Learning to Classify Marathi Questions and Identify Answer Type Using Machine Learning Technique

Sneha Kamble and S. Baskar

Abstract One of the budding fields of artificial intelligence is Question Answering (QA). QA is a type of information retrieval in which a set of documents is given, and a QA system attempts to search for the correct answer to the question posed in natural language. Question classification (QC), which is a part of QA system, helps to categorize each question. In QC, the entity type of the answering sentence for a given question in natural language is predicted. QC is a very crucial step in QA system as it helps to take the important decision. For example, QC helps to reduce the possible options of the answer, and thus the answers that match the question class are to be considered. This research takes the first step toward the development of QC system for English–Marathi QA system. This system analyzes the user's question and deduces the expected Answer Type (AType), for which a dataset of 1000 questions from Kaun Banega Crorepati (KBC) was scrapped and manually translated into Marathi. Right now, the result for translation approach for the coarse-grained class is 73.5% and the fine-grained class is 47.5%, and for the direct approach, it is 56.5 and 30.5% for coarse and fine, respectively. Experiments are going on to improve the results.

1 Introduction

Question Answering (QA) is an application of Natural Language Processing. A system which provides people with a convenient and natural interface for accessing information is Question Answering system. Nowadays, the need to develop accurate systems gains more importance due to available structured knowledge bases and

S. Kamble (✉) · S. Baskar
Goa University, Taleigao, Goa, India
e-mail: sneha311093kamble@gmail.com

S. Baskar
e-mail: baskar@unigoa.ac.in

D. Reddy Edla et al. (eds.), *Advances in Machine Learning and Data Science*,
Advances in Intelligent Systems and Computing 705,
https://doi.org/10.1007/978-981-10-8569-7_4

the continuous demand to access information rapidly and efficiently. QA system has applications in various domains like education, health care, and personal assistance. QA system retrieves the precise information from large documents according to the user question [1].

India is a multilingual society with 30 languages which are spoken by more than a million native speakers, and these languages are written in different scripts. However, users who do not use English as the first language are also significantly high in number. In India, most of the people speak in their regional language and hence find it difficult to express their queries in English and a huge amount of information on the Internet is available in English. Thus, this work might help to develop a search system in the Marathi language [2].

Question classification maps a question into a category that indicates which information should be present in the answer. QC is a very important step because the question category helps mainly in two cases. One is to reduce the number of possible answer options and other is to decide the strategy to search for answer depending on the question category.
Example:
Who is Mozart?
Is directly mapped into Human: Description
What is Australia's national flower?
The headword flower is identified and mapped into the category Entity: Plant.

The paper is organized as follows: Sect. 2 presents the background study; Sect. 3 discusses the data creation; Sect. 4 describes the QC system; Sect. 5 presents the results and experiments for various features used; Sect. 6 discusses error analysis; and Sect. 7 concludes the paper with the future work.

2 Background Research

2.1 "Learning Question Classifiers: The Role of Semantic Information"—Xin Li and Dan Roth

Li et al. [3] present a question classification system that is modeled with the machine learning approach with a multi-class classification task with 50 classes. A hierarchical classifier was used to classify questions in fine-grained classes. Results of [3] conclude that a learning approach is better as the problem can be solved with 90% of accuracy.

2.2 "From Symbolic to Sub-symbolic Information in Question Classification"—Joao Silva, Luisa Coheur, Ana Cristina Mendes, and Andreas Wichert

Silva et al. [4] present the evaluation of a question classifier which is rule based. Direct match to the question or identifying the question headword and then applying WordNet to map the question are the two scenarios that are followed [5]. Direct match uses a set of rules that are manually built for compilation. The second scenario also uses manually built rules, but these rules express a path in the parse tree for question headword. After headword extraction, the headword hypernyms are analyzed until it matches to the possible question category. Results obtained by [4] gave a precision of 99.9% for coarse-grained class and 98.1% for fine-grained class in case of a direct approach. Whereas a precision of 86.4% for coarse-grained class and 78.5% for fine-grained class was obtained for headword extraction approach with WordNet mapping.

