MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi

Deepak Gupta, Surabhi Kumari, Asif Ekbal and Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India-801106
{deepak.pcs16, surabhi.mtmc16, asif, pb}@iitp.ac.in

Abstract

In this paper, we assess the challenges for multi-domain, multi-lingual question answering, create necessary resources for benchmarking and develop a baseline model. We curate 500 articles in six different domains from the web. These articles form a comparable corpora of 250 English documents and 250 Hindi documents. From these comparable corpora, we have created 5,495 question-answer pairs with the questions and answers, both being in English and Hindi. The question can be both factoid or short descriptive types. The answers are categorized in 6 coarse and 63 finer types. To the best of our knowledge, this is the very first attempt towards creating multi-domain, multi-lingual question answering evaluation involving English and Hindi. We develop a deep learning based model for classifying an input question into the coarse and finer categories depending upon the expected answer. Answers are extracted through similarity computation and subsequent ranking. For factoid question, we obtain an MRR value of 49.10% and for short descriptive question, we obtain a BLEU score of 41.37%. Evaluation of question classification model shows the accuracies of 90.12% and 80.30% for coarse and finer classes, respectively.

Keywords: Multi-lingual Question answering, Answer extraction, Neural network, Question classification

1. Introduction

Question answering (QA) is an important area with a wide range of applicability in various Natural Language Processing (NLP) tasks, such as information retrieval, information extraction etc. The aim of a QA system is to automatically extract/generate the answer(s) for a given question from the data repository (e.g., web, document etc.). In a QA system questions are formulated in natural languages and answers are also dealing with natural languages. Unlike search engine instead of extracting information, here a QA system usually focus on extracting relevant and precise answer(s). In other words, we can say QA system is the extended modification in the search engine. For achieving QA system, usually three subprocesses are followed i) question classification ii) document(s)/passage(s) extraction and iii) appropriate answer(s) extraction. Most of the existing works focus on retrieving answers in the same language in which the questions are posed. However, with the rapid growth of multilingual contents on the web, it is necessary to build an automated system that retrieves information from the documents written in multiple languages.

Posing questions in multiple languages and retrieving answers accordingly is known as Multilingual Question Answering (MQA), which has emerged as an interesting research area in QA. This enables the situation where the question could be in a different language from the language of documents where the answer(s) lies. This allows users to interact in their native languages, facilitating multilingual information access, which is immensely useful in a country like India. MQA system can contribute to conserving the endangered languages which are losing their existence and prestige as mentioned in (Knott et al., 2001). Hindi is a widely spoken language in India, and in terms of native speakers, it ranks fourth all over in the world. In India, a sum of 53.60% of total population speak Hindi as compared to English (12.18%). English, on the other hand, is used for all kinds of official communications. There is often need to exchange information from Hindi to the other popular language(s) such as English.

In recent year several multilingual and cross lingual QA systems have been built. These systems are seeking to overcome the issue of accessing and retrieving information in multiple languages. The majority, however, are based on translating relevant sections of the question – usually with the aid of machine translation system - which is used to access to a collection containing relevant information. The basic goal of MQA framework is to set up a common system to evaluate both bilingual and cross-lingual question answering that process queries in either Hindi or English language and retrieve answer in either language from documents in Hindi or English. The main motivations and/or contributions of the current work are as follows:

1. Most of the existing works are in resource-rich languages such as the English. Indian languages are resource-scarce, and developing a multi-lingual QA system involving English and Hindi has the benefit of utilizing resources and tools available for the resource-rich language like English.

2. Creating a benchmark setup for multi-lingual QA involving Indian languages will be beneficial for multi-lingual information access. To the best of our knowledge, this is the very first attempt in this direction.

3. Question classification is an important step in Question-Answering (QA). We propose a method based on deep Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for question classification.

