Amrita_CEN@MSIR-FIRE2016: Code-Mixed Question Classification using BoWs and RNN Embeddings

Anand Kumar M
Center for Computational Engineering and Networking(CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham
Amrita University
m_anandkumar@cb.amrita.edu

Soman K P
Center for Computational Engineering and Networking(CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham
Amrita University
kp_soman@amrita.edu

ABSTRACT

Question classification is a key task in many question answering applications. Nearly all previous work on question classification has used machine learning and knowledge-based methods. This working note presents an embedding based Bag-of-Words method and Recurrent Neural Network to achieve an automatic question classification in the code-mixed Bengali-English text. We build two systems that classify questions mostly at the sentence level. We used a recurrent neural network for extracting features from the questions and Logistic regression for classification. We conduct experiments on Mixed Script Information Retrieval (MSIR) Task 1 dataset at FIRE2016[1]. The experimental result shows that the proposed method is appropriate for the question classification task.

CCS Concepts

• **Information systems~Clustering and classification** • *Information systems~Question answering* • **Computing methodologies~Learning latent representations**

Keywords

Question Classification ; BoW ; Recurrent Neural Nets ; Embeddings ; Code-Mixed Script

1. INTRODUCTION

Question answering systems can be viewed as an inevitable element of information retrieval systems, allowing users to ask questions in a natural language text and receive brief answers. Earlier research has shown explicitly that the correct classification of questions to the expected answer type is necessary to any successful question answering system. Question classification is to recognize the answer-type automatically to a given query written in the natural language text. For example, the query, "*What is the Capital of India?*", the task of a question classification system is to recognize the type "*Location*" to this question because the expected answer to this query is a named entity of type "*Location*". Classification of queries is also treated as an answer type prediction since the type of the answer is predicted. Many existing question answering systems used manually built sets of rules to map a question to a correct type, which is the language specific, not efficient in maintaining and upgrading. Machine learning approaches are often used to identify the expected answer types. The motivation of using the advantage of Recurrent Neural Network (RNN) based embedding is that RNN captures the contextual information in a better way.

2. RELATED WORKS ON QUESTION CLASSIFICATION

Basically, there are two different methods commonly used in question classification: knowledge-based and machine learning based. There are also some combined approaches which connect rule-based and the machine learning approaches (Huang et. al., 2008; Silva et. al., 2011; Ray et. al., 2010) [1,2,7]. Rule-based methods classify the questions with hand-crafted rules (Hull, 1999; Prager et. al., 1999) [3,4]. However, these approaches affected from too many rules (Li and Roth, 2004) [5] and only perform well on a particular dataset. Recent NLP research for Indian languages moving towards social media content which is informal and often code-mixed. Researchers focused on developing conventional Natural Language Processing (NLP) applications for handling Social media content. Standard shared tasks and workshops like FIRE and ICON[2] Tools contest are giving preferences to this new genre text. The large-scale use of code-mixed style in social media platforms motivates the researchers to carry out this type of research in Indian languages. The significant number of research is going on in social media text and code-mixed text. Notable areas are, language identification [8] , question answering [11] , POS tagging [15], polarity detection [13] and entity extraction for Indian languages [12, 14]. Barman et. al. [9] presented the challenges of Language Identification in code-mixed text and they claimed that code-mixing is common among users who are multilingual. Vyas et. al. [15] discussed the efforts taken to POS tag social media content from English-Hindi code-mixed text while trying to address the complexities of code-mixing. The impact of code-mixing on the effectiveness of information retrieval has been discussed by Gupta et. al. [16] in query expansion for mixed-script and code-mixed queries. Recently, Banerjee et. al. (2015) [17, 18] formally introduced the code-mixed cross-script question answering as a research problem. Banerjee et. al. [19] explains the use of growing user generated content to serve as information collection source for the question answering task on a low-resource language for the first time and explained their cross-script code-mixed question answering corpus

[1] http://fire.irsi.res.in/fire/2016/home

[2] http://amitavadas.com/Code-Mixing.html

3. TASK DESCRIPTION

The code-mixed cross-script question classification is subtask-1 in shared task on Mixed Script Information Retrieval (MSIR[3]) at FIRE 2016 [23].

