Code-Mixed Question Answering Challenge: Crowd-sourcing Data and Techniques

[⋆]Khyathi Raghavi Chandu [⋄]Ekaterina Loginova
[△]Vishal Gupta [⋄]Josef van Genabith [⋄]Günter Neuman
[▽]Manoj Chinnakotla [⋆]Eric Nyberg [⋆]Alan Black

[⋆]Language Technologies Institute, Carnegie Mellon University
[⋄]Deutsche Forschungszentrum für Künstliche Intelligenz,[△]IIIT Hyderabad,[▽]Microsoft,USA
{kchandu, ehn, awb}@andrew.cmu.edu
{ekaterina.loginova, Josef.van_Genabith, neumann}@dfki.de
vishal.gupta@research.iiit.ac.in, manojc@microsoft.com

Abstract

Code-Mixing (CM) is the phenomenon of alternating between two or more languages which is prevalent in bi- and multi-lingual communities. Most NLP applications today are still designed with the assumption of a single interaction language and are most likely to break given a CM utterance with multiple languages mixed at a morphological, phrase or sentence level. For example, popular commercial search engines do not yet fully understand the intents expressed in CM queries. As a first step towards fostering research which supports CM in NLP applications, we systematically crowd-sourced and curated an evaluation dataset for factoid question answering in three CM languages - Hinglish (Hindi+English), Tenglish (Telugu+English) and Tamlish (Tamil+English) which belong to two language families. We share the details of our data collection process, techniques which were used to avoid inducing lexical bias amongst the crowd workers and other CM specific linguistic properties of the dataset. Our final dataset, which is available freely for research purposes, has 1,694 Hinglish, 2,848 Tamlish and 1,391 Tenglish factoid questions and their answers. We discuss the techniques used by the participants for the first edition of this ongoing challenge.

1 Introduction

Code-Mixing (CM) is formally defined as the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language, which is commonly observed in multilingual communities ((Myers-Scotton, 1997), (Poplack, 1980), (Muysken, 2000)). Traditionally, some studies (Yow and Patrycia, 2011) have viewed the mixing of two independent codes as lack of fluency of the segment of population in either of the languages. However, an alternate perspective (Milroy and Muysken, 1995) argues that mixing of two traditionally isolated linguistic codes potentially creates a third legitimate code. Researchers (Crystal, 1997) have also presented several socio-cultural reasons and motivations for switching. There have been studies to depict the usage of particular language based on the emotional attachment and the sentiment of the person towards that topic (Rudra et al., 2016). In this paper, we adopt the perspective of descriptive linguistics and make an effort to describe this prevalent form of language as it occurs, without adopting a prescriptive approach.

Ubiquitous access to social media tools and platforms have also made CM the preferred choice for both formal and informal communication. In such settings, where the communication is either semi-formal or informal, researchers ((Bali et al., 2014), (Barman et al., 2014)) have observed a higher tendency for multilingual speakers to use CM. We studied a sample of conversation logs from a commercial chit-chat based conversational agent in the Indian market. The agent was trained to engage in informal chat conversations with the help of a database of Twitter conversations from the Indian market. Since India is a multilingual country with a large number of multilingual speakers, we notice that users often freely use each language individually or their CM versions while conversing with the agent. We notice that, in around 3% of overall conversations, users were found to be chatting with the agent in CM language such as *'hello, kya chal raha hai'* (Meaning: hello, what's up?). Interestingly, in cases where the response of the agent was in CM language such as *'sorry yaar'* (Meaning: sorry friend), users too

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 29–38
Melbourne, Australia, July 19, 2018.

responded back in CM language in 27% of those times. There have been other studies regarding the quantitative and qualitative aspects of code-switching on social media along similar lines (Hidayat, 2008). However, a large number of NLP applications, such as Question Answering (QA), Dialogue Systems, Summarization etc, still continue to be designed with the assumption of a single interaction language such as English (Brill et al., 2002), Hindi ((Kumar et al., 2005)), Chinese ((Yongkui et al., 2003), (Sun et al., 2008)). Such systems are most likely to break given a CM utterance which has multiple languages mixed at sentence, phrase or morphological level. Hence, it is highly imperative for researchers to focus on building more robust end-user NLP applications which can understand and process CM language.

