# Context Free Frequently Asked Questions (FAQs) Detection using Machine Learning Techniques

Fatemeh Razzaghi and Ali A. Ghorbani

## 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 classier.

## 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.

| Context | FAQ Question |
|---|---|
| Sport | What is volleyball?<br>Is soccer dangerous? |
| Food | Is maple syrup vegetarian?<br>How does food become hazardous? |
| Internet | What is the Internet?<br>How does e-mail work? |

| Context | Non-FAQ Question |
|---|---|
| Sport | Who is your favorite soccer player?<br>What is Michael Jordan doing now? |
| Food | How often do you go drinking?<br>Do you like food from other countries? |
| Internet | Which Internet browser is the best?<br>Have you tried Google Talk? |

## Feature set

Four sets of features have been used for the purpose of FAQ detection:

- Syntactic features
- Semantic
- Question words
- Bag of words

| Feature |
|---|
| The number of tokens in the question sentence. |
| The number of noun phrases in the question. |
| The number of prepositional phrases in the question. |
| The binary variable indicating the use of each wh-word. |
| The number of name entity tags of the question (Person, organization, Location). |
| The binary variable indicating the use of determiners in the question. |
| The binary variable indicating the use of any other question words (How many, How long, etc.) |
| The binary variable indicating if the question is yes-no question. |
| The list of Non-FAQ bag of words. |
| The binary variable indicating personal pronoun of the question. |
| The binary variable indicating the noun phrase as the last token of the question. |

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

We are using SVM and Naive 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)

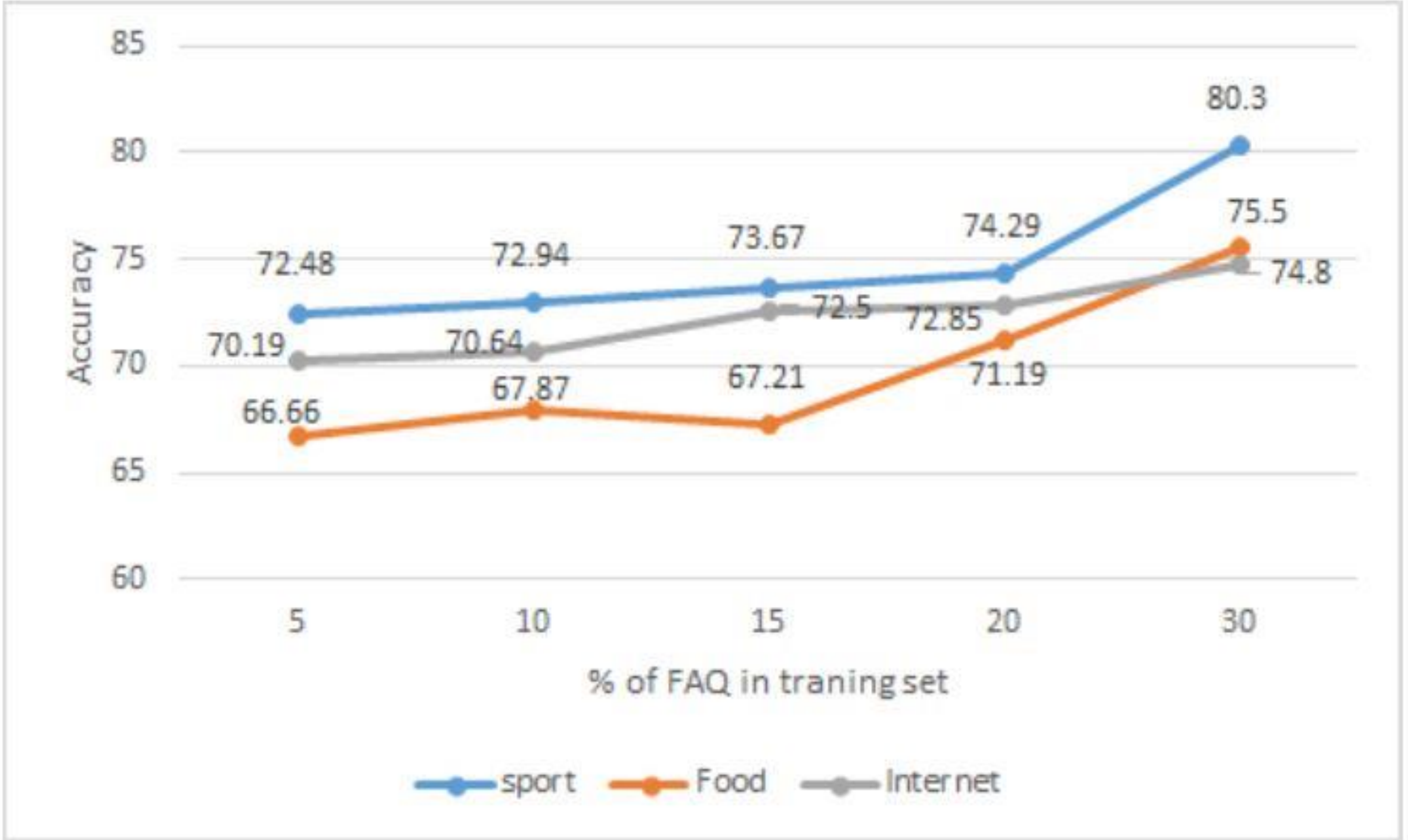
As a result, we chose top 5 features based on three feature selection methods in different contexts.