Let, Q = {q_1, q_2 , . . . , q_n} be a set of factoid questions written in Code-mixed Bengali-English Text (Romanized Bengali along with English). Let T = {t_1, t_2,...,t_n} be the set of question types. The task is to classify each given question q ∈ Q into one of the predefined coarse-grained question type t ∈ T. Example for code-mixed question classification task is given below,

Question: *last volvo bus kokhon chare ?*
[When is the last Volvo bus..]

Question Type: TEMPORAL

The number of queries, the total number of words and average words per query in Training and testing data are illustrated in Table 1. Totally, 9 different coarse-grained question types are used in this question classification task. The various question types and their corresponding frequency in training data are shown in Table 2. This table also reveals the percentage of each question type in training data. More than 65% of the training data set belongs to 4 primary query types which are Organization, Temporal, Person, and Number.

Table 1. MSIR Subtask-1 data facts

Model	Queries	Total Words	Average Words
Training	330	1756	5.321
Testing	120	858	7.15

Table 2. Question types and their counts

Types	Count	Percentage
ORG	67	20.3
TEMP	61	18.5
PER	55	16.7
NUM	45	13.6
LOC	26	7.9
MNY	26	7.9
DIST	24	7.3
OBJ	21	6.4
MISC	5	1.5
	330	1

4. QUESTION CLASSIFICATION FOR CODE-MIXED BENGALI ENGLISH TEXT

We have submitted two runs in the question classification for code-mixed text. In the first run, we used the traditional BoW model with logistic regression. In order to apply regression, we represent each word-type to random vectors of floating numbers using the *categorical variable* function in TensorFlow [10]. In the second run, we tried with on Recurrent Neural Network embedding with logistic regression. Since the dataset is a very small the RNN based method trails traditional methods. Even though RNN based method accuracy is less compared with other methods, the performance of RNN based embedding is significant for the very limited data. This gives an anticipation for applying RNN for code-mixed NLP related task.

4.1 Bag-of-Words Model for Question Classification (Run1)

We developed a question classification system with a BoW model using TensorFlow [10]. Here the maximum word length is fixed as 15 and embedding size as 50. Each word-type in the query is converted into 50-dimensional vectors. For the given 330 queries in the training set, we formed an input matrix of size 330 x 15, and for each word we substitute the random word embeddings (categorical word representation) and finally the size of the input tensor is 330 x 15 x 50. We used the max pooling concept and choose the maximum value across the max word length of 15. This reduced the tensor to the matrix of size 330 x 50 which is considered as query embeddings and given to logistic regression classifier with default parameters. Finally, we used Arg-max function to choose the best question type.

4.2 Recurrent Neural Net based Question Classification System (Run2)

Recurrent Neural Networks (RNNs) are successful models that have shown prominent improvement in many NLP applications. The idea behind RNNs is to make use of sequential information [21]. If you want to predict the subsequent word in a sentence you completely know which words appeared before it. RNNs are called recurrent because they carry out the same task for every element of a sequence, with the output being depended on the previous computations.

In our second submission, we developed a Recurrent Neural Network based question classification system using TensorFlow [10]. We followed the same produce of Run1 for creating the input tensor of size (330 x 15 x 50). This initial 15 x 50 matrix embedding of each query is reduced to 50-dimensional embedding vectors. This initial embedding vector is given to Gated Recurrent Unit, or GRU, a slightly variation on the LSTM introduced by [22].The resulting model is simpler than standard LSTM models, and has been growing increasingly popular. Finally, take encoding of the last step and pass it as features for logistic regression for training.

5. EXPERIMENTS AND RESULTS

In this section, detailed cross-validation results and the accuracy has been given by the task organizers are elucidated.

5.1 Cross-validation Results

We randomly split the 330 queries in training set into 281 and 49 and named as training and development set respectively. This data set used for validating our methods with two different parameters, embedding size, and maximum query length. We varied the maximum document size to 10, 15, 20, 25, and 30. We used only two different embedding sizes, 50 and 100. We tried BoW and RNN based methods for developing the code-mixed question classification system. Figure 1 explains the comparison between the BoW and RNN based methods with different query length and embedding size. We fixed the query length as 15 and embedding size as 50 in our experiments.