Article (English)		Article (Hindi)		
Shimla is the capital of Himachal Pradesh and was also the summer capital in pre-Independence India. Shimla derives its name from Shyamala Devi, an incarnation of the goddess Kali, whose temple existed in the dense forest covering the Pradesh Jakhu Hill in the early 19th century. Shimla is the capital of Himachal in pre-Independence India. Covering an area of 25 sq km at a height of 7,238 ft, Shimla is surrounded by pine, deodar and oak forests.		शिमला, एक ख़ूबसूरत हिल स्टेशन है जो हिमाचल प्रदेश की राजधानी है। समुद्र की सतह से 2202 मीटर की ऊँचाई पर स्थित इस जगह को 'समर रिझूज' और 'हिल स्टेशनों की रानी' के रूप में भी जाना जाता है। वर्तमान का शिमला जिला 1972 में निर्मित किया गया था। इस जगह का यह नाम 'माँ काली' के दूसरे नाम 'श्यामला' से व्युत्पन्न है। जाखू, प्रॉस्पैक्ट,ऑब्सर्वेटरी, एलीसियम और समर इस जगह की महत्वपूर्ण पहाड़ियाँ हैं।		
Question (English)	**Answer (English)**	**Question (Hindi)**	**Answer (Hindi)**	**Answer Availability**
What is the capital of Himachal Pradesh?	Shimla	हिमाचल प्रदेश की राजधानी क्या है?	शिमला (*Shimla*)	EN and HI
How much area is covered by Shimla?	25 sq km	शिमला का क्षेत्रफल कितना है?	-	EN
From which goddess name the Shimla word derived?	Shyamala Devi	शिमला शब्द किस देवी के नाम से लिया गया है?	श्यामला (*Shyamala*)	EN and HI
What is the height of Shimla from sea level?	7,238 ft	समुद्र के स्तर से शिमला की ऊंचाई क्या है?	2202 मीटर (*2202 meter*)	EN and HI
In which year Shimla established?	-	शिमला की स्थापना किस साल हुई?	1972	HI

Table 1: An example of comparable articles from English and Hindi and set of question-answer pairs created from the given articles.

Reasoning Types	Question & Answer Sentence	Descriptions
Interlingual semantic word	**Q:** What was introduced in the Sixth Five-Year Plan? **Answer Sentence:** छठी पंचवर्षीय योजना (1980-1985) ने भी आर्थिक उदारीकरण की शुरुआत को चिन्हित किया । (*Chhati panchvarshiya yojna (1980-1985) ne bhi arthik udarikaran ki shuruaat ko chinhit kiya.*)	An interlingual semantic word knowledge is required to provide the answer. Only translation may not help to extract the answer.
Multiple sentence reasoning	**Q:** What is the part of the Adam's Bridge? **Answer Sentence: Pamban Island** is situated in the Gulf of Mannar between India and Srilanka. It is a part of the Adam's Bridge.	There is an anaphora, or fusion of multiple sentences is required to answer.
Word matching	**Q:** What is the collection of Vedic hymns or mantras called? **Answer Sentence:** The collection of Vedic hymns or mantras is called the **Samhita**.	Word matching between question and answer sentence can provide the answer.
Interlingual syntax variation	**Q:** प्लेग किसके लिए जिम्मेदार है ? (*Plaque kiske liye jimmedar hai?*) **Answer Sentence:** Plaque deposited on the teeth and under the gumline irritates the gum tissue, and causes **gingivitis**.	Syntactic structure of question and answer sentence vary across the language.
Single Sentence	**Q:** What is Sustainable development? **Answer Sentence:** Scale defines the relationship between distance on a map and on the earth's surface. Sustainable development: **Development that does not exploit resources more rapidly than the renewal of those resources**, ...	Answer of short descriptive question can be a single sentence from a paragraph.
Multiple Sentence	**Q:** Why India is considered to be an eastern country? **Answer Sentence:** ... **India lies east of the Prime Meridian. Therefore India is considered to be an eastern country because of its situation in the Eastern Hemisphere**....	Answer of short descriptive question can be multiple sentence from a single or multiple paragraph.

Table 2: The set of possible reasoning types with the corresponding question-answer pair example and descriptions. Reasoning types show the difficulty of the question in terms of finding their answer. The answer in answer sentence has been shown in bold font.

Problem Definition

Given a natural language question Q (factoid or short descriptive) in either language, English or Hindi. The QA system should return the answer A for the given question Q from the comparable English and Hindi documents. The returned answer should be in the same language[1] as the question Q.

