Project 1 – Evaluation Forum
Natural Language

Ana Ferro, 86375 Joana Alvoeiro, 89469

APPROACH = hybrid (coarse model: similarity measure, fine model: N-grams and similarity measure)

1 Introduction

The main goal of this project is to simulate an evaluation forum, a competition in which participants test their systems in specific tasks.

In this specific project, we were tasked with classifying questions according to Li and Roth's questions' taxonomy.

In order to train our model and get the best accuracy we opted to use two different models, one for the coarse and a different one for the fine taxonomy, both of them use tokenization and word manipulation.

2 Models' Description

To train our model, we first defined a baseline model with no accentuation and punctuation, with all words in lowercase, and using the Jaccard distance measure, which is a statistic used for gauging the similarity and diversity of sample sets. With our baseline model we got an accuracy of 70.977% for coarse and 16.509% for fine. From here, we further experimented with various pre-processing techniques and another distance measure to optimize our model.

It was in the pre-processing that we decided that using two different models would guarantee better results.

<table>
<tr><th>Coarse Model</th><th>Fine Model</th></tr>
<tr><td colspan="2">Tokenization & Standardization</td></tr>
<tr><td>Accuracy [at this point]: 70,977%</td><td>Accuracy [at this point]: 16.509%</td></tr>
<tr><td colspan="2">Stop-words removal
The way we chose our stop-words was through obtaining the 20 most common words in the training set, from that we evaluate how these words work together. if considered stop-words. Such common words could create noise in our models, so we decided to try removing them. Using this approach, we obtained the following sets of stop-words:

Set: [‘world’, ’go’, ’one’, ’four’] | Set: [‘first’, ‘one’, ‘four’]

In addition, we use the NLTK package of English stop-words. However, from this set of stop-words we kept the question words because they provide information as to what kind of question it is. Information that is important in the moment of the matching.</td></tr>
<tr><td>Accuracy [at this point]: 73.081%</td><td>Accuracy [at this point]: 19.769%</td></tr>
<tr><td>Snowball Stemmer
The idea behind stemming is that it reduces words to their base form, so that similar words will match each other.</td><td>Bigrams
Bigrams help provide the conditional probability of a token given the preceding token when we apply this conditional probability.</td></tr>
<tr><td>Accuracy [at this point]: 73.607%</td><td>Accuracy [at this point]: 64.458%</td></tr>
</table>

3 Other metrics & Error report

To get the highest accuracy possible using Jaccard distance we trained our model with other metrics such as: **Lemmatization**, **Porter Stemmer** and **RSLP Stemmer** but we did not get better results. We also combined **Cosine Similarity** with ***td-idf*** but the same happened.

The aspect we think we could most improve upon is stop word testing and selection. Initially, we added stop-words based on observation, without testing, which we paid for by having to manually check their effects later on. We also could have spent more time analyzing the training set and checking for superfluous words.

4 References

- Coursework materials ("Sebenta" and slides)
- NLTK documentation reference - https://www.nltk.or