NLP Spell Checker Assignment

August 30, 2011

1 Assignment statement

Students are expected to submit a spellchecker code that would be run against test cases provided by us. A test case may be a word, phrase, sentence, or paragraph. For every test case, a maximum of 5 ranked suggestions can be made. Test cases would evaluate the submitted program at four levels of operation:

1. Word level - how good is your program at correcting standalone misspelled words?
2. Phrase level - how good is your program at correcting misspelled words in phrases?
3. Sentence level - the ability of your program to correct misspelled words in sentences.
4. Paragraph level - the ability of your program to correct misspelled words in paragraphs.

Test cases of the kinds (2), (3) and (4) would test the ability of your program to take into account contextual information. A set of sample test cases would be provided for you to work with, and after submission your programs will be evaluated on a seperate set of test cases. Test cases would be provided in the following general XML format:

```
<tests>
    <case id=1> test case </case>
    <case id=2> test case </case>
</tests>
```

There would be four test case files provided, one for each type. All test cases would adhere to the above format. In cases of type (1), a test case would be a single word. In cases of type (2), a test case is a string of space-separated words (representing a phrase). In cases of type (3), a test case is a string of space-separated words (representing a sentence). In test cases of type (4), a test case is collection of sentences, representing a paragraph, with each sentence of the paragraph on a separate line. Example:

```
<tests>
    <case id=12>
Two opinins about programming date from those days.
```

```
I mentin them now, I shall return to them later.
    </case>
</tests>
```

It may be noted that in test cases of types (2), (3) and (4) there may be multiple misspelled words per test case. The section on Output format details how these are to be represented in the output of your program.

1.1 Output

The output file is a comma-separated file with the following specific format:

```
Case Id_Number, misspelled word, correction1, correction2,
correction3, correction4, correction5
Case Id_Number, misspelled word, correction1, correction2
```

Note that a maximum of five corrected words is allowed. This is treated as a ranked list. i.e., give your best guess first, and then the second best guess, and so on. Also the submitted program must identify the misspelled word for itself (of course in test cases of type(1) this would be the only word in the test case. In cases of types (2), (3) and (4), the misspelled word(s) need to be identified.)

In cases of type (2), (3) and (4), each misspelled word per case and its suggested corrections must appear on a separate line. For example, given the paragraph test case input above, this is an output:

```
12, opinins, opinions, opines, openings
12, mentin, mention, minting, mentis, mentions
```

2 Resources/Methodology

Students may use any methodology they see fit, to solve the problem. They are expected to use the following resources:

- Wordnet as a dictionary. Wordnet is downloadable from `http://wordnet.princeton.edu/`
- For contextual training:
  - 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) e.g. corpora, that may be required, students may approach the TAs.

3 Sample test cases

Following are a few sample test cases of Types 1, 2 and 3:

Type 1:

hwere
where

taht
that

convnvcing
convincing

teh
the

ocrrect
correct

Type 2:

Securities and Echane Commission
Securities and Exchange Commission

international capitan markets
international capital markets

educational nsitutions
educational institutions

mewn and women
men and women

United tates
United States

Type 3:

The impending takeover bid increased the stock value of the bak.
The impending takeover bid increased the stock value of the bank.

The fugitive oldiers from the military were finally arrested today.
The fugitive soldiers from the military were finally arrested today.

The fisherman went fishing close to the river bak.
The fisherman went fishing close to the river bank.

All divisions of the ramed forces participated in the parade.
All divisions of the armed forces participated in the parade.

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

Note: During actual evaluation the input format would be as described in section 1 (Assignment Statement).

4 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,

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

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.