“Answer ka type kya he?” Learning to Classify Questions in Code-Mixed Language

Khyathi C. Raghavi
IIIT Hyderabad, India
chandukhyathi.raghavi@research.iiit.ac.in

Manoj Chinnakotla
Microsoft, India
manojc@microsoft.com

Manish Shrivastava
IIIT Hyderabad, India
m.shrivastava@iiit.ac.in

ABSTRACT

Code-Mixing (CM) is defined as the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language. CM is a natural phenomenon observed in many multilingual societies. It helps in speeding-up communication and allows wider variety of expression due to which it has become a popular mode of communication in social media forums like Facebook and Twitter. However, current Question Answering (QA) research and systems only support expressing a question in a single language which is an unrealistic and hard proposition especially for certain domains like health and technology. In this paper, we take the first step towards the development of a full-fledged QA system in CM language which is building a Question Classification (QC) system. The QC system analyzes the user question and infers the expected Answer Type (AType). The AType helps in locating and verifying the answer as it imposes certain type-specific constraints.

We learn a basic Support Vector Machine (SVM) based QC system for English-Hindi CM questions. Due to the inherent complexities involved in processing CM language and also the unavailability of language processing resources such POS taggers, Chunkers, Parsers, we design our current system using only word-level resources such as language identification, transliteration and lexical translation. To reduce data sparsity and leverage resources available in a resource-rich language, in stead of extracting features directly from the original CM words, we translate them commonly into English and then perform featurization. We created an evaluation dataset for this task and our system achieves an accuracy of 63% and 45% in coarse-grained and fine-grained categories of the question taxonomy. The idea of translating features into English indeed helps in improving accuracy over the uni-gram baseline.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Question-answering (fact retrieval) systems; I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms

Algorithms, Experimentation

Keywords

Code-Mixing, Question Answering, Machine Learning, Support Vector Machines, Classification

1. INTRODUCTION

India is a multilingual society with 30 languages which are spoken by more than a million native speakers. Although, Hindi and English have been designated as the official languages by the Union Government, many speakers are bilingual or trilingual and have knowledge of Hindi/English in addition to their mother tongue[1]. *Code-Mixing (CM)* is defined as the *“the embedding of linguistic units such as phrases, words, and morphemes of one language into an utterance of another language”* [20, 19]. It is a natural phenomenon which is observed in multilingual speakers. CM allows speakers to speed up communication especially when they are short of words or do not know the appropriate word in a native language [7]. Due to this, CM has also become a popular mode of expression in social media like Facebook, Twitter etc.

Modern search engines and Information Retrieval (IR) systems have evolved from supporting telegraphic user queries which return the top relevant URLs to also support question queries, expressed in Natural Language (NL), which retrieve the precise answer. Besides, with the steady increase in mobile, smart-phone users and voice based search, Question Answering (QA) is turning out to become the most convenient and natural way to get quick and accurate answers for certain kinds of user information needs. For example, one can ask a search engine or virtual assistant on phone (Such as Siri, Cortana, Google Now) - “Who is the president of India?” and get the precise answer “Pranab Mukherjee” in stead of browsing through the ten blue links to actually figure it out.

Current QA research [15, 16, 11] and systems only support interaction in a single language such as English, French and German etc. This assumption 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 Hindi speaker wants to know his driving license number, he is more likely to express it as - *“mera driving license*

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743006.

[1] http://goo.gl/yfumm8

number kya he?" (Translation: what is my driving license number?) where *driving license* are English words mixed in Hindi. Hence, to increase the reach, impact and effectiveness of QA systems in multi-lingual societies, it is highly imperative to support understanding of CM language. In this paper, we take the first step towards the development of a full-fledged QA system for CM language which is building a Question Classification (QC) system. QC is a crucial component of QA system which analyzes the user question and infers the expected Answer Type (AType). The AType imposes certain type-specific constraints that help in locating and verifying the precise answer. For example, for the question, *Which Indian City is also known as "Pink City"?*, the QC system will classify it as LOCATION:CITY implying that only candidate answers which are cities need to be considered for further processing and analysis.