2.3 "Answer ka type kya he?" Learning to Classify Questions in Code-Mixed Language—Khyathi C. Raghvavi, Manoj Chinnakotla, and Manish Shrivastava

Raghvavi et al. [6] present a QC system for English–Hindi code-mixed questions which is Support Vector Machine (SVM) based. The language resources like chunker, parser does not exist for code-mixed languages. [6] made use of word-level resources such as language identification, transliteration, and lexical translation to the features in a single common language and then trained the SVM model to predict the question classes. Translating into a resource-rich language such as English would be helpful since it has QC training data. [6] created a code-mixed question dataset for English–Hindi language pair from college students and evaluated the approach to it. [6] achieves a coarse-grained average accuracy of 63% and fine-grained accuracy of 45%

Table 1 summarizes the discussed papers.

3 Data Creation

Predicting the entity type of the answering sentence for a given question is performed in question classification. QC [7] is performed by classifying the question to a category from a set of predefined categories.

Table 1 Literature review

Paper	Author name and year	Experiments (Classifier-features)	Results
From symbolic to sub-symbolic information in question classification	Joo Silva Lusa Coheur Ana Cristina Mendes Andreas	SVM-Headword, categories, Unigram	Coarse: 95.00% Fine: 90.8%
Learning Question Classifiers: The Role of Semantic Information	Xin Li and Dan Roth (2009)	VM - POS, NER, Chunk Tags	Coarse: 92.5% Fine: 85.00%
Answer ka type kya he? Learning to Classify Questions in Code-Mixed Language	Khyathi C. Raghavi, Manoj Chinnakotla, Manish Shrivastava (2015)	SVM Unigram, Adjacent	Coarse: 63% Fine: 45.00%

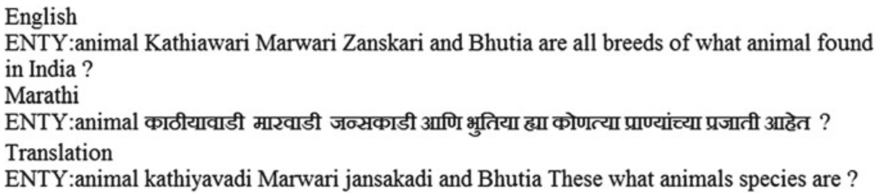

English
ENTY:animal Kathiawari Marwari Zanskari and Bhutia are all breeds of what animal found in India ?
Marathi
ENTY:animal काठीयावाडी मारवाडी जन्सकाडी आणि भुतिया ह्या कोणत्या प्राण्यांच्या प्रजाती आहेत ?
Translation
ENTY:animal kathiyavadi Marwari jansakadi and Bhutia These what animals species are ?

Fig. 1 Dataset example

For research and development for QC system, data was obtained. The dataset was created through KBC questions from[1] and comprises a total of 1000 questions. For each question, the dataset includes the questions and their question class as the class for the classifier. These questions were tagged into their respective classes using the hierarchical question ontology defined by [3] which is shown in Table 2. All question have been annotated with syntactic features. Further, the complexity of the dataset was analyzed. Since the data is of KBC, it had English–Hindi code-mixed words in it, which were taken care separately.

For Marathi questions, the same questions were translated in Marathi manually and under linguist guidance. So, total parallel dataset of 1000 questions in English and Marathi has been generated. The dataset created is expected to contribute to others for their research and also simulate other research in the field of question classification. Example of the dataset is given Fig. 1.

[1] http://www.smsoye.in/quiz.php.

Table 2 Taxonomy for question classification

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, Period, Percent, Speed, Temperature, Size and Weight

4 Question Classification System

This section describes the QC system in detail. For classification, the taxonomy of question classes proposed by [3] is used. This research mainly focuses on the classification of Marathi questions for which two approaches were used. One is translation and second is direct, i.e., using syntactic features on Marathi words as it is [8].

