Spell Correction using LM[*]

Hamid Fazli Khojir
University of Windsor
fazlikh@uwindsor.ca

1 INTRODUCTION

Finding the correct spelling of a word can be done using minimum edit distance approach. For example, if we write 'Hmaid' instead of 'Hamid', minimum edit distance can suggest 'Hamid' as the correct spelling. However, this approach cannot predict the next word of the sentence. To address this problem, we can use language models. For example, if the user has typed 'This is a beautiful', then the language models can suggest words like 'city' and 'flower'. Additionally, language models can be used for spell correction as well. In this assignment, we are going to use N-gram language model for spell correction.

2 MOTIVATION

Writing is one of the main ways of communication. However, we may write a misspelled word because of lack of knowledge, not paying attention, and other reasons. The main purpose of our work is to reduce these mistakes by suggesting the correct spelling of a misspelled word.

3 PROBLEM DEFINITION

Given a text corpus C, train N-gram language models using $n = \{1, 2, 3, 5, 10\}$, and evaluate the model using s@k measure for $k = \{1, 5, 10\}$ on a spelling error corpus $\mathcal{E}$.

3.1 Example

For example, if error corpus has 'arrengiments arrangements to make *', it means that the correct sentence is 'to make arrangements', but the user has typed 'to make arrengiments'. Therefore, if n=2, then N-gram language model should suggest some words by receiving 'make' as the input. If n=3, then 'to make' will be passed to the model. Imagine that the model has suggested words like 'money', 'table', 'plan', 'arrangements', 'commands', respectively. Therefore, s@1, s@2, and s@3 are 0. s@4 and s@5 are 1 because the model could produce the correct spelling of the expected word at fourth suggestion.

4 EXPERIMENT

4.1 Datasets

The Brown Corpus is relatively old corpus and the first million-word electronic English corpus. It is consisted of 500 different sources and includes many different genres like news, reviews, and religion. It has 57340 sentences and 1161192 words. For this assignment, we have only considered the news genre of this corpus, which has 4623 sentences and 100554 words. This corpus is used as the training corpus for N-gram models.

APPLING1DAT.643 is a part of Birkbeck corpus which has almost 200 sentences which have one misspelled word. For example, 'companys companies * started to work' is one line of this corpus. This corpus is used as the test to evaluate trained models.

4.2 Results

n-gram	s@1	s@5	s@10
1	0	0	0
2	0	3	4
3	2	5	6
5	2	5	6
10	2	5	6

The above table shows total number of s@k for each n-gram model. Please consider that if you need to know the average, instead of total number, those numbers should be divided by 196, according to the number of sentences in test corpus. For example, average of s@5 for 2-gram model is 0.0153. Because the average number is small for all n-gram models, total number of success is used to be easier to read. This problem can be explained by two different reasons.

The first reason is that the test corpus has a lot of sentences like '* between', '* sensations', 'my * hurt'. The n-gram model that is used in this experiment is not bidirectional. Therefore, the words after the misspelled word will not help the model to suggest a good word. For example, a sentence like 'is * proved' is equivalent to 'is * than', and 'is *' for the model.

The second reason is that the training corpus and the test corpus are not highly related to each other. Therefore, the model cannot perform well on the test context.

As you can see, for 1-gram model, there is not success even after 10 predictions. When 1-gram is used, it means that the model does not receive any words of the sentence that it wants to correct. Therefore, the output of model is completely independent of the sentence.

By increasing the number of n, n-gram model receives more words of a sentence. Therefore, it can suggest better words according to the context. This is why we see increase in s@5 and s@10 from 1-gram to 2-gram model. The same reason can explain increase for all s@k from 2-gram to 3-gram.

s@k does not change from 3-gram to 5-gram and 10-gram. The reason is that the test corpus is consisted of small sentences with almost 3-4 words per sentence. Therefore, the potential of 5-gram and 10-gram model will not be used, and the history that they remember is wasted.

5 CONCLUSION

In this work, we successfully trained some n-gram language models for the spell correction task. Our results show that s@k may increase by increasing n in n-gram model, however, there are some limits and increasing n might not help the model to suggest better words. Therefore, n should be selected according to the training and test corpus, and there is not a general rule for selecting n.

[*] https://github.com/feknall/spell-correction-ngram