Building a good evaluation dataset for Factoid QA in CM is wrought with challenges such as a) ensuring that the annotators are unbiased in anyway to artificially use CM b) recruiting a good team of native bi-lingual speakers as annotators c) maintaining a good quality and diversity of questions across intents, answer types and entities. In this paper, we describe our experience in dealing with the above challenges while creating the dataset. We used a crowd-sourcing platform for collecting data where the crowd workers were restricted to only native language (Hindi, Telugu and Tamil) speakers. We shared a detailed set of guidelines and instructions about the task with the crowd workers and also ran them through some basic quality checks before collection of actual data. Finally, we were able to collect around 1,694 Hinglish, 2,848 Tamlish and 1,391 Tenglish factoid questions along with their answers. We have organized a Code-Mixed Question Answering challenge based on this data for the first edition of this challenge. There are 7 teams that registered and took the data from us. In this paper, we discuss the preliminary techniques that 2 of these groups used. To summarize, the following are the main contributions of this paper:

- We curated an evaluation dataset for the task of Factoid QA in CM languages with more than 5000 QA pairs for Hinglish, Tamlish and Tenglish languages. We also make it freely available for research purposes.
- We share our experiences related to eliciting lexically unbiased CM questions by using images as anchor points.
- We present the techniques used in the first edition of the CM QA challenge.

2 Related Work

Early work in this domain include investigating CM phenomenon in a formal and computational framework (Joshi, 1982) and developing formalisms (Goyal et al., 2003), (Sinha and Thakur, 2005). Recent years have seen attention towards part of speech tagging for CM languages and gathering corpora ((Vyas et al., 2014), (Solorio and Liu, 2008), (Jamatia et al., 2015), (Soto and Hirschberg, 2017)) for it. Language identification in mixed language scenarios has also been studied recently ((Barman et al., 2014), (Chittaranjan et al., 2014)) and has also been aggressively addressed as a shared task at major conferences ((Solorio et al., 2014), (Sequiera et al., 2015)). Some of the other applications that were picked up in research in CM over the past few years include Named Entity Recognition (Zirikly and Diab, 2015), semantic parsing (Duong et al., 2017), dependency parsing (Partanen et al., 2018) and shallow parsing (Sharma et al., 2016). While the above work focusses on important language processing challenges in CM, we are more interested in end-user NLP applications which support CM such as Factoid QA in CM languages.

Eliciting a corpus of CM questions by paraphrasing an English question was used to perform question classification (Raghavi et al., 2015). While this method has the advantage of having a ground truth parallel text, the possibility of lexical bias from the English question while framing the code-mixed question exists. An extension to this work was proposed by building an end-to-end web based CM question answering system named WebShodh (Chandu et al., 2017). Efforts have been made to develop cross lingual QA systems that take questions in English and answer back in English but search for candidate answers in Hindi newspapers (Sekine and Grishman, 2003) along with other machine learning approaches (Nanda et al., 2016). There has been some work in the early 2000s to generate a dialog based QA system in Telugu to support Railway inquiries (Reddy and Bandyopadhyay, 2006). This kind of cross language QA system is being researched for European languages as well (Neumann and Sacaleanu, 2003). A dataset of 506 questions from messages from Facebook was proposed in the Bengali-English CM domain (Banerjee et al., 2016). Our dataset is over ten times larger than this data and takes into account the lexical variation brought in by collecting questions from images and code-mixed articles.

3 Dataset Collection

In order to study differences between lexical bias from entrainment and elicit lexically diverse questions, we employ two modes of data collection: eliciting code-mixed questions from a) images and b) code-mixed articles. The former are general questions and the latter are context specific questions (similar to machine reading). Techniques of collecting queries for a dialog system by presenting scenarios symbolically and diagrammatically was previously used (Black et al., 2011) in order to minimize supplying lexical and phrasal cues. For collection of Hinglish data, we used both these approaches whereas for collecting Tenglish and Tamlish data, we used only images. This is because for Hinglish, we could find informative blogging websites based on which it is easier to frame factoid code-mixed questions. However, to the best of our knowledge, during the time of our annotation, such fact based code-mixed content was still not available in Tenglish/Tamlish. It is also noted that it is less likely to get questions that have abstract answers (beyond the realm of physical entities) when they are collected based on images.

3.1 Challenges in Code-Mixed Factoid Questions Collection

We faced the following challenges during the data collection task.