2. Related Work

In literature, we found a very few existing works related to question-answering (QA) in Hindi or English (Sekine and Grishman, 2003; Kumar et al., 2005; Sahu et al., 2012; Stalin et al., 2012). However, none of these focuses on multilingual QA. (Kumar et al., 2005) implemented the Hindi search engine. The task of the search engine is to retrieve relevant passages from the collection of the passages. In the proposed architecture various modules were introduce. Automatic Entity Generator module identified domain related entities from which user can ask questions. Question classification module has several categories of question. An answer extraction module extracts the answer. By using ranking, answer selection module selects the answer among the candidate answers.

(Sahu et al., 2012) discussed an approach for question answering system for the Hindi language. This work deals with four types of questions when, where, how many and what time. For given question, the answer was retrieved from Hindi text. Each sentence in the text was analyzed to understand its meaning. In this work, they represent the questions using query logic language(QLL) which is a subset of Prolog. For identification of the noun, verb and question word Hindi shallow parser was used.

(Stalin et al., 2012) implemented the web based Hindi question answer. In this work the question and answer deal with only Hindi language, if the answer was not presented in Hindi document then it was retrieved from Google.

(Sekine and Grishman, 2003) proposed a question answering system for Hindi and English. The questions were created in Hindi language and the answers retrieved from Hindi newspaper in the Hindi language. These answers were then converted into the English language. In this work, an English Hindi bilingual dictionary was used to find top 20 Hindi articles which were used to find candidate answers.

(Reddy and Bandyopadhyay, 2006) proposed question answering system in the Telugu language. The system was dialogue based and railway specific domain. The architecture was based on the keyword approach. The query analyzer generates the tokens and keywords. From tokens, SQL statements were generated. Using SQL query the answer was retrieved from the database.

(Reddy and Bandyopadhyay, 2006) develop the question answer system in English and Punjabi language. In this work a pattern and matching algorithm was introduced to retrieve the most relevant appropriate answer from multiple sets of answers for a given question.

3. Resource Creation

We create QA dataset MMQA (Multi-domain Multilingual Question Answering) in Hindi and English languages covering multiple domains. We focus on creating *factoid* and

[1]Whenever required, a in-house language identification module (Gupta et al., 2014) and translation are used.

Domains	English/Hindi			
	# Articles	# Paragraphs	# Sentences	# Words
Tourism	112/112	1,569/1,186	5,077/3,799	90,222/71,863
History	68/68	563/597	2,518/2,224	40,368/46,418
Diseases	31/31	441/298	1,932/1,171	33,787/28,128
Geography	16/16	81/171	304/520	8,915/9,443
Economics	13/13	146/144	667/477	10,633/10,875
Environment	10/10	64/54	290/272	5,319/6,109
Total (EN/HI)	250/250	2,864/2,450	10,788/8,463	189,244/172,836
Total (EN+HI)	500	5,314	19,251	362,080

Table 3: Statistics of comparable English and Hindi articles from various domains.

short descriptive questions. The MMQA dataset is created in three different stages: *Comparable Article Curation*, *Question Answer Formulation* and *Validation*.

3.1. Comparable Article Curation

Since our objective was to build a multilingual QA dataset, therefore we curated comparable articles from the various web sources covering different domains. Rather than translating an article from one language to the other, we curated comparable articles, mainly for two reasons: **(i)** to bridge the information gap between two different language articles, and **(ii)** to assess the challenges of dealing with two syntactically divergent texts to retrieve answers for the given questions. From each of these articles, we extracted individual paragraphs, and removed images, links, and tables. We curated a total of 500 articles from 6 different domains. We curated these texts from the different web sources using a web crawler[2]. We provide the statistics of the comparable articles in Table 3.

3.2. Question and Answer Formulation

We engage annotators[3] to formulate the question-answer pairs in their own word. The annotators are provided with a interface displaying domain name, article name and the comparable article in parallel. They were asked to formulate questions in English and Hindi by looking into both the comparable articles. When they formulate a question by looking a paragraph in one (English or Hindi) article, they also have to verify whether the question of interest is available in the comparable article or not. In particular, they have to provide the *question, answer, answer source (sentence or paragraph where the answer exists), type of question (factoid, descriptive)* in both the languages, if exist. Additionally, annotators were encouraged to set questions in their own words. Statistics[4] of question and answers in both the languages are shown in Table 4.

