Report Of Assignment 1—Spell Correction

1 Task Description

Chances are high that people make typos during the typing process, and what's even worse, people generally find it difficult to catch their own typos. This task aims to spot the typos in a given sentence and correct them by implementing the Noisy Channel Model. Of the 1000 sentences from the test data, 94.5% are properly corrected.

2 Method

- Corpus: Gutenberg Corpus and Reuters Corpus from NLTK package
- Model: Noisy Channel Model
  $\hat{w} = \underset{w \in V}{argmax} P(w|x) = \underset{w \in V}{argmax} \frac{P(x|w)P(w)}{p(x)} = \underset{w \in V}{argmax} P(x|w)P(w)$
  - Language Model: Kneser–Ney smoothing with bigram probabilities.
  - Channel Model: the confusion matrix is generated from Peter Norvig's list of errors.

There are two kinds of errors in the given sentence, namely the non-word spelling errors and the real-word spelling errors. To deal with both of them, first the non-word errors are detected and replaced by the candidate with the highest probability given the misspelling. If there are still errors in the sentence, a set of candidates for each word will be generated, and each time one of the words in the sentence is replaced by its candidate. The sentence is then replaced with the one that maximizes P(W). For all the misspelling, candidates are generated within 1 Damerau Levenshtein edit distance.

3 Evaluation Result

The overall accuracy of the task is 94.5%, with running time of approximately 680 seconds. Even if only the non-word spelling errors are solved, the accuracy reaches up to 88.7%. After correcting the real-word spelling errors, the accurary increases to 94.5%. For the unsolved sentences, except for errors which should be corrected with candidates of edit distance greater than 1, most of the remaining misspellings are real-word errors. For instance, 'estimated' is mistakenly replaced by 'estimates'.

4 Future Works

To increase the accuracy, we could generate candidates with edit distance greater than 1, or try to implement n-grams Kneser-Ney smoothing to produce a better result when solving the real-word spelling errors.

* To run the program and evaluate the results, please run:

```
python hw1.py
python eval.py
```

Student ID: 16307090185 Name: Lin Yawen