1. How would we eliminate the bias towards using English in general scenarios while using a search engine etc.,? In other words, we need to encourage crowd workers to provide us with data that is neither biased to English monolingual questions due to preconceptions of the language preference while interacting with a computer, nor bias them to provide mixed language data if it does not feel natural to them.
2. How do we eliminate responses from people who are not native speakers? To mitigate this problem, we have given the instructions to each of these target languages in romanized code-mixed version of the corresponding languages mixed with English. This has the dual advantage of being understood only by those who have enough competence in the matrix language as well as easing them into code-mixing and making them comfortable with it.
3. How do we elicit factoid questions? This is a trivial issue. We had to explain what a factoid question is while providing sufficient examples of factoid and non-factoid questions.
4. How do we collect questions that are general enough that they could be answered without providing the context of the images (for image based questions)? The design of the task to collect questions based on images, in order to study the comparative lexical bias when a code-mixed article is given, has resulted in a lot of questions that are related to multi-modal reasoning. For example, Tenglish question *'image lo entha mandi unnaaru?'* (Meaning: How many people are there in the image?) requires a visual in order to answer. We removed such questions in the post processing after data collection.

We ensured a good mix of categories while selecting the target fulcrum entity images (example: guitar, bicycle), location (example: Eiffel Tower, Golden Gate Bridge), person (example: Roger Federer, Eminem), event (example: World War 2, Dandi March). Out of these, we manually selected 80 images from which factoid questions can be asked. To gather questions based on articles, we first scraped documents from *hinglishpedia.com*, randomly selected 80 articles from them and made sure that all of them were code-mixed. The crowd-workers are then requested to form factoid questions based on these articles such that the answers to the questions are present in the corresponding article.

3.2 Crowd-sourcing for Question & Answer Collection

We engaged with two streams of demographics while collecting the data: university students and crowd-workers. Each participant is allowed to provide us with only 20 questions to avoid idiolectic biases i.e, biases of each individual. In the first step, we performed the activity in a more controlled environment in university classrooms. The instructors of the classrooms were requested to give a brief presentation we made about what code-mixing is along with some example questions. This was performed to alleviate the bias against mixing while interacting with a machine. The students (with native languages among Hindi, Telugu and Tamil) were given clear explanations about factoid and non-factoid questions in order to elicit the right kind of questions for our task.

In the second phase, we migrated this setup to Amazon Mechanical Turk task, where crowd-workers were redirected to our interface[1] of mixed language instructions based on their native language. Each accepted Human Intelligence Test

[1] https://docs.google.com/document/d/1CTFTjmU6RKUwsNH1z0Sjl_8EZt8dzl-VGF5VwPLIpy0/edit?usp=sharing(toanonymize)

Category of Questions	Num	Multilingual Index	Language Entropy	Integration Index	Avg Length
Hinglish image questions	1,419	0.72	0.61	0.25	7.50
Hinglish article questions	275	0.88	0.66	0.29	8.90
Tamlish questions	2,848	0.69	0.59	0.24	5.56
Tenglish questions	1,391	0.80	0.64	0.28	5.90

Table 1: *Data Statistics: The code-mixing metrics for Hinglish (Hindi+English), Tamlish (Tamil+English) and Tenglish (Telugu+English) questions*

(HIT) was compensated with $2.50 on the Turk setup. We have got a lot more responses for Tamlish questions as compared to both Hinglish and Tenglish together as reflected in Table 1. In this scenario, although the turkers have not been formally explained about what code-mixing is apart from providing them with instructions and examples, most of the data we received included mixed language Romanized questions. While in the former scenario, the collected questions were explicitly moderated to remain within bilingual and multilingual environments, like in India, the latter scenario does not ensure this, as they do not have to be present in India. The extent of code-mixing and fluency of the questions may vary as compared to the questions collected from Indian classroom environment due to the difference in their socio-cultural environments. Though we have mentioned in the instructions to provide us with questions that sound natural to the participants, we acknowledge that it may not have been completely natural as they were explicitly requested to mix two languages. The third phase involves collecting answers to all the questions. Monolingual questions and image-based questions that contain referring expressions, such as *'in this figure'* were removed before collecting answers. To filter out noisy and random answers, the set-up includes a qualifying CM question for which we clearly knew the answer. When collecting answers, we only accept them from workers who correctly answer the qualifying question.

