Introduction

FAQs are the lists of common questions and answers on particular topics. Today one can and them in almost all web sites on the internet and they can be a great tool to give information to the users. Questions in FAQs are usually identified by the site administrators on the basis of the questions that are asked by their users. While such questions can respond to required information about a service, topic, or particular subject, they can not easily be distinguished from non-FAQ questions. Here we describes machine learning based parsing and question classification for FAQs and we showed that SVM and NB reach the accuracy of 80%. Identification of specific features is the key to obtaining an accurate FAQ classifier.

Contributions

- We hypothesize that FAQs have similar co-occurrence patterns through syntax and structure and could be correlated.
- This model is designed to be domain independent since it is not using any context dependent and lexical features.
- With detecting the candidate FAQs automatically by finding associations and similarity among questions, there is no need of continuous maintenance of FAQs.
- We introduce and evaluate a set of features in a setting that uses real and noisy data to classify Candidate FAQs.

Dataset

To the best of our knowledge, there is no publicly available FAQ dataset. Therefore, after choosing the suitable sources, we manually created our own dataset for the evaluation purposes. In our experiments, we collected FAQs and Non-FAQs about Sports, Foods, Computers (internet specifically) from the web, which contains approximately more than 3000 questions. In order to conduct more general FAQs and to be independent of the context, we try to choose a variety of domains with minimal overlapping; so that we can draw more general inferences. We manually labeled all questions as FAQ or Non-FAQ with the help of two annotators.

Feature set

Four sets of features have been used for the purpose of FAQ detection:
- Syntactic features
- Semantic
- Question words
- Bag of words

How to Classify

FAQ detection as a sub category of question classification has the aim of classifying question types into FAQ and Non-FAQ classes based on the sentence structure, keywords, syntactic, and semantic information.

The question classifier is responsible for taking the question as input and identifying which class it belongs to.

We are using SVM and Naïve Bayes classification approaches and both classifiers are trained using a set of predefined questions for each category.

Finally the classifiers are tested using a test-set containing unseen questions in each class.

Feature Selection

We have used following empirical feature selection methods:
- Information Gain (IG)
- ChiSquared Attribute Evaluation(Chi)
- CfsSubset (Cfs)

As a result, we chose top 5 features based on three feature selection methods in different contexts.

Evaluation

Our evaluation experiment investigates whether the performance of our classifier is comparable to the actual user-based rating questions. From this perspective we have implemented an experiment in which 50 English speaking graduate students and faculty members rated the questions generated from 4 different topics. We used Spearman’s rank correlation to compare the ranking score of our classifier and human user’s ranks obtained in identical number of questions:

\[ \rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \]

Therefore, we compared our classifier ranked questions with user rates and almost 81% similarity of the question ratings gives some confidence.