[3] https://msir2016.github.io/

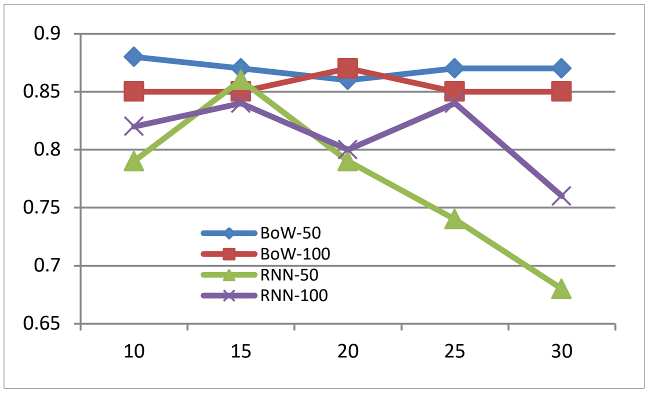

Figure 1. Cross-validated results with different query length and embedding size

5.2 MSIR Sub Task-1 Results

Here, the accuracy has been given by the task organizers are explained. Organizers evaluated submitted systems based on the accuracy. Overall performance and in-depth accuracy per question type are also released by the organizers [20]. The overall accuracy of our submission is shown in Table 3. The highest accuracies of other teams are shown in Table 4. IINTU team positioned to first followed by Anuj, BITS, and our Team (Amrita_CEN). Figure 2 explains the query types and their corresponding accuracy for our submissions. It is interesting to note that RNN based model outperforms the BoW in the ORGANIZATION type questions which count is higher in the training dataset. At the same time, the OBJ and MISC type, which are less in a count, accuracies are comparably low in RNN based model.

Table 3. Overall Accuracy of our two submissions

Runs	Run1 (BoW)	Run2 (RNN)
Correct	145	133
Incorrect	35	47
Accuracy	80.55556	73.8888889

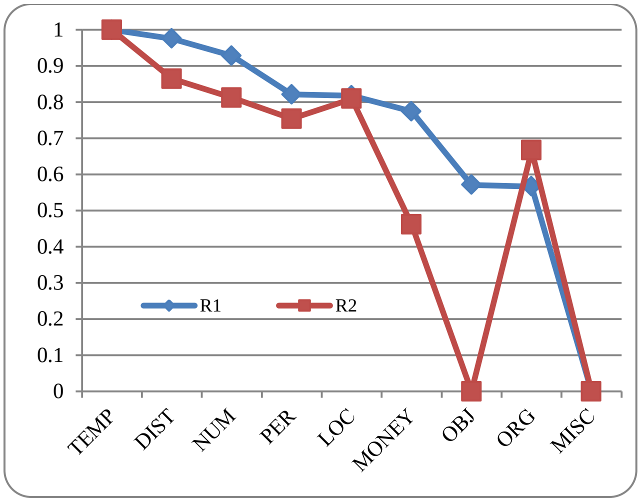

Figure 2. Query types and accuracies.

Table 4. Top accuracies of team irrespective of run

Runs	Correct	Incorrect	Accuracy
AmritaCEN	**145**	**35**	**80.55556**
AMRITA-CEN-NLP	143	37	79.44444
Anuj	146	34	81.11111
BITS_PILANI	146	34	81.11111
IINTU	**150**	**30**	**83.33333**
IIT(ISM)D*	144	36	80
NLP-NITMZ	142	38	78.88889

6. CONCLUSION

Question classification is an inevitable module in the question answering system. This working note presents code-mixed question classification system using BoWs and RNN embeddings. To our knowledge, this is the first time that RNN embedding is applied to question classification task. Since the training corpus is small and unavailability of unsupervised code-mixed data, the performance of the RNN based system trails the traditional BoWs method. The performance of the RNN based embedding is not that poor and paves the way in future to apply for code-mixed script analysis. It is exciting to note that RNN based model outperforms the BoWs in the ORGANIZATION type questions which occurrence is high in the training dataset. At the same time for OBJ and MISC type queries, which are less in a count, accuracies are comparably low in RNN based model. Finally, our team (Amrita_CEN) positioned third place in the overall performance.