3.3 Curation and Post-processing

After data collection, we removed duplicate entries and also performed one step of human verification. This responsibility was divided into two phases. The first step was employing certain post processing steps in order to remove the questions that did not match the presented specifications and rejecting them. One major problem is the use of referring expressions and determiners corresponding to the images about which the questions were asked. In each of the three languages, we made a list of all possible spelling variants of referring expressions like *'image/picture mein'* (Meaning: in the image), *'ye'* (Meaning: this), *'iss'* (Meaning: this) and separated the questions that contain these expressions. The same process was not done for questions collected based on code-mixed articles. This is because referring expressions corresponding to the given text do not hinder searching for an answer in the given snippet. Lexical level language identification is performed to remove the questions that do not have atleast one word from both the languages. These selected questions after filtering are then curated and gone through manually to add back the questions that made sense before rejecting the HITs. The next level of curation was performed during the answer collection phase. This was necessary because it was still possible to bypass these curation conditions. For example, there were some entries that seemed like English queries with an additional suffix belonging to the corresponding native language at the end of some of the words. For example, *'Whatil isil waterfalla borderil America and Canada?'* (Tamlish question for 'What is the waterfall in the border of America and Canada'). On the other hand there are queries that seemed to have been translated using an online translation tool into the matrix language and randomly inserting some English words in between. For example, *'Mein which Indian state did Mother Teresa kaam kiya?'* (Hinglish question). This example seems to be a lexical level translation of first, eighth and ninth words of the English question 'In which Indian state did Mother Teresa work (past-tense)?' into Hindi. 67.87% of the data collected from Turk was acceptable and passed our curation tests. Among the remaining, about 21% were rejected due to the use of referring expressions, 11% due to erroneous attempts by typing junk words. All the questions passed the curation tests for more than 90% of the accepted HITs and some of the questions were acceptable for the remaining 10%. This implies that the instructions provided for the task were sufficiently clear to elicit CM factoid questions. The above are the statistics for the data corresponding to the crowd sourced platform that might provide a baseline estimate for collecting useful data for this domain on such platforms.

Since the number of curators is much less (approximately 6 people) than the number of crowd-workers, we need to understand that the curation process is much more expensive in terms of manual effort. The above steps are taken to elicit quality data for our purposes. The tasks of collecting questions and answers were deliberately separated for two reasons. One is to ensure clarity of the task and make sure that the users are giving naturally code-mixed questions and asking them to provide answers in English within the same task might lead to unnecessary biases or confusion. The second reason is that when asked to provide questions and answers together, one might tend to ask the simplest questions to which answers are already known to them. This in turn might have reduced the variety of questions for the same anchor point. We have also collected feedback about each question whether it is a factoid and if additional multi-modal information is needed to answer it, during the answer collection phase.

4 Data Analysis

Recent studies have focused on empirical measurements of code-switching (Guzmán et al., 2017). The multilingual index(M-Index), Language Entropy and Integration index(I-index) measure the extent of mixing and switching frequency. Table 1 presents the statistics of out dataset along with these metrics for mixing. The average number of words per question is higher in Hinglish compared to Tamlish and Tenglish. The M-index for all the 3 language pairs are in comparable ranges while it is slightly less for Tamlish. The questions are provided along with the following information: (1) language information (2) question type annotated as 'context dependent' (for article based questions) and 'context independent' (for images based questions) and (3) corresponding article for article based questions.

4.1 Answer Type Distribution

In order to analyze the distribution of question types in our dataset, we sampled questions, from Tenglish and Hinglish, which contain either of the following two selected images - 'Taj Mahal' and 'Hiroshima'. We use the coarse and fine-grained type hierarchies proposed by (Li and Roth, 2006) for annotating the questions. For the first image: 'Taj Mahal', we had 91 Tenglish questions and 71 Hinglish questions. For Tenglish questions, the distribution of coarse level types were 34 PERSON, 8 ENTITY, 30 LOCATION and 19 NUMERIC. For Hinglish questions, the coarse-type distribution was found to be 25 PERSON, 9 ENTITY, 22 LOCATION and 15 NUMERIC. An interesting observation we noticed was that - there were 23 Tenglish and 17 Hinglish variants of the question *'Who built Taj Mahal?'* and similarly there were 12 Tenglish and 8 Hinglish variants of the question *'In which city is Taj Mahal located?'*. A similar analysis for the other focus entity 'Hiroshima' gave us 21 Tenglish questions distributed as 7 NUMERIC, 9 LOCATION, 4 ENTITY, 1 PERSON type questions and 14 Hinglish questions distributed as 10 NUMERIC, 2 LOCATION, 1 ENTITY and 1 PERSON type questions. Among these we observed 4 and 8 variants in Tenglish and Hinglish respectively for the question *'In which year did attack on Hiroshima and Nagasaki take place?'*. These statistics reveal that, for the same image shown to them, participants issued questions resulting in a variety of answer types in the target code-mixed languages. Figure 1 shows the percentage distribution of the question types for 'Taj Mahal' and 'Hiroshima' across the three language pairs.