4.1 Translation Approach

The architecture of the system for translation approach is shown in Fig. 2. Given a question in the Marathi language, we initially perform word-to-word translation. The questions are translated into English. 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 Part Of Speech (POS), Named Entity Recognition (NER), and Chunk as the features for all the words in the question [9]. The classifier according to training data classifies the test questions.

4.2 Direct Approach

In this approach, the process starts from the second step of the architecture. Given a question in the Marathi language, we directly transform the Marathi question into

Table 3 Results

Dataset	SVC		SVM		Random forest	
	Coarse (%)	Fine (%)	Coarse (%)	Fine (%)	Coarse (%)	Fine (%)
English	72.5	47.5	73.5	41.5	69.0	42.5
Marathi	56.0	30.5	59.0	31.5	61.0	30.5
Translated	73.5	47.5	67.0	36.0	67.0	35.0

a feature vector, rather than translating it, and pass it through all the SVMs and output the SVM class which outputs the maximum score [10]. We use POS, NER, and, Chunk as the features for all the words in the question. Later, the classifier according to training data classifies the test questions.

To compare and study the results, the English questions were classified using the existing question classification system.

5 Results and Experiments

The overall results are presented in Table 3. It shows the results obtained for English and Marathi dataset for coarse and fine class.

Along with this, we would like to mention the challenges faced in translation. For translation, Goslate API, Google translator, and Shata Anuvaadak[2] were tried.

The results obtained by Goslate API and Google translator are almost same. But Goslate API is available only limited times a day. None of the above gave better results for an entire question at once; hence, word-level translation was experimented and better results were obtained.

From the example given in Fig. 3, we can see that word-level translation gives a better result. When the results were compared to the results of Shata Anuvaadak, it was observed that few words remained as it is which it could not translate. In addition, the coverage of the translation model was less.
On the other hand, Google translator translates most of the word; the words that were difficult to translate were transliterated.

6 Error Analysis

During the course of our experiments, a few observations stood out that highlight the errors made by the approach used to extract the features. We observed that when the features were passed on word level, the results were poor.

[2] http://www.cfilt.iitb.ac.in/~moses/shata_anuvaadak/resource.phptext.

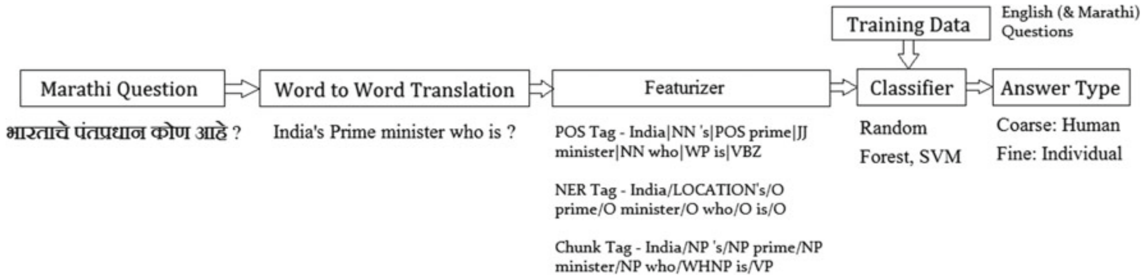
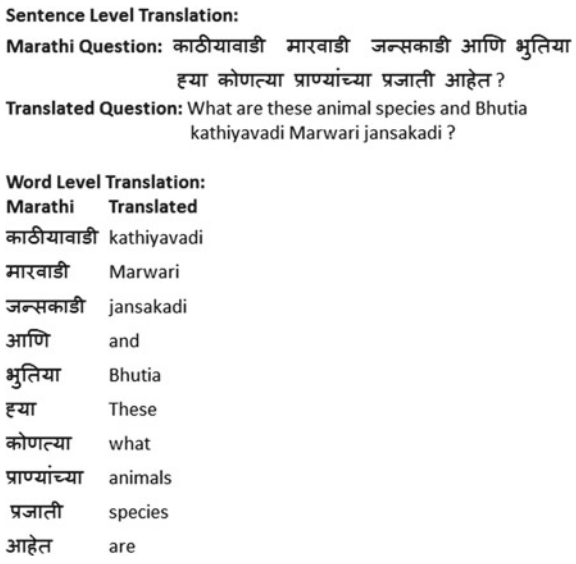