7. REFERENCES

[1] Zhiheng Huang, Marcus Thint, and Zengchang Qin. Question classification using headwords and their hypernyms. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP ’08), pages 927–936,2008.

[2] Joao Silva, Luısa Coheur, Ana Mendes, and Andreas Wichert. From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 35(2):137–154, 2011.

[3] E. Voorhees. The TREC-8 Question Answering Track Report. In Proceedings of the 8thText Retrieval Conference (TREC8), pp. 77-82, NIST, Gaithersburg, MD, 1999.

[4] John Prager, Dragomir Radev, Eric Brown, and Anni Coden. The use of predictive annotation for question answering in trec8. In NIST Special Publication 500-246:The Eighth Text Retrieval Conference (TREC), pages 399–411. NIST, 1999.

[5] Xin Li and Dan Roth. 2004.Learning question classifiers: The role of semantic information. COLING,pp. 556-562.

[6] Zhiheng Huang, Marcus Thint, and Asli Celikyilmaz.2009 Investigation of question classifier in question answering . EMNLP , pp. 543-550.

[7] Santosh Kumar Ray, Shailendra Singh, and B. P. Joshi. A semantic approach for question classification using wordnet and Wikipedia. Pattern Recogn. Lett.,31:1935–1943, 2010.

[8] Rahul Venkatesh Kumar, R.M., Anand Kumar, M., Soman, K.P. AmritaCEN-NLP @ FIRE 2015 language identification for Indian languages in social media text (2015) CEUR Workshop Proceedings, 1587, pp. 26-28.

[9] Barman, A. Das, J. Wagner, and J. Foster, "Code Mixing: A Challenge for Language Identification in the Language of Social Media," in First Workshop on Computational Approaches to Code Switching, 2014, pp. 21–3

[10] Abadi, Martın, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* (2016).

[11] 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 the 24th International Conference on World Wide Web* (WWW '15 Companion). ACM, New York, NY, USA, 853-858.

[12] Devi, G.R., Veena, P.V., Kumar, M.A., Soman, K.P. Entity Extraction for Malayalam Social Media Text Using structured Skip-gram Based Embedding Features from Unlabeled Data (2016) Procedia Computer Science, 93, pp. 547-553.

[13] Nivedhitha, E., Sanjay, S.P., Anand Kumar, M., Soman, K.P. Unsupervised word embedding based polarity detection for Tamil tweets (2016) International Journal of Control Theory and Applications, 9 (10), pp. 4631-4638.

[14] Anand Kumar, M., Se, S., Soman, K.P. AMRITA-CEN@FIRE 2015: Extracting entities for social media texts in Indian languages (2015) CEUR Workshop Proceedings, 1587, pp. 85-88.

[15] 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

[16] 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

[17] Banerjee, S., Bandyopadhyay, S.: Ensemble Approach for Fine-Grained Question Classification in Bengali. In: Proceedings of 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC), Taiwan, pp. 75–84 (2013).

[18] Banerjee, S., Bandyopadhyay, S.: An Empirical Study of Combining Multiple Models in Bengali Question Classification. In: Proceedings of International Joint Conference on Natural Language Processing (IJCNLP), Japan, pp. 892–896 (2013).

[19] Somnath Banerjee, Sudip Kumar Naskar, Paolo Rosso, and Sivaji Bandyopadhyay. The first cross-script code-mixed question answering corpus. In Modelling, Learning and mining for Cross/Multilinguality Workshop, 38th European Conference on Information Retrieval (ECIR), pages 56-65, 2016.

[20] Somnath Banerjee and Sudip Naskar and Paolo Rosso and Sivaji Bandyopadhyay and Kunal Chakma and Amitava Das and Monojit Choudhury, MSIR@FIRE: Overview of the Mixed Script Information Retrieval, Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, CEUR Workshop proceedings, CEUR-WS.org, 2016.

[21] http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/

[22] Cho, Bahdanau, Dzmitry, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate.arXiv:1409.0473 [cs.CL], September 2014.

[23] S. Banerjee, K. Chakma, S. K. Naskar, A. Das, P. Rosso, S. Bandyopadhyay, and M. Choudhury. Overview of the Mixed Script Information Retrieval at FIRE. In Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation, Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings. CEUR-WS.org, 2016.