Computational processing and understanding of CM data has been known to be linguistically challenging [3, 13, 24]. Although, building language analysis tools for CM is currently an active area of research [24, 22], it is still *resource-scarce* and there are no resources such as Chunkers, Shallow Parsers etc. available, which many QC systems [15, 8, 26, 25, 18] heavily rely on. Moreover, roman script is used to express content in languages which have non-roman scripts such as Hindi, Bangla, Chinese, Arabic etc. [23]. This brings in the additional challenge of accurately identifying the language at a word-level.

In this paper, we learn a Support Vector Machine (SVM) based QC system for English-Hindi CM questions. Since, no language resources such as Chunkers, Parsers exist for CM, we just make use of word-level resources such as language identification, transliteration and lexical translation to mine features in a single common language and then train a SVM model to predict the Question Classes. The motivation to translate word features is mainly to reduce data sparsity and also to harness any monolingual resources available in the common language. In this context, we believe translating into a resource-rich language such as English would be helpful since it has QC training data. We created a CM question dataset for English-Hindi language pair from college students and evaluate our approach on it. We achieve a coarse-grained average accuracy of 63% and fine-grained accuracy of 45% on English-Hindi question corpus using the Li and Roth [15] question classification hierarchy. The idea of translating features into a common languages helps in improving accuracy over the uni-gram baseline.

The paper is organized as follows: Section 2 discusses the related work in this area. In Section 3, we describe the methodology used to create the dataset along with its characteristics. Section 4 describes the architecture of our system and feature generation techniques. Section 5 describes the experimental results and error analysis. Finally, Section 6 concludes the paper.

2. RELATED WORK

CM is a well studied topic in linguistics literature [17, 1, 2]. However, they have mainly studied the sociological, conversational motivation behind CM and also its linguistic nature. Li *et. al.* [14] and San *et. al.* [21] have studied code-mixing for Chinese-English languages in Hong Kong and Macao and reported that there are linguistic motivations causing such behavior. Dey *et. al.* [7] present an analysis of English-Hindi code mixing corpus developed out of student interviews and also discuss the grammatical contexts in which such behaviors are triggered. Linguists also distinguish the phenomenon at a intra-sentence (Code-Mixing) and inter-sentence (Code Switching) level. Hidayat [10] reports that 45% of the switching was due to real lexical needs, 40% was used for talking about a particular topic, and 5% for content clarification.

Language	Hindi-English CM
Number of Questions	1000
Number of Words	13276
Percentage of English Words	0.3458
Percentage of Hindi Words	0.6529
Avg. CM Words (Eng) per Question	5
Avg. Length of Questions	11

Table 1: Dataset Details

In recent years, there is a surge of interest in tackling code-mixed language input for Indian Languages [3, 24, 6, 12]. This increased interest may be attributed to the large scale use of code-mixed language by urban youth of India in various social forums. Barman *et. al.* [3] presented the challenges of Language Identification in code mixed text. They also forward the claim that code-mixing is frequent among speakers who are multilingual. Vyas *et. al.* [24] describe their "initial efforts to POS tag social media content from English-Hindi bilinguals while trying to address the challenges of code mixing, transliteration and non-standard spelling, as well as lack of annotated data". The authors conclude that while CM is a common phenomenon in all multilingual societies, transliteration still remains an issue. There are some recent studies on the impact of code-mixing on the effectiveness of IR. In this context, Gupta *et. al.* [9] have worked on query expansion for mixed script and code mixed queries.

A lot of work has been done on QC for QA in a single language [15, 8, 26, 25, 18]. To the best of our knowledge, our work is the first one which attempts to build a QC system for CM languages based on the Li and Roth question hierarchy [15]. We envision this as the first step towards building a full-fledged QA system for CM language.