We ask two other annotators to verify the questions and answers generated in both the languages. Annotators were given a free hand to correct the answers to some extent, or by eliminating the question-answer pairs, if found not fitting.

[2]Tourism (EN):www.india.com/travel
Tourism (HI): https://hindi.nativeplanet.com
Diseases (EN, HI): https://simple.wikipedia.org/wiki/List_of_diseases,
rest of the domains are curated from http://www.jagranjosh.com/

[3]The annotators are equally proficient in both the languages

[4]Total of 7120 questions (English+Hindi) for which the answer exists in either of two language documents.

3.3. Validation

Validation stage is performed to ensure that we obtain a high quality datasets at the end. We ask two other annotators to verify the questions and answers generated in both the languages. Annotators were given a free hand to correct the answers to some extent, or by eliminating the question-answer pairs, if found not fitting. The validation stage is applicable for the question-answer pair of both the languages.

3.4. Analysis

We analyze the questions and answers of the proposed MMQA dataset. It is required to understand its property and usefulness as a multilingual dataset. Our analysis focuses on studying the difficulty level of questions and diversity of answers. We provide some examples in Table 2 to give some ideas about the difficulty levels associated. For better understanding and thorough analysis of various answer types, similar to Rajpurkar et al. (2016) and Trischler et al. (2016), we categorize the answers of factoid questions into 8 entities and phrases. Statistics of the answer types for English and Hindi QA pairs are provided in Table 7.

An example of QA pairs formulated from a comparable articles is given in Table 1. Some examples of short descriptive QA pair from our dataset are given in Table 5. The direct comparison of our dataset with the Cross-Language Evaluation Forum (CLEF) datasets (Pamela et al., 2010) is not possible because we have created question answers pair in both language (MQA) in contrast the CLEF dataset have the question and answer pair in the different languages. However, we have shown the comparison in various terms as shown in Table 6.

4. Evaluation: Proposed Approach

We develop a translation based approach for multilingual QA. As English is a resource-rich language, we translate Hindi question and articles into English. Our proposed model comprises of *Knowledge Source Preparation*, *Question Processing* and *Answer Extraction*, We describe the details of each component in the following.

4.1. Knowledge Source Preparation

In this step, an information source (articles) from which answers are to be derived was set-up. We translate Hindi questions and articles into English by Google Translate[5]. The complete English articles are indexed at passage level using inverted indexing mechanism. We use the Lucene[6] implementation of inverted indexing.

4.2. Question Processing:

The question processing (QP) step is responsible for analyzing and understanding the questions posed to the QA system. We perform question classification with the question classes proposed by Li and Roth (2002). Question class provides us the semantic constraint on the sought-after answer. We propose a deep learning based question classification

[5]https://translate.google.com

[6]https://lucene.apache.org/

Domains	QA pair in only English (Fact/Desc)	QA pair in only Hindi (Fact/Desc)	QA pair in Both Languages (Fact/Desc)	Total QA pair (Fact/Desc)	Total QA pair
Tourism	456/14	403/5	422/10	1,281/29	1,310
History	110/75	126/78	1,118/588	1,354/741	2,095
Diseases	81/54	33/26	48/40	162/120	282
Geography	55/7	29/10	174/202	258/219	477
Economics	25/4	14/5	682/218	721/227	948
Environment	9/3	2/1	226/142	237/146	383
Total	736/157	607/125	2,670/1,200	4,013/1,482	5,495

Table 4: Statistics of QA pairs for factoid and short descriptive questions in English and Hindi.

Question (English): Why did Alexander marched back in 325 BC?
Question (Hindi): अलेक्जेंडर 325 ईसा पूर्व में क्यों चला गया?
Answer (English): After Alexander's last major victory in India as his forces refused to go any further. They were too tired to carry on with the Alexander's expedition and wanted to return home. Moreover, the might of Magadhan Empire (the Nanda Rulers) also dissuaded them. Alexander marched back in 325 BC after making necessary administrative arrangement for the conquered territories. He died at the age of 33 years when he was in Babylon.
Answer (Hindi): हालांकि, यह जीत भारत में उसकी आखिरी बड़ी जीत साबित हुई क्योंकि उसकी सेना ने इसके बाद आगे जाने से इनकार कर दिया था। वे सिकंदर के अभियान के साथ जाने से काफी थक गए थे और वापस घर लौटना चाहते थे। इसके अलावा, मगधियन साम्राज्य (नंदा शासक) की ताकत से भी वो भयभीत थे। विजय प्राप्त प्रदेशों के लिए आवश्यक प्रशासनिक व्यवस्था करने के बाद सिकंदर 325 ईसा पूर्व वापस चले गया।

