Question Classification and Context-Sensitive Response Generation Using Statistical Natural Language Processing

Idea:

The idea is to develop a system that gives meaningful and contextual answers to questions. We present a machine learning model which achieves this goal using various surface features present in a question. The algorithm would take as input a question from the user and would then try to extract meaning from the sentence, search the web for answers to the question, and finally display the answer in the correct context. The project is an alternative approach to web search, which uses page ranking for text mining.

The NPCEditor uses a statistical learning algorithm that can be used well for learning a generalized natural language model. It is similar to cross-language information retrieval task. Our algorithm will use KL-Divergence to compare the two probability models. Combining this approach with the algorithm used in [1], we can find the category under which the given question falls. Also, another model will be trained on the answer set, describing the type of answers. This can be used to better map questions to the apt responses. Different statistical models will be trained on these questions, answers and their types.

Data Set:

The data set contains question class definitions, the training and testing question sets, examples of preprocessing the questions, feature definition scripts and examples of semantically related word features.

Dataset Link:

Experimental Data for Question Classification
The link contains a repository of 5500 questions with the type labels assigned.

Software:

We plan to implement the algorithm in python/Matlab.
Libraries:
Python – NLTK, NumPy, SciPy
Matlab – MatlabNLP

Reading:

[1] Question Classification
[2] The N-best Algorithm: An Efficient and Exact Procedure for finding the n most likely sentence hypotheses
[3] NPCEditor: A Tool for Building Question-Answering Characters
[4] A Statistical Approach for Text Processing in Virtual Humans

Team:

Gaurav Narang (902896153)
Jayashree Chandrasekaran (902907974)
Rashmi Avancha (902927476)

Milestone:

By November 15 2012, we plan to have implemented the model and trained it with the data set that we have. We will then proceed to design the classifier and validate it with the available test data.