In order to gain better intuitions on the *'why and how'* of code-mixing, the collected questions are studied with respect to idiolectic language preferences, i.e the idiosyncrasies of mixing the languages and the extent of code-mixing across languages. To study the individual mixing biases in the data, *Multilingual Index* is calculated for each individual in each of the language pairs. Figure 2 shows histograms for idiolects of Hinglish, Tenglish and Tamlish. As can be observed from the figure, Hinglish and Tenglish has many crowd-workers towards the higher end of multilingual index whereas Tamlish has a rather smoother distribution except for the last range.

4.2 Lexical Bias in Article based Questions

The bias of copying the words was mitigated to an extent by the usage of images as anchor points to collect questions. However, studying the lexical bias when a code-mixed article is given acts as a proxy to study entrainment. The variant of expressing the questions from code-mixed articles serves two purposes. One is to study the differences in difficulty of down stream task of retrieving for question answering as compared to the image based questions. In this category of questions, the answer is present in the snippets that are given and the focus is primarily on retrieving the answers from within the given text. The second is to study the varying lexical biases to frame a question when code-mixed content is given versus when it is not. To study this empiri-

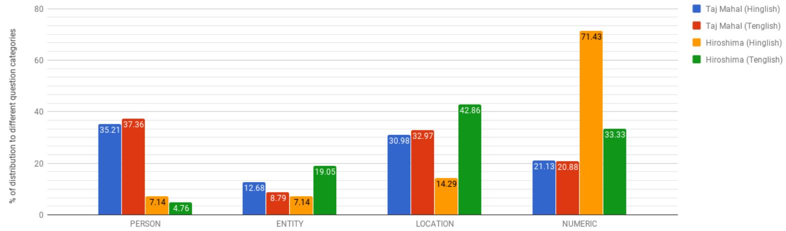
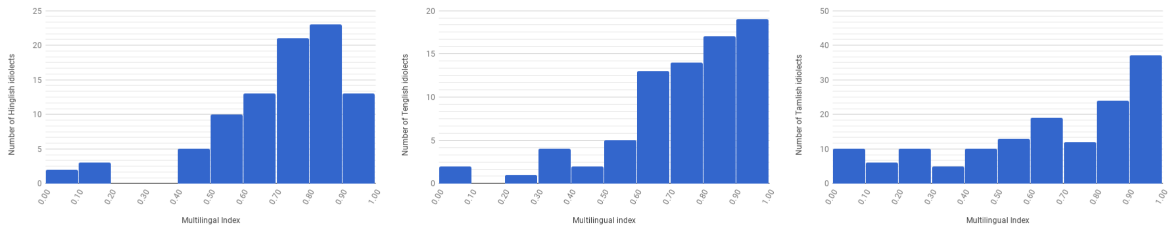


Figure 1: Distribution of question types in Hinglish and Tenglish for 2 topics: Taj Mahal and Hiroshima



(a) Hinglish idiolectic mixing distribution (b) Tenglish idiolectic mixing distribution (c) Tamlish idiolectic mixing distribution

Figure 2: Histograms for Multilingual index for idiolects for Hinglish, Tenglish and Tamlish

cally, we calculated the percentage of intersection of words between question and articles. The average of overlapping words is 54.20%, while the minimum and maximum are 12.5% and 92.31% respectively. Similarly, the longest overlapping subsequences have a mean of 2.24 with a minimum of 1 word and a maximum of 16 words.

4.3 Mixing Phenomena observed in the Data

One of the interesting categories of mixing that is observed in the data is mixing gender information of the native and the mixed form of the word. For example, consider the question from the data, *'earth kab form hui thi?'* (English meaning: When was Earth formed?). A paraphrase of the same question is *'dharthi kab form hui thi?'*. In Hindi, the gender of the verb has to agree with the subject. While the gender of Earth is masculine (which should have agreed with *'hua'* and not *'hui'*), the gender of *dharthi* (which agrees with *'hui'*) is feminine as perceived by a native speaker. But as observed in the question, feminine form *'hui'* is used with 'Earth' which is mixed word from English. Sebba (2009) refers to this as one type of 'harmonization strategy' in language mixing and it is one that he says might be typical of highly literate bilinguals. We believe the naturalness of the data is highly dependent on the nativity of annotators. Throughout our process, we took as much care to ensure that we use native speakers of the language for our annotation. However, there were still a few exceptions. We also tried avoiding completely random, spurious and noisy inputs by checking if they were simply permutations of the original input and their lexically translated words.