Question (English): What does Buddhist texts such as Jatakas reveal?
Question (Hindi): बौद्ध ग्रंथों जैसे जतकस क्या बताते हैं?
Answer (English): ABuddhist texts such as Jatakas reveal socio-economic conditions of Mauryan period while Buddhist chronicles Mahavamsa and Dipavamsa throws light on the role of Ashoka in spreading Buddhism to Sri Lanka.
Answer (Hindi): Not Available

Table 5: Examples of short descriptive QA pairs from the dataset.

model to classify the question at coarser and finer level. The proposed question classification model is described as follows:

4.2.1. Question Classification

This is a vital component of any scoring based answer extraction technique. Its performance is the major concern as the errors in this component can propagate through the next stage and can affect the subsequent stages. In general question classification categorizes a question at coarser and finer level based on the answer type. For example, when considering the question Q: *When did Mandi become a part of India?*, we wish to classify this question as coarse class: *Numeric* and finer class: *date*, implying that only candidate answers that are dates need to be considered. With the recent developments in deep learning, neural network models have shown promise for QA. Deep neural network being perform exceptionally well in other NLP problem. Inspired by the success of deep neural network we adapt neural network architecture to develop our question classification model. Our question classification model is based on CNN and RNN. The model comprises of *Question embedding layer, Convolution layer, Recurrent layer, Softmax classification layer*. Our question classification model is inspired from (Kim, 2014) and (Xiao and Cho, 2016). The input to the model is an English question. Now we describe each component of the model:

- **Question embedding layer:** It is responsible for obtaining the sequence of dense, real-valued vectors, $E = [v_1, v_2 \dots v_T]$ of a given question having T tokens. We keep the maximum size of token $T = 15$ in this layer. The distributed representation $v_i \in R^k$ is the k-dimensional word vector. The distributed representation v is looked up into the word embedding matrix W. In our experiment we have used the pre-trained word embedding [7] matrix by (Mikolov et al., 2013).
- **Convolution layer:** This layer performs convolution operation. Similar to (Xiao and Cho, 2016) and (Kim, 2014) we obtain convolution feature c_t at given time t. Then we generate the feature vectors $C = [c_1, c_2 \dots c_T]$. The convolution operations are performed with the filter size of 3, 4 and 5.
- **Recurrent layer:** This layer performs recurrent operations over the convolution output c at given time t. Similar to (Xiao and Cho, 2016) we obtained the forward and backward hidden states at every step time t using the gated recurrent unit (GRU) (Cho et al., 2014). Xiao et al. (2016) have used LSTM unit, however we have employed GRU (Cho et al., 2014) due to its less complex architecture compare to long short term memory (LSTM).

$$\mathbf{z}_i = \sigma(\mathbf{W}_z c_i + \mathbf{V}_z \mathbf{h}_{i-1} + \mathbf{b}_z)$$
$$\mathbf{r}_i = \sigma(\mathbf{W}_r c_i + \mathbf{V}_r \mathbf{h}_{i-1} + \mathbf{b}_r)$$
$$\mathbf{c}_i = tanh(\mathbf{W} c_i + \mathbf{V}(\mathbf{r}_i \odot \mathbf{h}_{i-1}) + \mathbf{b})$$
$$\mathbf{h_i} = z_i \odot \mathbf{h}_{i-1} + (1 - \mathbf{z}_i) \odot \mathbf{c}_i$$

[7] https://code.google.com/archive/p/word2vec/

	2003	2004	2005	2006	2007	2008	2009	**Our data**
Target lang.	3	7	8	9	10	11	9	2
Collection	News 1994		+ News 1995		+ Wikipedia Nov. 2006		JRC-Acquis	Web
No. of questions	200						500	**7120**
Type of questions	200 Factoid		+ Temp. restrict +Defn	-Type of question +List	+ Linked question +Closed lists		-Linked +Reason +Purpose +Procedure	Factoid Descriptive
Supporting info.	Document			Snippet			Paragraph	Document

Table 6: Comparison of our dataset with the various released Cross-Language Evaluation Forum (CLEF) dataset

where $\mathbf{z}_i$, $\mathbf{r}_i$ and $\mathbf{c}_i$ are update gate, reset get and new memory content, respectively. c_i is the convolution output at time t.The final output of recurrent layer h is obtained as the concatenation of the last hidden state of forward and backward hidden states.