3. DATASET CREATION

For this research, we collected code-mixed question data from 30 student volunteers at IIIT Hyderabad who were native speakers of Hindi. We initially downloaded approximately 1200 questions from all the episodes of "Kaun Banega Crorepati (KBC)" - a popular Indian television game show based on the UK game show "Who Wants to Be a Millionaire?". To this, we added some school level questions from the Central Board of Secondary Education (CBSE) - a Board of Education for public and private schools under the Government of India. The above questions were then filtered to retrain only factoid questions having a unique answer. This resulted in a total of 1080 factoid questions out of which we chose 1000 questions randomly. For each of these 1000 questions, we asked the student annotators on how they would pose this question in a code-mixed language. In order to avoid any individual bias in the way the code-mixing happens, each question was assigned to three annotators who

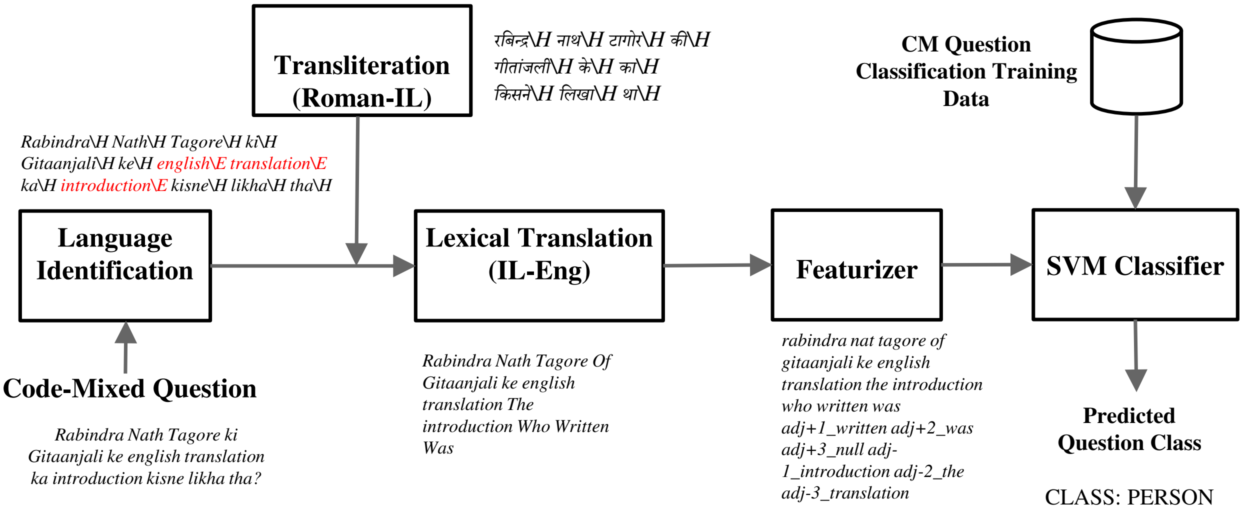
Transliteration (Roman-IL)
रबिन्द्र\H नाथ\H टागोर\H की\H गीतांजली\H के\H का\H किसने\H लिखा\H था\H
CM Question Classification Training Data
Rabindra\H Nath\H Tagore\H ki\H Gitaanjali\H ke\H english\E translation\E ka\H introduction\E kisne\H likha\H tha\H
Language Identification
Lexical Translation (IL-Eng)
Featurizer
SVM Classifier
Code-Mixed Question
Rabindra Nath Tagore ki Gitaanjali ke english translation ka introduction kisne likha tha?
Rabindra Nath Tagore Of Gitaanjali ke english translation The introduction Who Written Was
rabindra nat tagore of gitaanjali ke english translation the introduction who written was adj+1_written adj+2_was adj+3_null adj-1_introduction adj-2_the adj-3_translation
Predicted Question Class
CLASS: PERSON

