Report_Team17

Kevin Alex Mathews[1] and Varalakshmi Nenavath[1]

Indian Institute of Technology Madras

1 Introduction

For word spell check, the noisy channel model of spell checking [1] has been employed.

For phrase spell and sentence spell check, our method makes use of the context words within a certain a distance of the confusing word.

Bayesian framework has been followed throughout the assignment for the calculation of probabilities.

2 Training Data

We have used the Brown corpus (without tags) [2] as the primary data set. We have tried other corpora such as the reuters dataset [3] and the big corpus [4]. In general, we got better results with the Brown corpus rather than with the others.

We use a word count file [5] consisting of 100, 000 words. In general, we achieved considerably good performance with this word list. However since it is a just list of common words, rather than a dictionary, it consists of many meaningless terms like 'll', 'b' and 'london'.

Stop words do not contribute much to remove the ambiguity of target words. Hence, we use a list of stop words [6] in order to filter out the stop words from the training data and the input text. This improves the performance of the spell checker in terms of execution time and space.

3 Word Spell Check

For word spell check. we use an inverted index to generate the candidates for an incorrect word. The index consists of character n-grams and corresponding words in the dictionary that contain the n-gram. We tried with both bigram and trigram inverted indicies. Although trigram inverted index performed better by

[1] http://www.cs.ubc.ca/ carenini/TEACHING/CPSC503-04/spelling90.pdf
[2] http://nltk.googlecode. com/svn/trunk/nltk_data/packages/corpora/brown.zip
[3] http://www.daviddlewis.com/resources/testcollections/reuters21578/reuters21578.tar.gz
[4] http://norvig.com/big.txt
[5] http://norvig.com/ngrams/count_1w100k.txt
[6] http://xpo6.com/list-of-english-stop-words/

generating a lesser number of candidates, it missed out on words without any matching trigram which are in common with the intended word. For example, if 'over' is incorrectly typed as 'ovre', trigram inverted index would not add the intended word 'over' to the candidate list because they do not share any trigram. In this case, a bigram inverted index could solve the problem. But, in general, bigram inverted index has low performance in terms of execution time because it generates a larger number of candidates.

The candidate list is pruned by removing all words which are at an edit distance greater than 3 from the incorrect word. This is based on the intuition that 99% of all spelling errors are at an edit distance of at most 2. To be on the safe side, we chose to keep all candidate words at an edit distance at most 3.

The prior probability is calculated from the training data. A smoothing of 0.5 is done to accommodate for words which are not present in the training data.

The probability of likelihood is calculated from the confusion matrices generated by Kernighan et.al.. All the four confusion matrices - addition matrix, deletion matrix, substitution matrix and reversal matrix - are employed for the purpose. A smoothing of 0.5 is done to accommodate for characters tuples with value 0 in the confusion matrices.

The words in the candidates list with the highest posterior probabilities are suggested as corrections for the wrong entry.

4 Phrase Spell Check and Sentence Spell Check

For phrase spell check and sentence spell check, we maintain a list of most confusing words. For example, the words 'advise' and 'advice' occur in the same confusion set in the list since they are often confused with each other.

We also maintain a list of confused words (pivot word) with corresponding context words. The list is generated from the training data offline. The half-width of the context window is set as 6. For example, typically the pivot word 'desert' would have context words such as 'arid', 'camel', 'sand' and 'oasis'.

Whenever a word in any confusion set is encountered in the input text, for each word in the corresponding confusion set, probabilities for the word to co-occur with the context words in the input text is calculated. We have employed Bayesian framework [7] for calculating the probability.

The words in the confusion set with highest posterior probabilities are suggested as corrections for the confused word.

We are also handling words such as 'handleoffame', 'graincell' which got combined incorrectly.

5 How to make the spell checker better?

More than one list of words can be used so that the probability of suggesting spurious terms as corrections can be minimized.

[7] http://www.aclweb.org/anthology/W95-0104

Word stemming can be done to reduce the size of the list of context words by getting rid of redundant terms. For example, if 'manufacturing' and 'manufacturer' occur in the context of the pivot word 'factory' in the training data, then word stemming can be employed so that only one word 'manufactur' would get stored instead of two words 'manufacturing' and 'manufacturer'.

The phonetic algorithm Metaphone was used to generate more candidate words. But since we faced difficulty in deciding how to incorporate it in the noisy channel model of spell check it was not included in the final version of the spell checker.

WordNet [8] can be used in phrase and sentence spell check for expanding the list of context words. For example, the word 'waterless' occurs in the context of the pivot word 'desert' in the training data. The context word 'waterless' can be looked up in the WordNet database to get similar words like 'arid' or 'dry'. These synonyms can be used to expand the list of context words of the pivot word 'desert'.

The performance of the pivot word-context word matrix is entirely dependent on the training data. Since it was not exhaustive, generating constructive context words for pivot words proved to be a difficult task. A solution to this problem is to make use of a huge data set like Wikipedia [9]. Since Wikipedia is an encyclopedia, there is a very high chance for constructive words to co-occur in the same article. Those words can be used as context words for the topic of the Wikipedia article. For example, the Wikipedia article on 'desert' [10] consists of constructive words such as 'arid', 'dry', 'camel', 'oasis' and 'sand-dune'.

[8] http://wordnet.princeton.edu/
[9] https://www.wikipedia.org/
[10] https://en.wikipedia.org/wiki/Desert