- **Softmax classification layer:** Finally, the fixed-dimensional vector h is fed into the softmax classification layer to compute the predictive probabilities for all the question classes (coarse or fine).

4.2.2. Query Formulation

In order to form the query, we remove all the stop word, punctuation symbol from the question. We tag the question with Stanford PoS tagger (Toutanova et al., 2003). Then we concatenate all the noun, verb and adjective in the same order in which it appears in the question.

4.3. Passage Retrieval

The candidate passage that contains the answer(s) to the given question(s) are extracted in this stage. We exploit the Lucene's text retrieval functionality to retrieve passage. It retrieves and ranks the passages using a combination of a Boolean model and the BM25 vector space model (Zaragoza et al., 2004). The query obtained from the *question processing* stage, serve as an input to the scorer module. The most relevant 30 passages were retrieved for subsequent processing.

4.4. Candidate Answer Extraction

This depends on the output of question classification. For factoid question, the coarse class and finer class guide this stage to extract the appropriate entities from the candidate passage(s). We tag the candidate passage with Stanford named entity tagger (Finkel et al., 2005). We utilize the coarse class and finer class of a question to extract the suitable candidate answers. For a descriptive question, candidate answers are extracted by segmenting the relevant passage.

4.5. Answer Scoring and Ranking

Each candidate answer is assigned a score using the candidate answer extraction phase. We segment the candidate passage into several candidate answer sentences. Thereafter, we calculate the score for each of the candidate answer sentences.

1. **Term coverage (TS):** It calculates the number of query terms appearing in the candidate answer sentence. This is normalized w.r.t the number of terms present in the given query.

2. **Proximity score (PS):** It calculates the length of the shortest span that covers the query contained in the candidate answer sentence. This is again normalized in the same way.

3. **N-Gram coverage score (NS):** We compute the n-gram coverage till $n = 4$. Finally, the n-gram score between a query (q) and a candidate answer sentence (S) is calculated based on the following formula.

$$NGCoverage(q, S, n) = \frac{\sum_{ng_n \in S} Count_{common}(ng_n)}{\sum_{ng_n \in q} Count_{query}(ng_n)} \quad (1)$$

$$NGScore(q, S) = \sum_{i=1}^{n} \frac{NGCoverage(q, S, i)}{\sum_{i=1}^{n} i} \quad (2)$$

4. **Semantic Similarity Score (SS) :** Query and candidate answer are represented using the semantic vectors. Cosine similarity is then computed between the query and candidate answers.

$$\texttt{VEC}(X) = \frac{\sum_{t_i \in X} \texttt{VEC}(t_i) \times \text{tf-idf}_{t_i}}{\textit{number of look-ups}} \quad (3)$$

where X is query q or candidate answer sentence S, $\texttt{VEC}(t_i)$ is the word vector of word t_i. *number of look-ups* represents the number of words in the question for which pre-trained word embeddings[8] are available.

5. **Pattern matching score (MS):** This score is used in the descriptive question only. We design a set of patterns similar to the (Joho, 1999) to match a query against the candidate answers. We setup a score for each pattern according to their importance.
For factoid and descriptive questions the weighted aggregate score for each candidate answer (A) is calculated as:

$$\begin{aligned} S_f(Q, A) &= W_1^f * TC + W_2^f * PS + W_3^f * NS + W_4^f * SS \\ S_d(Q, A) &= W_1^d * TC + W_2^d * PS + W_3^d * NS \\ &\quad = + W_4^d * SS + W_5^d * MS \end{aligned} \quad (4)$$

