IIS — hw3

Chu–Cheng Lin `chuchenl`

October 2014

1 Error Analysis

Didn't capture the relationship among related words. The retrieval algorithm didn't take the similarity between 'bury' and 'destroy'.

Inadequate tokenization Query 2 does not have '`Jordan--one`' tokenized as 3 distinct tokens. This also affects Query 5. Query 7 and 11 could have suffered this problem, too.

Not recognizing the question type. Query 3 asks for the time Alaska was purchased; documents that contain temporal information should have higher scores. The same also goes for Query 4. Query 6 is affected by this problem, too. Also Query 8.

Not understanding the question. This type of error is probably hard to deal with, as it requires comprehension of the question to correctly answer it. For example, Query 9 has two potential answers 'Luna 2' and 'Eagle'. To answer this question, one must have the knowledge that a manned spacecraft falls in the spacecraft category.

2 Improvement

Based on the error analysis, I have done the following improvment to the system:

Better tokenization. Using the rule–based Penn Treebank tokenizer provided by the Stanford CoreNLP package, MRR increases from 0.4375 to 0.5125. This is an obvious effective improvement.

Lemmatization. Lemmatization may help the word relationship problem described before, as we will get the same lemma for several derived words. Combined with the Penn Treebank tokenizer, we get an MRR of 0.6375.