APPROACH: MLClassification

Report MP1 Natural Language

Taíssa Ribeiro - 86514

Madalena Galrinho - 87546

1. A description of your model (for instance, the evaluation metrics you have tested or the preprocessing you have done)

We opted for a machine learning classification approach, using TF-IDF and Support Vector Machine classifier.

With TF-IDF we can determine how important a word is by weighing its frequency of occurrence in the document, in our case each question will be the document, and computing how often the same word occurs in other documents. Using this method, we can see how relevant a word is. For instance, if a certain word appears in almost all of the questions then it will be of low importance.

Relatively to the prediction of the taxonomy of these questions, we decided to use a Support Vector Machine, since it is known that has a lot of benefits for text classification [1]. This algorithm is capable of determining the best decision boundary between vectors that belong to a given group, based on the aforementioned vectors with TF-IDF.

The preprocessing was composed by elimination of stop words, where we used the English stop words list but we remove words such as "what", "where", "how", "when" and "why", since they are essential to predict the labels. Moreover, we used tokenization and lemmatization, to avoid data sparseness. Furthermore, we also applied lowercasing and removed punctuation.

2. Accuracy resulting from evaluating your model in the development set

Coarse = 81.28%

Fine = 72.77%

3. Short error analysis

The fact that there is ambiguity since words can have multiple meanings and different words can have the same meaning seems to be a problem. It would require recognizing the context in which the words appear, and some other techniques to eliminate the ambiguity.

Also, TF-IDF does not recognize semantics. It can only capture at lexical level.

Lastly, another problem that we do not take into account are the words that never appeared before, so if a question in the test set encounters a new word no data will be available for this specific word.

4. Bibliography

[1] – SUN A., Lim E. *On strategies for imbalanced text classification using SVM: A comparative study* (Institutional Knowledge at Singapore Management University, 2009)