Training Data
English (& Marathi) Questions
Marathi Question
भारताचे पंतप्रधान कोण आहे ?
Word to Word Translation
India's Prime minister who is ?
Featurizer
POS Tag - India|NN 's|POS prime|JJ minister|NN who|WP is|VBZ
NER Tag - India/LOCATION's/O prime/O minister/O who/O is/O
Chunk Tag - India/NP 's/NP prime/NP minister/NP who/WHNP is/VP
Classifier
Random Forest, SVM
Answer Type
Coarse: Human
Fine: Individual

Fig. 2 Architecture

Fig. 3 Translation example



Sentence Level Translation:

Marathi Question: काठीयावाडी मारवाडी जन्सकाडी आणि भुतिया ह्या कोणत्या प्राण्यांच्या प्रजाती आहेत ?

Translated Question: What are these animal species and Bhutia kathiyavadi Marwari jansakadi ?

Word Level Translation:

Marathi	**Translated**
काठीयावाडी	kathiyavadi
मारवाडी	Marwari
जन्सकाडी	jansakadi
आणि	and
भुतिया	Bhutia
ह्या	These
कोणत्या	what
प्राण्यांच्या	animals
प्रजाती	species
आहेत	are

Example:
How far is Goa from Gujarat?
numeric : distance
How are you?
Entity: other

Here is an example in which the word how is categorized into two, and in the test example, the word might get wrongly classified. So, it is better to consider the whole sentence for better results.

7 Conclusion and Future Work

This paper presents our proposal of question classification for Marathi questions using translation and direct approach.

For this, we have used syntactic features like POS, NER, Chunk. The experimental results show that for a given set of questions in English and Marathi, the results are almost the same, whereas while using Marathi syntactic features directly gives a poor result.

Further, the features that are available and are useful to improve results can be experimented using better techniques.

References

1. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building watson: an overview of the DeepQA project. AI Mag. **31**(3), 59–79 (2010)
2. Zhang, D., Lee, S.W.: Question classification using support vector machines. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 26–32. ACM (2003)
3. Li, X., Roth, D.: Learning question classifiers. In: Proceedings of the 19th International Conference on Computational Linguistics, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
4. Silva, J., Coheur, L., Mendes, A.C., Wichert, A.: From symbolic to sub-symbolic information in question classification. Artifi. Intell. Rev. **35**(2), 137–154 (2011)
5. Zhiheng, H., Marcus, T., Zengchang Q.: Question classification using head words and their hypernyms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 927–936. Association for Computational Linguistics (2008)
6. Raghavi, K.C., Chinnakotla, M.K., Shrivastava, M.: Answer ka type kya he? Learning to classify questions in code-mixed language. In: Proceedings of the 24th International Conference on World Wide Web, pp. 853–858. ACM (2015)
7. Metzler, D., Croft,W.B.: Analysis of statistical question classification for fact-based questions. Informat. Retr. **8**(3), 481–504 (2005)
8. Moschitti, A., Chu-Carroll, J., Patwardhan, S., Fan, J., Riccardi, G.: Using syntactic and semantic structural kernels for classifying definition questions in jeopardy! In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 712–724. Association for Computational Linguistics (2011)
9. Vyas, Y., Gella, S., Sharma, J., Bali, K., Choudhury, M.: Monojit: Pos tagging of English-Hindi code-mixed social media content. EMNLP **14**, 974–979 (2014)
10. Jinzhong, X., Yanan, Z., Yuan, W.: A classification of questions using svm and semantic similarity analysis. In: 2012 Sixth International Conference on Internet Computing for Science and Engineering (ICICSE), pp. 31–34. IEEE (2012)