Figure 1: Architecture of Question Classification Engine for CM Language

Coarse	Fine
Abbreviation	Abbreviation, Explanation
Entity	Animal, Body, Color, Creative, Currency, Disease, Event, Food, Instrument, Language, Letter, Other, Plant, Product, Religion, Sport, Substance, Symbol, Technique, Term, Vehicle, Word
Description	Definition, Description, Manner, Reason
Human	Group, Individual, Title, Description
Location	City, Country, Mountain, Other, State
Numeric	Code, Count, Date, Distance, Money, Order, Other, Period, Percent, Speed, Temperature, Size, Weight

Table 2: Hierarchical Question Ontology defined by Li and Roth [15]. Coarse grained categories are italicized. Fine grained categories are a further division of each Coarse grained category.

gave the code-mixed version of the question in roman script. Hence, we collected around 3K different code-mixed variants of the initial 1K question set. In order to avoid repetition of questions and also any individual bias, for each question, we randomly picked a code-mixed version from one of the three annotations and made the current dataset of 1K questions. The details of the dataset are shown in Table 1. We can see that on an average about one-third of a given CM sentence comprises of English words. Also, we notice that even though English words are used, the structure of the sentence remains unaffected. This property plays a crucial role in designing features for QC.

4. QUESTION CLASSIFICATION SYSTEM

In this section, we describe our QC system in detail. For classification, we used the taxonomy of question classes proposed by Li and Roth [15] shown in Table 2. The architecture of our system is shown in Figure 4. The central idea of our system is to uniformly represent all the features in a single language to enable better generalization during learning. Given a question in CM language, we initially identify the language of each word in the question. Later, each non-English word is transliterated from roman script to Indian Language (IL) script. The words in IL script are then translated into English using Google Translate API. Once all the question words are uniformly translated into English, we transform it into a feature vector for performing QC using SVMs.

For language identification, we use the Language Identification tool that was developed for FIRE 2014 shared task[4] which has an accuracy of 79.2% . This system is based on a SVM trained for each language pair which makes use of character-based smoothed n-gram language models trained separately for each language. The output of the system is the sequence of code mixed words annotated with their corresponding language.

Once the language is identified at a word-level, we transliterate the non-English words from roman script to IL script (Devanagari, in case of Hindi). This is done so that we can use the off-the-shelf translation tools from IL-English which accept words only in Devanagari script. For this, we use the transliteration engine developed by Chinnakotla *et. al.* [5] and take the top output from this. The accuracy of the transliteration system was around 70% at rank 1.

The transliterated IL words are individually fed to a lexical-translation engine which produces the corresponding English translated word. At present, we only restrict ourselves to word level translation since in CM text, the phrase-level structures usually contain a mix of Hindi and English words. Since, the translation is at word-level and without context, we use the most frequent translation of the word which might be error-prone. We use the Goslate [2] - a free Google Translate API for translation. We found that the world-level accuracy of the system is around 79%. The Hindi words in the code mixed data are replaced by these translated words. By the end of this step, we will have the Hindi-English code mixed question transformed into a sequence of English words.

4.1 SVM Based Classification

We use SVMs for learning the QC model. We train a one-vs-rest SVM model for each of the coarse-grained and fine-

[2] https://pypi.python.org/pypi/goslate

System	Accuracy (Coarse-Grained)	% Impr. on Baseline	Accuracy (Fine-Grained)	% Impr. on Baseline
Baseline	55%	-	43.50%	-
Baseline + ADJ	57%	2	43.50%	
Baseline + ADJ + Translation + Linear Kernel	62%	***7***‡	44.50%	1
Baseline + ADJ + Translation + RBF Kernel	63%	***8***‡	45.00%	1.5

Table 3: Overall Question Classification Results on Coarse-Grained and Fine-Grained Classes. Results marked as ‡ indicate that improvement was statistically significant at 95% confidence level ($\alpha = 0.05$) when tested using a paired two-tailed t-test.

