLP, pp. 712–724 (2011)
15. Xu, J., Zhou, Y., Wang, Y.: A classification of questions using SVM and semantic similarity analysis. In: ICICSE, pp. 31–34 (2012)
16. Li, X., Roth, D.: Learning question classifiers. In: International Conference on Computational Linguistics-Volume 1, pp. 1–7 (2002)
17. Chandu, K.R., Chinnakotla, M., Shrivastava, M.: Answer ka type kya he? Learning to classify questions in code-mixed language. In: International Conference on World Wide Web, pp. 853–858. ACM (2015)
18. Majumder, G., Pakray, P.: NLP-NITMZ@ MSIR 2016 system for CodeMixed crossScript question classification. In: ECIR, pp. 7–10 (2016)
19. Banerjee, S., et al.: The first cross-script code-mixed question answering corpus. In: ECIR (2016)
20. Bhat, I.A., et al.: IIIT-H system submission for FIRE 2014 shared task on transliterated search. In: FIRE, pp. 48–53 (2014)
21. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 26–32 (2003)