11-791: Homework 4

Instructed by *Dr. Eric Nyberg*

Due on Oct 28, 2013

Troy, Zhenhao Hua zhua

1 Create Bag-of-Word Tokens

In `DocumentVectorAnnotator`, the main task is to generatel token list for each sentence. The basic method is to split words by space. After get the token, there are many prepossing methods we can use, such as making tokens case-insensitive, stemming and to remove stop-words. We will discuss the result of each method in the later section.

2 Retrieval Model

In `RetrievalEvaluator`, most of the Retrieval Model and evaluation logic lies here. Standard vector cosine similarity, dice coefficient and Jaccard coefficient are developed and evaluated. The results will be discussed later.

In this retrieval model, I also create a `QuerySet` class to store the indexes of the question and answer candidates of each query.

3 Result Analysis with different methods

3.1 Preprosessing Methods

The baseline method is case-sensitive, does not use stemming and does not ignore stopwords. Then we sequencially add the following methods and the final MRR scores are as following:

For the stemming, I use `Morphology.stem()` in `stanford-corenlp:3.2.0`.

From the results, we can see the best performence comes from the combination of case-insensitive and stemming. Considering the stopwords actually decreases the score. Considering the test set contains only 5 queries, these scores may not make perfect sense.

3.2 Different Retrieval Models

As we have the best performence when we consider case-insensitive tokens and stemming of the tokens, we keep this settings for different ranking algorithms: From the results, we can see that neither dice coefficient

Method	MRR score
Baseline	0.767
Case-insensitive	0.767
stemming	0.8
ignore stopwords	0.7

Method	MRR score
cosine similarity	0.8
dice coefficient	0.767
Jaccard coefficient	0.767

nor Jaccard coefficient improve the result compared with cosine similarity. Again, consider the small size of testing set, these numbers may not mean anything.

4 Error Analysis

The task of this homework is to measure the similarity in semantic meaning. So no matter how we change the prepossing and ranking methods, there is a gap between bag-of-word model and semantic meaning. For example, for the 4th query, "The shortest distance between new friends is a smile", its meaning cannot be fully understood unless we understand the analogy or the metaphor. "Wear a smile and have friends; wear a scowl and have wrinkles" is similar to the query in word level and even in the semantic level, but the relation between two metaphors has be understood if we want to fully understand this sentence.

5 Possible Methods beyond bag-word-model

We can use a semantic parser to get the semantic meaning of each sentence and compute the similarity based on the semantic meaning.