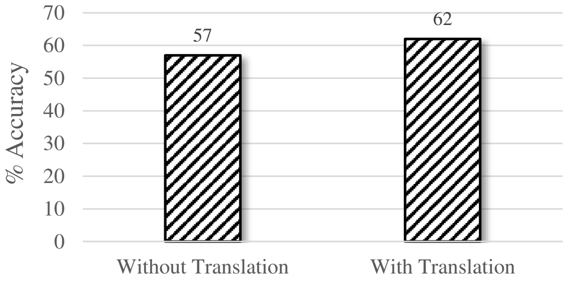

Figure 2: Contribution of Translation in our QC System

grained categories. Given a translated question, we transform it into a feature vector and pass it through all the SVMs and output the SVM class which outputs the maximum score. We use uni-gram word features from all the words in the question. In English-Hindi CM questions, the words adjacent to the question word carry a lot of information about the expected answer. For example, consider the following CM question: *"U.S. system ka* kaunsa unit *0.45 kgs ke barabar hai?" (English Translation: Which unit of U.S. system is equivalent to 0.45 Kgs).* Here the question word is *"kaunsa" (English Translation: which)* and the adjacent word *"unit"* tells that the expected answer type is a *"Unit"*. Similarly, for *"Purushon mein sabse frequently paya jaane wala cancer kaunsa hai?" (English Translation: In men, which cancer is found most commonly?).* The answer type is found in word *"cancer"* adjacent to the question word *"kaunsa" (English Translation: which).* Unlike English, in Hindi, a Wh-nominal does not have to undergo any movement in the sentence. We leverage this phenomenon by including adjacency features.

For each question, besides uni-gram word features, we also include adjacency features for the three surrounding words from the question word. For example, for the same question mentioned above - *"Purushon mein sabse frequently paya jaane wala cancer kaunsa hai?" (English Translation: In men, which cancer is found most commonly?)*, which is transformed after transliteration and translation into *"men in most frequently found go of cancer which is"*, the following adjacency features will be included - *ADJ-3_go, ADJ-2_of, ADJ-1_cancer, ADJ+1_is, ADJ+2_null, ADJ+3_null.* The adjacency feature value is *null* if no word is found in that position.

5. EXPERIMENTS AND RESULTS

In this section, we describe our experimental set-up, evaluation metrics and results. We train a uni-gram feature based classifier on the original CM question words. We consider this to be baseline with which we compare our results. For tuning the parameters of SVM (C and γ for RBF Kernel), we performed five-fold cross-validation on the dataset and chose the parameters which produced the highest average accuracy. The best values turned out to be - C (420 - Coarse Grained, 460 - Fine Grained) and γ for RBF (0.0002 for both).

5.1 Results and Discussion

The overall results are presented in Table 3. The baseline accuracy with just uni-gram features on the original question words is 55% for coarse-grained and 44% for fine-grained. The final accuracy of our system is 63% for coarse-grained and 45% for fine-grained. The idea of translating the word features into a single language and adding adjacency features helps in significantly improving the accuracy in the coarse-grained category which reaches 62%. Figure 2 shows the improvement due to translation. However, these features didn't help as much in improving the accuracy in fine-grained category. The improvement in the accuracy of fine-grained class was only 1.5% over the baseline. Since, the question word along with a few surrounding words is enough to determine the coarse-level category in most cases, it is relatively easy to improve the accuracy there. However, for accurately identifying the fine-grained class, semantic features and understanding are indispensable. In both cases, we noticed improvements due to changing the kernel from linear to RBF (Radial Basis Functions).