[8] https://code.google.com/archive/p/word2vec/

Answer type	Proportion (English/Hindi)	Examples (English/Hindi)
Person	12.22 / 14.28	Krishna /तुलसीदास (*Tulasidas*)
Location	17.26 / 14.89	Madurai /भुवनेश्वर (*bhubaneswar*)
Organization	7.69 / 8.96	International Monetary Fund/ टाटा मूलभूत अनुसंधान संस्थान (*Tata Mulbhut Anusandhan Sansthan*)
Noun Phrase	23.79 / 24.57	Hotel Apsara / स्टील प्लांट (*Steel plant*)
Verb Phrase	2.58 / 1.57	planned economic development/ तनाव से छुटकारा (*Tanav se chhutkara*)
Adjective Pharse	1.98 / 1.02	Smiling Buddha/ 14 प्रमुख भारतीय बैंक (*14 pramukh Bhartiya bank*)
Date / Numbers	32.43 / 33.17	580 / 7 किलो (*7 kilo*)
Other	2.05 / 1.54	at least two / सापूतरा का मतलब है 'नागों का वास'। (*Saputra ka matlab hai 'Nagon ka vas*)

Table 7: Set of various answer type categories (only for factoid questions) from the dataset with their proportion (in %) for English and Hindi answer.

Here, W_k^f and W_k^d are the learning weights for factoid and descriptive question, respectively. Optimal values [9] [10] are determined through the validation data. For factoid question, the candidate having the maximum score is returned as an answer to the given question. Answers to the descriptive questions may sometimes cover multiple sentences. At first, we consider the sentence having the maximum score, and then include the other sentences which have scores closer to the highest one.

5. Experiments, Result and Analysis

The experiments performed on the benchmark English-Hindi dataset can be categorized in two-fold: English question classification and answer extraction.

5.1. Question Classification

We perform the experiment on coarse and fine class set of the questions using the model discussed in Section 4.2.. For training, we use three datasets

1. A dataset[11] of $5,452$ questions collected from Hovy et al. (2001), TREC 8 and TREC 9 questions dataset,

2. A dataset of 500 questions from TREC 10[12].

3. We also manually label $1,022$ questions at coarse and finer labels with the taxonomy guidelines provided by Li and Roth (2002). These questions were randomly taken from the set of curated questions.

We perform 5-fold cross-validation to evaluate the question classification model. We obtain the accuracies of 90.12% and 80.30% for question classification under coarse (i.e. 6 classes) and fine classes (i.e. 63 classes), respectively. This model is used to classify the incoming questions while we perform answer extraction.

[9] optimal weights for factoid (0.31, 0.18, 0.39, 0.12)
[10] optimal weight for descriptive (0.21, 0.09, 0.23, 0.19,0.28)
[11] http://cogcomp.org/Data/QA/QC/
[12] http://cogcomp.org/Data/QA/QC/TREC_10.label

Network Training and Hyper-parameters

We have applied the rectified linear units (ReLu) (Nair and Hinton, 2010) as the activation function in our experiment. We use the development data to fine-tune the hyper-parameters. In order to train the network, the stochastic gradient descent (SGD) over mini-batch is used and Backpropagation algorithm (Hecht-Nielsen, 1992) is used to compute the gradients in each learning iteration. In order to prevent the model from over-fitting, we employed a dropout regularization (set to 50%) proposed by (Srivastava et al., 2014) on the penultimate layer of the network. We have used cross-entropy loss as the loss function.

5.2. Answer Extraction

We perform experiments for the factoid and descriptive questions using the model proposed in Section 4.. We use 10% of the total dataset of factoid and descriptive QA pairs, shown in Table 4, as the validation dataset to fine-tune the weight parameters. Mean reciprocal rank (MRR) and exact match (EM) (Trischler et al., 2016) are used to evaluate the model performance on factoid question. For descriptive questions, we use the well-known machine translation evaluation metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).While evaluating we translate Hindi answer to English and create a gold answer set by combining the actual English answer and the translated English answer for each question. Performance of the system is reported in Table 8.

