Spell Checker - NLP Assignment

August 20, 2013

1 Problem statement

All of you might have experienced the power of the Google Spell Checker. This is an opportunity to get a hands-on experience into building such a system. The assignment consists of 3 parts:

- Word spell check - standalone words are given and you are supposed to suggest corrections.
- Phrase spell check - words present in a phrases need to be checked for spelling
- Sentence spell check - an entire sentence needs to be checked for spelling

A practice dataset is provided along with the specification. The dataset is given in the form as follows

```
query1 <tab> suggestion1 <tab> suggestion2
query2 <tab> suggestion1 <tab> suggestion2
```

Think of the various issues that come into play when you make your spell checker. What distance would you choose for the word correction and to what level? Can you intelligently prune down the search space of candidate replacements? For the phrase and sentence spell check the context comes into play.

Remember that a good performance on the sample dataset is a necessary but not sufficient condition for your algorithm, since the results on the test might still surprise you.

Towards the end of the deadline for the assignment, the test sets will be provided to you.

1.1 Output

The output will be in the following specific format:

```
query1 <tab> suggestion1 <tab> score <tab> suggestion2 <tab> score
query2 <tab> suggestion1 <tab> score
```

Here score refers to the scores that you have obtained while ranking your suggestions.

2 Resources/Methodology

Some suggestions regarding the background knowledge that you can use:

- Free frequent n-grams data(to be downloaded offline) based on Corpus of Contemporary American English. You need to register your email to get the resources. The url is `http://www.ngrams.info/`.
- Wordnet as a dictionary. Wordnet is downloadable from `http://wordnet.princeton.edu/`
- The Brown Corpus (this is tagged with part-of-speech): `http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/brown.zip`
- Reuters dataset: `http://www.daviddlewis.com/resources/testcollections/reuters21578/reuters21578.tar.gz`

For any other resource(s)/help that may be required, you may approach the TAs.

3 Evaluation

The evaluation will be done using the Mean Reciprocal Rank measure. Please go through the url `http://en.wikipedia.org/wiki/Mean_reciprocal_rank` for more information. For example,

```
Problem:
The departments of the institute offer corses, conducted by highly qualified staff.
courses

Result
corses,courses,horses,corset - courses - rank 1, so Mean Reciprocal Rank is 1.
corses,horses,courses,corset - courses - rank 2, so Mean Reciprocal Rank is 1/2.
corses,horses,corset,courses - courses - rank 3, so Mean Reciprocal Rank is 1/3.
corses,horses,corset,scores - correct result not there, so Mean Reciprocal Rank is 0.
```

The MRR for all the test cases will be summed up to get a measure of the performance of the spell checker system you designed. Time taken by your code will also be taken into consideration.

Along with the code, you will also have to submit a report containing the details of your algorithm and observations you have made.

Intermediate deadlines for the assignment are as follows:-

- Aug 30 - Word Spell Check
- Sept 11 - Phrase Spell Check
- Sept 18 - Sentence Spell Check
- Sept 22 - Report and Code Submission

Considering the quiz week, the deadline has been set at **22nd Sept, 23:55 hrs and this is a hard deadline and non-negotiable.**