A known problem in dealing with code-mixed text is non-standardized Romanization of native language when mixed with English. Phonological perceptions of a syllable can be represented differently. For example, from the data a couple of the very frequent such variations are *'kon'* and *'kaun'* for 'who' in Hindi, and *'he'* and *'hai'* for 'is' in Hindi. For both these words, the latter variants are closer to the pronunciation of the Hindi words, but the other sounds are in colloquial usage frequently as well. Consider the question, *'Friends serial ke kitne seasons banaye ja chukein hain?'* (Meaning: How many seasons were made for Friends serial?). Using *'n'* in *'chukein'* indicates that the person literally transliterated the Hindi spelling into Roman spelling because colloquially the *'n'* sound is often omitted while speaking. A similar observation applies to the word *'kartein'* (Meaning: do). Similarly, 'pe' is a more colloquial usage of the word 'par' (Meaning: on). Though 'pe' is never used in standard written Hindi, the data collected has both variants of the words. Similar observations in Tenglish data include variations for *'cheyinchaadu'* and *'ceyincadu'* (both the words mean 'did'). This problem compounds in Tenglish since Telugu is an agglutinative language. For example, in the variants *'chesthu unaadu'* and *'chesthunnadu'* (meaning: have been doing (masculine form)), the two words can be written together as a single word or separately.

Some of the examples show forcible mixing since the instructions specifically mentioned to provide code-mixed questions. For example, *'Android ko Google ne kab buy kiya tha?'* (Meaning: 'When did Google buy Android?'). The word *'khareed'* which means 'buy' is a very common Hindi word and in such cases, the native word is used more naturally as opposed to the mixed word. 10 such examples were selected from Hinglish and shown to 5 native speakers of Hindi to annotate if they seem natural, unnatural or neutral. All these examples were marked as either unnatural (36%) or neutral (64%) and none of them were marked as natural. This shows that there is some pattern or notion of mixing words for native speakers.

In some other examples, we also observed what can be considered an opposite of forced mixing. For example, in question *'1994 mein premier kiya hua pramukh American comedy TV saathiyon ka naam kya hai?'* (Meaning: What is the name of the famous American comedy TV show Friends that was premiered in 1994?), words like *'pramukh'* (Meaning: famous) and *'saathiyon'* (Meaning: friends) are less common in common usage compared to their English mixing counterparts. Also, note that this is uncommon since the named entity 'Friends' is translated to the Hindi counterpart. Another known phenomenon is the mixing of languages at morphological level which was observed very commonly in the data. This poses a problem for word level modeling or formulation for addressing the down stream tasks such as our current case of question answering. For example, in the Tenglish question *'Eiffel Tower ni entha mandi architectlu design chesaru?'* (Meaning: How many architects designed Eiffel Tower?), the word *'architectlu'* (Meaning: architects) is mixed at morphological level by English word 'architect' and Telugu suffix *'lu'*, which is a plural marker.

5 CM QA Challenge: Techniques

The CM QA challenge was announced and broadcasted during the summer of 2017. The task is to provide a ranked list of relevant answers for given CM queries. The image based questions are annotated as 'general' and the article based questions are provided with the corresponding articles. While 7 teams have registered to take part in the challenge and have collected data from the organizers, a couple of teams have successfully completed participating in the challenge. In this section, we discuss the techniques used by two participating groups to address the first edition of this challenge.

As discussed in Section 3, there are 2 categories of questions; (1) general questions where there is no context, and (2) article based questions for which the answers are retrieved from a given context. To address the latter type, paragraphs from Wikipedia are leveraged as general context. One team (from Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)) addressed this by identifying the named entities in the CM query and look them up in the summaries of Wikipedia articles [2]. These summaries typically contain 5 sentences. The second team (from IIIT Hyderabad) trained a similarity model using DSSM (Huang et al., 2013) to retrieve and rank the answer bearing sentences from Wikipedia. Both the groups have worked along similar lines to address questions with general context.