QUESTION	CURRENT RESULT	EXPECTED RESULT	ANALYSIS
Question: Tania Sachdev aur Koneru Chawla bharat ke liye kaunsa khel khelte hai? **Gloss Translation:** Tania Sachdev and Koneru Chawla India the for which sport play is the? **Meaning:** Which sport do Tania Sachdev and Koneru Chawla play for India?	ENTITY	ENTITY	The Q-word "kaunsa" is translated coorectly to "which", which is an important feature for ENTITY classification.
Question: Richard Attenborough ki oscar jeetne wali film "gandi" ke geet George Fenton Sahit kisne banaye the? **Gloss Translation:** Richard Attenborough of the oscar win the film "gandi" the song George Fenton Sahit who build were a? **Meaning:** Who created the song George Fenton Sahit for Richard Attenborough's oscar winning film "gandi" ?	HUMAN	HUMAN	The Q-word "kisne" is translated correctly to "who", which is an important feature for HUMAN classification.
Question: Indo-China war ke shaheedon ko shradhhanjali dene ke liye, Kavi Pradeep ne kaunse gana ki rachna ki thi? **Gloss Translation:** Indo-China war the fallen to tribute offering the for, poet Pradeep the what song of the composition of the was? **Meaning:** To give a tribute to the fallen of Indo-China war, which song is composed by poet Pradeep?	ENTITY: Creative	ENTITY: Creative	The lexical features of "song" and "composed" are features that classify the Answer Type to be ENTITY:Creative.
Question: Kaun se Taapmaan pe Fahrenheit aur Celsius scale samaan reading dikhate hai? **Gloss Translation:** who by the temperature pay Fahrenheit and Celsius scale equal reading show is the? **Meaning:** At which temperature, Fahrenheit and Celsius scale show the same reading?	HUMAN	NUMERIC	Lexical translation of Q-word "kaun" is "who", which classified the question as HUMAN.
Question: Narmada kaun si pahadi sharankla me nirmit hoti hai? **Gloss Translation:** Narmada who c the hill srncle in built up there is the? **Meaning:** What is the mountainous series where Narmada is situated?	HUMAN: Individual	LOCATION: Mountain	The Q-word "kaun" is translated as "who". This lexical feature "who" classifies this question to answer type "HUMAN:Individual".

Table 4: Qualitative Analysis of the results of our QC System

Within coarse-grained class, maximum accuracy was obtained in the HUMAN (89.19%) and LOCATION (65.22%) categories because they are easy to determine based on a few words such as HUMAN (who, player, character, cricketer etc.) and LOCATION (country, state, place etc.). However, ABBREVIATION (0%) and DESCRIPTION (0%) were the hardest classes. Since, the training data is only 1000 samples, some of the these classes were not adequately represented. Due to this, the model could not learn anything significant. Similarly, in fine-grained category, classes such as ENTITY:Substance (100%), HUMAN:Individual (90.48%), LOCATION:City (69%) were some of the better performing ones whereas a number of classes in the DESCRIPTION, ABBREVIATION and ENTITY received 0%.

5.2 Qualitative Analysis

In Table 4, we present a qualitative analysis of our results. We show both positive and negative examples where our algorithm performs better or worse than the baseline. In the first example, the presence of question word *"kaunsa (which)"* leads to classifying it as ENTITY. In the second case, the question word *"kisne (who)"* is a strong indicator for the HUMAN class. The third example is a case where the model leverages both question word *"kaunsa (which)"* and adjacency features *"composed, song"* to correctly classify it was ENTITY:creative.

In the negative example rows, the first example was misclassified because the question word *"kaunsa (which)"* was written as two separate words *"kaun sa "* during annotation. Due to this, we fired the *"kisne (who)"* feature which is a strong indication of HUMAN. The second example also has the same variation in annotation. However, besides misfiring for *"kisne (who)"*, we also could not identify the relation between hill and mountain due to lack of semantic features.

6. CONCLUSION AND FUTURE WORK

Code-Mixing is a natural phenomenon observed in multilingual societies. It speeds-up communication and allows wider varieties of expression due to which it has become a popular mode of expression in social media conversations like Facebook and Twitter. However, current QA research and systems only support expressing a question in one single language which is an unrealistic and hard proposition. In this paper, we presented our initial efforts towards the development of a full-fledged QA system in CM language. Processing CM language also presents hard challenges due to its linguistic complexities and the lack of language resources required to process them.

In this paper, we learn a basic Support Vector Machine (SVM) based QC system for English-Hindi CM questions using word-level resources such as language identification, transliteration and lexical translation. Although, we improved over the uni-gram baseline, we still have a lot of ground to cover. As part of future work, we plan to include semantic features based on WordNet for improving the accuracy in fine-grained classes.

