CS6370 Natural Language Processing
Spell Check Assignment

Tanmay Dhote CS11B051
S Vishnu Vardhan Reddy CS11B050

September 28, 2014

1 Introduction

This assignment involves making a Spell Checker. The spell checker has been divided into three parts:

- Word spell check - suggesting corrections for standalone erroneous words
- Phrase spell check - finding and suggesting corrections for misspelled words present in a phrase
- Sentence spell check - finding and correcting the erroneous word(s) that are present in a sentence

An important that has been made during the implementation is that the misspelled word is not present in the dictionary.

2 Word spell check

It is a two step process :

1. Candidate Generation - Creating a list of possible replacements
2. Scoring Candidates - Likelihoods of the candidates generated in the previous stage are estimated and the most likely candidates are chosen

2.1 Candidate Generation

The generation of candidates is done by using the edit distance. Edit distance is a way of quantifying how similar two strings are by finding out the minimum number of transformations needed to transform one string into another.

All words present in the dictionary within an edit distance of 3 are considered as possible candidates. Since the dictionary is huge, it has to be pruned to reduce the search space and improving efficiency. This was done by creating bins in which words we put words base on the size. 20 bins were made as with the last bin containing all words with length greater than 19. This value 20 was chosen since more than 99% words lie within 20 words. Only bins in the range of +- 3 of the length of the word are considered as only they can be the possible candidates.

Another method that was considered was to generate candidates ourselves based on 4 basic modifications - insertion, deletion, substitution and transposition. This was ultimately scrapped as the generation of candidates with edit distance 3 became too computationally intensive.

2.2 Scoring candidates

The scoring was done depending on the Maximum A Posteriori estimate using Bayes' theorem :

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

where $P(A|B)$ is known as the posterior, $P(B|A)$ is known as the likelihood, P(A) is known as the prior and P(B) is known as the evidence.

The priors were estimated to be the frequencies of the words seen previously. The likelihood was estimated using the edit distances with edit distance 1 being more likely than edit distance 2 which in turn is more likely than edit distance 3.The costs for the operations causing the transformation are based on the cost tables given by Kernighan, Church, and Gale in "A Spelling Correction Program Based on a Noisy Channel Model".

3 Phrase spell check

Since context was now available, it was made use of to have a context sensitive spell check. It is also a two step process similar to the word spell check

3.1 Candidate Generation

The generation of candidates was by using the candidates obtained by applying Word Spell check on the misspelled words.

3.2 Scoring Candidates

The scoring of candidates was done by using two methods

3.2.1 Context Words

All words in the context of the words (words spatially around the word in the phrase) were chosen and the most likely candidates were chosen depending on it. For phrase the whole phrase itself was considered as the context due to it's small size.

3.2.2 Collocations

Collocations not only use the context words but also the relative positions of the words with respect to the misspelled words. This was also implemented with the size of context as +-2.

Collocations were found to be much more computationally intensive than context words without giving a significantly higher performance. As a result, the context words was chosen to score phrases.

4 Sentence spell check

Sentence spell check was considered as a multiple phrase check problem with the only difference being that the size of the context was increased as a sentence may have long range dependencies. Also, since it may have multiple corrections, multiple parses of the sentences are done as a corrected word may then become a significant context for some other misspelled word.

5 Future Scope of Improvement

- The word spell check can be improved by using an indexing structure to further increase the pruning during candidate generation
- In context based spell check, PoS tags can be used