The team from DFKI dealt with article based questions as well. At this stage, both the categories of questions contain query and information about relevant paragraph. A pre-trained Document Reader model DrQA proposed by Chen et al. (2017) on a popular machine reading QA dataset SQuAD (Rajpurkar et al., 2016) is used for this domain. This model answers open domain factoid questions using Wikipedia by not considering document retrieval. An open source implementation of this model[3] is used and our results are lower than we expected: average EM is 0.0691 and average F1 is 0.1001 on the training dataset. In the category of general questions (image based) where the relevant paragraph is not given, the predicted answer is similar in meaning to the ground truth but can be broader. For instance, when the Hinglish answering 'eminem ka profession kya hai?' (Meaning: What is Eminem's profession?), this system gives 'rapper, record producer, and actor', as compared to 'Rapper'. Though the system is correct, the answer gathered included only 'rapper' which most data collection techniques for QA face an issue. To train these models, Hindi embedding space is mapped into the English one. A standard approach in relation to Hindi was investigated by (Bhattacharya et al., 2016) involving finding a translation matrix (using linear regression) that minimizes the reconstruction error between target language embeddings and translated embeddings.

[2] https://pypi.org/project/wikipedia

[3] https://github.com/facebookresearch/DrQA

This idea is developed by using a neural network and a random forest regression to find translation matrix. By using Polyglot Hindi and English embeddings with Universal Word-Hindi Dictionary we achieve MSE score of 0.057.

6 Challenges observed in CM QA

Gathering more data: The training subset contains 1295 unique question-answer pairs, which poses a significant challenge to train complex models from scratch. As an alternative, a transfer learning technique can be used, using a model pre-trained on a large-scale open-domain factoid dataset, such as SQuAD. For instance, community question answering forums can be used, which naturally contains a lot of code-mixed language due to the extensive borrowing of technical terms. Such setup has two benefits: it eases the problem of collecting new data and alleviates the need to manually label it.

Spell Checking: Since the question-answer pairs are coming from an informal background, some of them are misspelt. Language identification is an overhead to deal with this using traditional spell checking techniques. An extensive use of dictionaries is the most obvious approach, but a more practical solution might be to use character-based methods and introduce artificial noise to make models more robust.

Romanization variability: It should also be noted that apart from spell checking, there is variability in romanization output. For example, the Hindi word 'jidhar'(Meaning: where) can either be written as 'jidhr'. As it is unclear which of the models of transliteration would a user prefer, a developer needs to keep all options open.

Poor translation from open source tools: In many cases, translation tools completely distort the meaning of the sentence. An illustrative example of this is: 'Sun ka colour kya hai?' (What is the colour of the Sun?) - 'What is the color of listening?' and so on. As one can see, an English collocation 'full name' is not preserved, but translated into 'Fullham'. In some cases it can be explained by the incorrect use of capitalization: 'niagara falls kaunse desh mein hai?' is translated into 'What is the name of the person who is suffering from diabetes?', but using capitalized 'N' gives correct translation. It is worth noting that incorrect query translation contributed to approximately 35% of errors.

Answer granularity: Moreover, while performing error analysis, we have found a few cases where a level of required granularity for an answer was unclear. A common type of error for the model was to output 'Champ de Mars in Paris, France' when asked 'Eiffel Tower kahan hai?' (Where is the Eiffel Tower?), while the ground truth answer was 'france'. Errors like that account for approximately 7% of all the wrong predictions in the development set. Such cases suggest that considerable attention must be paid during labeling of a corpus. One can either keep a list of acceptable answers or provide refined guidelines for both annotators and developers. In the latter, it might help to analyze human performance on the same dataset to understand what is the most common answer granularity level.

Cross-lingual embeddings: Finally, when working with neural models, we have to carefully approach the construction of embedding spaces. While in the current version we have worked only with English translations, a neater approach would be to directly use both languages. (Ruder, 2017) provides an extensive survey of the available approaches. Whereas more and more resources are emerging for Hindi, such as MUSE (Conneau et al., 2017), few researchers have addressed the task for Telugu and Tamil.

7 Conclusions

In this paper, as a first step towards fostering research in the area of Factoid QA in CM languages, we present our evaluation dataset consisting of more than 5000 crowd-sourced questions along with their answers in three CM languages - Hinglish, Tenglish and Tamlish. We received a lot more Tamlish questions on crowd sourced platform compared to the other two languages. We also shared our experiences while curating this evaluation dataset such as usage of images as anchor points to avoid lexical biasing towards CM. We have looked at the extent of lexical biasing of the words in article based questions. In future, we would also like to see if the participants are inverting the language for the words present in the articles. The dataset features a diverse range of answer types across all the CM languages. We also shared some interesting properties of this dataset related to lexical bias and other phenomenon related to code mixing. In future, we would like to explore techniques to generate synthetic CM data from large-scale datasets.We plan on continuing the data collection process to elicit more data. This paper also reports the first edition of the challenge and plan on continuing it in the coming years as well. We have made our dataset freely available for research purposes to encourage more research work and result in significant advances in the area of Factoid QA in CM languages.