For factoid questions, we obtain the maximum MRR value of 65.72 for the domain *Environment*. We obtain the lowest EM and MRR values for the domain *Diseases*. One possible reason could be that most of the factoid answers are the phrases and the PoS tagger could not extract these correctly. The system achieves the maximum BLEU of 48.51 and ROUGE-L of 45.72 scores for the domains *Diseases* and *History*, respectively. Our model could not perform well for the descriptive questions of the domain *Tourism*. However, it is to be noted that *Tourism* contains only a few (29) short descriptive questions. Our close analysis reveals that the system suffers due to the errors encountered in the linguistic components such as PoS tagger and named entity (NE) tagger. The NE tagger could not detect some of the

Domains	Factoid		Descriptive	
	EM	MRR	BLEU	ROUGE-L
Environment	39.13	65.72	45.81	42.56
History	29.53	57.19	42.84	45.72
Geography	35.55	52.27	43.02	44.61
Diseases	23.29	34.78	48.51	39.19
Economics	26.28	46.89	45.12	44.77
Tourism	27.68	37.79	22.96	24.29
Total	30.24	49.10	41.37	40.19

Table 8: Performance (in %) of the proposed model for factoid and descriptive questions

entities present in the translated Hindi passage, may be due to the errors encountered during translation.

6. Conclusion

In this paper, we propose a new multilingual QA dataset: MMQA. The dataset has wide coverage of various entities as the answer. It can be used to build a monolingual (EN: English, HI: Hindi), cross-lingual (EN $\rightarrow$ HI, HI $\rightarrow$ EN) and multilingual (EN $\leftrightarrow$ HI) QA system. We have collected $5,495$ QA pairs from 500 articles covering various domains. Our analysis yields divers answer types and a significant proportion of questions that require some reasoning ability to solve. We expect that MMQA will facilitate research in multilingual QA, involving Indian languages. We have also built a deep CNN-RNN based model for question classification. Our scoring based answer extraction module will serve as a useful baseline for further research. In future, we would like to extend the dataset by adding more QA pairs from various languages and different types of questions such as list and complex questions. We would also like to propose an end-to-end model for multilingual QA in the near future.

Acknowledgements

Asif Ekbal greatfully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

7. Bibliographical References

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Gupta, D. K., Kumar, S., and Ekbal, A. (2014). Machine learning approach for language identification & transliteration. In *Proceedings of the Forum for Information Retrieval Evaluation*, pages 60–64. ACM.

Hecht-Nielsen, R. (1992). Neural networks for perception (vol. 2). chapter Theory of the Backpropagation Neural Network, pages 65–93. Harcourt Brace & Co., Orlando, FL, USA.

Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-Y., and Ravichandran, D. (2001). Toward semantics-based answer pinpointing. In *Proceedings of the first international conference on Human language technology research*, pages 1–7. Association for Computational Linguistics.

Joho, H. (1999). *Automatic detection of descriptive phrases for Question Answering System: A simple pattern matching approach*. Ph.D. thesis, University of Sheffield, Department of Information Studies.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Knott, A., Bayard, I., de Jager, S., Smith, L., Moorfield, J., and O'Keefe, R. (2001). A question-answering system for english and ma_ri. In *Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES), University of Otago*, pages 223–228.

Kumar, P., Kashyap, S., Mittal, A., and Gupta, S. (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.

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics, COLING 2002*, pages 1–7. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Pamela, F., Danilo, G., Bernardo, M., Anselmo, P., Rodrigo, Á., and Sutcliffe, R. (2010). Evaluating multilingual question answering systems at clef. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002).

Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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. Association for Computational Linguistics.

Reddy, R. R. N. and Bandyopadhyay, S. (2006). Dialogue based question answering system in telugu. In *Proceedings of the Workshop on Multilingual Question Answering*, pages 53–60. Association for Computational Linguistics.

Sahu, S., Vasnik, N., and Roy, D. (2012). Prashnottar: A hindi question answering system. *International Journal of Computer Science & Information Technology*, 4(2):149.

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

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958.

Stalin, S., Pandey, R., and Barskar, R. (2012). Web based application for hindi question answering system. *International Journal of Electronics and Computer Science Engineering*, 2(1):72–78.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2016). Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Xiao, Y. and Cho, K. (2016). Efficient character-level document classification by combining convolution and recurrent layers. *CoRR*, abs/1602.00367.

Zaragoza, H., Craswell, N., Taylor, M. J., Saria, S., and Robertson, S. E. (2004). Microsoft cambridge at trec 13: Web and hard tracks. In *TREC*, volume 4, pages 1–1.