## Experiments and Results

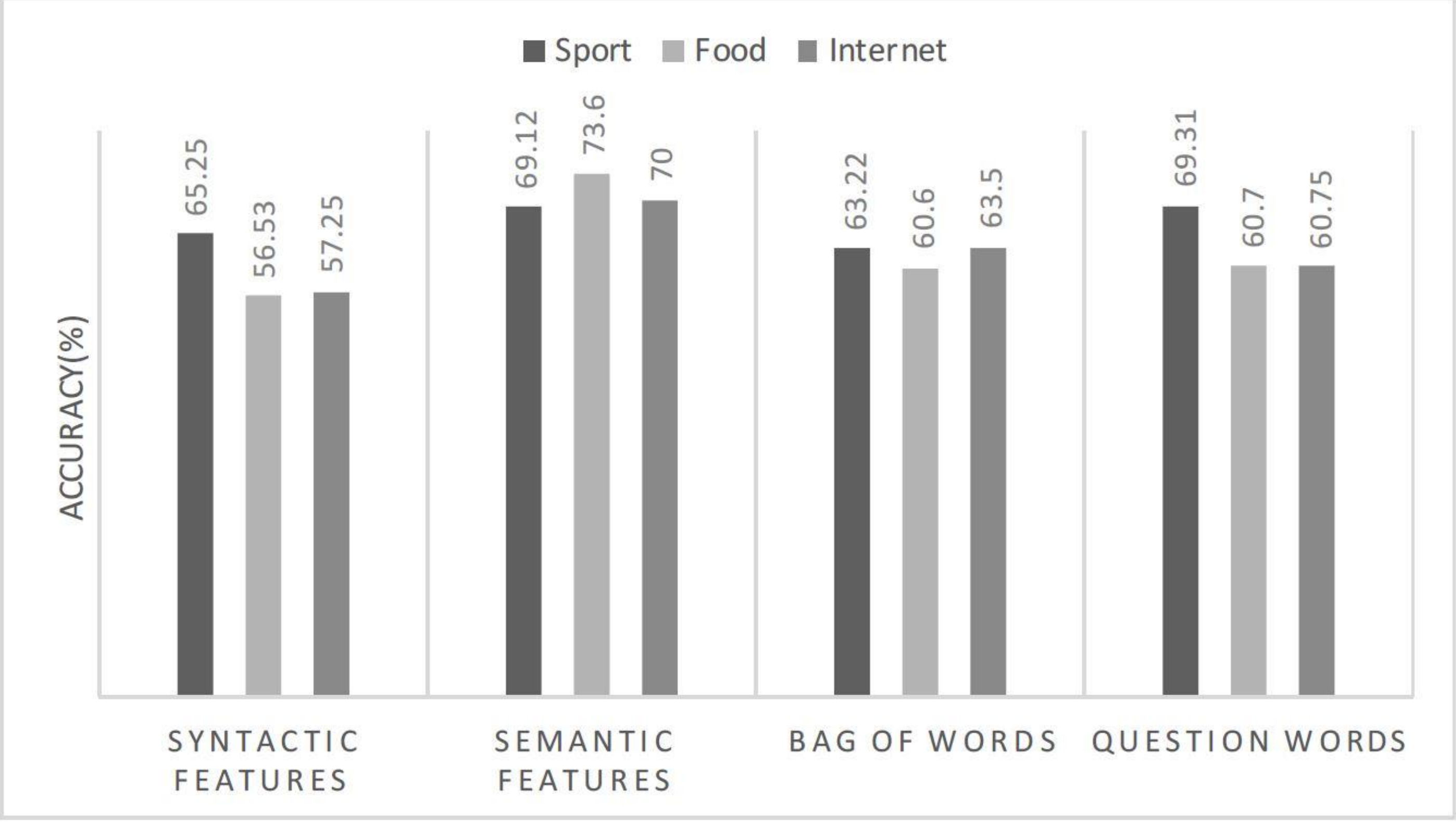
Validation results on three datasets (Sport, Food, and Internet) are given in the following table. We have used SVM and NB methods for FAQ classification. The overall best performance was achieved using the sport dataset.

| Dataset | NB | | | | | | SVM | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | PPV | TPR | Fscore | ROC | FP | Acc | PPV | TPR | Fscore | ROC | FP |
| Sport | 80.3 | 0.83 | 0.86 | 0.8 | 0.87 | 0.15 | 80.3 | 0.86 | 0.71 | 0.78 | 0.8 | 0.1 |
| Food | 75.5 | 0.78 | 0.69 | 0.73 | 0.84 | 0.18 | 77.2 | 0.86 | 0.63 | 0.72 | 0.77 | 0.09 |
| Internet | 74.8 | 0.75 | 0.74 | 0.74 | 0.84 | 0.23 | 74.5 | 0.76 | 0.73 | 0.74 | 0.75 | 0.23 |

Since we do not have access to the actual statistics of the total number of questions that form the bases of the real world FAQs, we consider different ratios of FAQs to Non-FAQs for training sets.



In order to analyze the impact of each group of feature on the classification task in the first run, the system was trained on each dataset (sport, food, and the Internet). However, we used each feature group separately to check their effect on increasing the accuracy.



## Evaluation

Our evaluation experiment investigates whether the performance of our classier 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 classier and human user's ranks obtained in identical number of questions:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

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

| Context | Spearman's rho |
|---|---|
| Food safety | 0.85 |
| Volleyball | 0.81 |
| Service Canada | 0.80 |
| Internet | 0.79 |

13th Annual research Exposition for the Faculty of Computer science April 8 2016, UNB Fredericton, Wu Conference Center