7. REFERENCES

[1] B. Alex. *Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing.* Ph.D Thesis, School of Informatics, The University of Edinburgh, UK, 2008.

[2] P. Auer. *Code-Switching in Conversation: Language, Interaction and Identity.* Routledge, 2013.

[3] U. Barman, A. Das, J. Wagner, and J. Foster. *Code mixing: A Challenge for Language Identification in the Language of Social Media.* In ACL 2014, pages 13–23.

[4] I. A. Bhat, V. Mujadia, A. Tammewar, R. A. Bhat, and M. Shrivastava. *IIIT-H System Submission for FIRE 2014 Shared Task on Transliterated Search.* In Forum for Information Retrieval Evaluation (FIRE), 2014.

[5] K. M. Chinnakotla and P. O. Damani. *Experiences with English-Hindi, English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009.* In ACL 2009, pages 44–47.

[6] G. Chittaranjan and Y. Vyas. *Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System.* In EMNLP 2014, page 73.

[7] A. Dey and P. Fung. *A Hindi-English Code-Switching Corpus.* In the 9th International Conference on Language Resources and Evaluation (LREC), 2014.

[8] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. *Building Watson: An Overview of the DeepQA Project.* AI magazine, 31(3):59–79, 2010.

[9] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso. *Query Expansion for Mixed-Script Information Retrieval.* In SIGIR '14, pages 677–686, ACM, 2014.

[10] T. Hidayat. *An analysis of Code Switching used by Facebookers.* 2012.

[11] L. Hirschman and R. Gaizauskas. *Natural Language Question Answering: The view from here.* Natural Language Engineering, 7:275–300, 12 2001.

[12] N. Jain, I.-H. LTRC, and R. A. Bhat. *Language Identification in Code-Switching Scenario.* In EMNLP 2014, page 87.

[13] N. M. Kamwangamalu. *Code-Mixing Across Languages: Structure, Functions, and Constraints.* PhD thesis, University of Illinois at Urbana-Champaign, 1989.

[14] D. C. S. Li. *Cantonese-English Code-Switching Research in Hong Kong: A Y2K Review.* World Englishes, 19(3), 2000.

[15] X. Li and D. Roth. *Learning Question Classifiers.* in COLING '02, 2002.

[16] D. Metzler and W. B. Croft. *Analysis of Statistical Question Classification for Fact-Based Questions.* Information Retrieval, 8(3):481–504, May 2005.

[17] L. Milroy and P. Muysken. *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching.* Cambridge University Press, 1995.

[18] A. Moschitti, J. Chu-Carroll, S. Patwardhan, J. Fan, and G. Riccardi. *Using Syntactic and Semantic Structural Kernels for Classifying Definition Questions in Jeopardy!* In Association for Computational Linguistics, 2011, pages 712–724.

[19] Myers-Scotton, Carol. *Duelling Languages: Grammatical Structure in Code-Switching.* Claredon. Oxford., 1993.

[20] Myers-Scotton, Carol. *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes.* Oxford University Press, 2002.

[21] H. K. San. *Chinese-Engilsh Code-Switching in Blogs by Macao Young People.* Master's Thesis, The University of Edinburgh, UK, 2009.

[22] T. Solorio and Y. Liu. *Part-of-Speech Tagging for English-Spanish Code-Switched Text.* In Association for Computational Linguistics, 2008, pages 1051–1060.

[23] P. Virga and S. Khudanpur. *Transliteration of Proper Names in Cross-Lingual Information Retrieval.* In Association for Computational Linguistics, 2003, pages 57–64.

[24] Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. *POS Tagging of English-Hindi Code-Mixed Social Media Content.* In EMNLP 2014 pages 974–979, October 2014.

[25] J. Xu, Y. Zhou, and Y. Wang. *A Classification of Questions using SVM and Semantic Similarity Analysis.* In IEEE, 2012, pages 31–34.

[26] D. Zhang and W. S. Lee. *Question Classification using Support Vector Machines.* In ACM, 2003, pages 26–32.