Question Classification in a code-mixed language

Archit Rathore - 12152
Prabuddha Chakraborty - 15111027

October 5, 2015

1 Introduction

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 [1]. Analysing code-mixed text is important for natural language processing tasks especially in bilingual or multilingual communities. And understanding the type of answer expected for a particular question is the initial step towards building an automatic question-answer system. Now for a code-mixed question, inferring the type of answer expected, will require require a solution for the above two problems.

2 Related Work

The problem of question classification for code mixed languages in Indian context was presented by Raghavi *et al*[1]. Li and Roth[2] have presented the widely used two-layered taxonomy for question type hierarchy that comprises of 6 coarse classes and each coarse class has 50 non overlapping fine classes which they classified using SVM with a modified tree kernel that preserves syntactic relations.

3 Dataset Creation

For this project we shall be generating our own dataset which shall contain "English and Hindi Code-mixed Questions-Answers". We shall be using the monolingual "English Question-Answer Set" created by Xin Li and Dan Roth [3] as basis for this dataset.

4 Proposed Method

Once we have the dataset we intend to perform the classification task first using the methods proposed in [2]. For the tree kernel method mentioned in [2] we would be using Siva Reddy's dependency parser [4] for generating the syntax tree. For generation of the feature vectors for the tree kernel method, we shall be using our own implementation. As this is still an open problem we plan to use all mainstream classifiers and compare their results. If it is feasible, we might also extend our work with distributional models for our code mixed data.

References

[1] Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. Answer ka type kya he?: Learning to classify questions in code-mixed language. *Proceedings of the 24th International Conference on World Wide Web*, pages 853–858, 2015.

[2] Xin Li and Dan Roth. Learning question classifiers. *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[3] Xin Li and Dan Roth. Experimental data for question classification. `http://cogcomp.cs.illinois.edu/Data/QA/QC/`.

[4] Siva Reddy. Hindi dependency parser. `https://bitbucket.org/sivareddyg/hindi-dependency-parser`, 2014.