References

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 First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. 2016. The first cross-script code-mixed question answering corpus. In *MultiLingMine@ ECIR*, pages 56–65.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2016. Using word embeddings for query translation for hindi to english cross language information retrieval. *Computación y Sistemas*, 20(3):435–447.

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, pages 2–7. ACL.

Eric Brill, Susan Dumais, and Michele Banko. 2002. An analysis of the askmsr question-answering system. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 257–264. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Manoj Chinnakotla, Alan W Black, and Manish Shrivastava. 2017. Webshodh: A code mixed factoid question answering system for web. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 104–111. Springer.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

David Crystal. 1997. *The Cambridge encyclopedia of language*, volume 1. Cambridge University Press Cambridge.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389.

P Goyal, Manav R Mital, A Mukerjee, Achla M Raina, D Sharma, P Shukla, and K Vikram. 2003. A bilingual parser for hindi, english and code-switching structures. In *10th Conference of The European Chapter*, page 15.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. *Proc. Interspeech 2017*, pages 67–71.

Taofik Hidayat. 2008. An analysis of code switching used by facebookers.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

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 the International Conference Recent Advances in Natural Language Processing*, pages 239–248.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

Praveen Kumar, Shrikant Kashyap, Ankush Mittal, and Sumit Gupta. 2005. A hindi question answering system for e-learning documents. In *Intelligent Sensing and Information Processing, 2005. ICISIP 2005. Third International Conference on*, pages 80–85. IEEE.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):229–249.

Lesley Milroy and Pieter Muysken. 1995. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Garima Nanda, Mohit Dua, and Krishma Singla. 2016. A hindi question answering system using machine learning approach. In *ICCTICT 2016*, pages 311–314. IEEE.

Günter Neumann and Bogdan Sacaleanu. 2003. A cross–language question/answering–system for german and english. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 559–571. Springer.

Niko Partanen, KyungTae Lim, Michael Rießler, and Thierry Poibeau. 2018. Dependency parsing of code-switching data with cross-lingual feature representations. In *Proceedings of the Fourth International Workshop on Computatinal Linguistics of Uralic Languages*, pages 1–17.

Shana Poplack. 1980. Sometimes ill start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. Answer ka type kya he?: Learning to classify questions in code-mixed language. In *Proceedings of WWW 2015*, pages 853–858. ACM.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rami Reddy Nandi Reddy and Sivaji Bandyopadhyay. 2006. Dialogue based question answering system in telugu. In *Proceedings of the Workshop on Multilingual Question Answering*, pages 53–60. Association for Computational Linguistics.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of EMNLP 2016*, pages 1131–1141.

Mark Sebba. 2009. *On the notions of congruence and convergence in code-switching*. Cambridge University Press.

Satoshi Sekine and Ralph Grishman. 2003. Hindi-english cross-lingual question-answering system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):181–192.

Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, et al. 2015. Overview of fire-2015 shared task on mixed script information retrieval. In *FIRE Workshops*, volume 1587, pages 19–25.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Srivastava, Radhika Mamidi, and Dipti M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. *arXiv preprint arXiv:1604.03136*.

R Mahesh K Sinha and Anil Thakur. 2005. Machine translation of bi-lingual hindi-english (hinglish) text. *10th Machine Translation summit (MT Summit X), Phuket, Thailand*, pages 149–156.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing universal part-of-speech tags for code-switching. *arXiv preprint arXiv:1703.08537*.

Ang Sun, Minghu Jiang, Yifan He, Lin Chen, and Baozong Yuan. 2008. Chinese question answering based on syntax analysis and answer classification. *Acta Electronica Sinica*, 36(5):833–839.

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

ZHANG Yongkui, ZHAO Zheqian, BAI Lijun, and CHEN Xinqing. 2003. Internet-based chinese question-answering system. *Computer Engineering*, 15:34.

W Quin Yow and Ferninda Patrycia. 2011. Challenging the linguistic incompetency hypothesis: Language competency predicts code-switching.

Ayah Zirikly and Mona Diab. 2015. Named entity recognition for arabic social media. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 176–185.