Real-word spelling correction : Experiments on assumptions made in the Mays, Damerau, and Mercer model

Aakash Nandi
McGill ID: 260741007
McGill University
aakash.nandi@mail.mcgill.ca

Abstract

The trigram-based noisy-channel model of real-word spelling-error correction that was presented by Mays, Damerau, and Mercer in 1991 took into consideration two critical assumptions on probability of a word likely to be not changed, to be constant and equal distribution of remaining probability mass among different words in confusion set. The author performs experiments by building models that question these assumptions.

1 Introduction to Real-word correction problem

Sometimes mistakes made by a typist cannot be determined by simply checking for the existence of a word present in sentence , in the vocabulary. While committing a mistake ,the typist may transform a word from what he/she meant to some other word that exists in the dictionary.For eg.He/She might replace "while" with "whale" as a typing mistake where "whale" might not fit into context.Such spelling mistakes require real-word correction.

2 The Mays, Damerau and Mercer Model

(Mays et al., 1991) (hereafter,MDM) proposed a statistical model based on Bayes theorem to tackle the real-word correction model.They framed this problem as a noisy channel problem in which an intended statement undergoes introduction of error by typist and then the system predicts the most likely statement that the typist had intended.

The probability that a word typed by typist is the word that he intended, is determined by a parameter that is denoted by α. In other words it encodes how confident is a typist. α in this model remains the same for all the words that the systems encounters while correcting.The remaining probability mass ie.(1-α) is equally distributed among all the words in the confusion set of the word being examined.The confusion set comprises of the all the words that are in the vocabulary and are within one edit distance from the word being examined.As stated in (O'Hearn et al., 2008) ,the probability that an intended word w is typed as x is

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ (1-\alpha)/|w_c| & \text{if } x! = w \end{cases} \quad (1)$$

where $|w_c|$ represents the size of confusion set comprising of all the words that are within one edit distance and exists in the vocabulary.This probability is used with a tri-gram based language model to compute probabilities of candidate statements and later the statement with the highest probability is considered to be the intended statement.

3 The Test-set and Evaluation

(Mays et al., 1991) did not test their implementation against a standard data-set and also did not report their performance in terms of precision,recall and F score.(O'Hearn et al., 2008) came up with three test-sets built on Wall Street Journal(WSJ) Corpus namely **MAL, T20, T62** to test (Mays et al., 1991)'s implementation.

Each test-set comprised of 300,000 words reserved from WSJ corpus of 30 million words, with 1

out of every 1000 words replaced with another word within one edit distance that exists in the vocabulary.The test-set differ only in the words being replaced.

T20: A replacement word exists in top 20,000 most frequent words.

T62: A replacement word exists in top 62,000 most frequent words.

MAL: A replacement word exists in lexicon of ispell® and the word being replaced is tagged as a noun in WordNet®.

Keeping in mind the time constraint and the computational resources, I test my models against a smaller corpus named Brown which has 1 million words.My test-set is a T20 subset of Brown corpus containing 10,000 words with an erroneous word inserted approximately after every 10 words to imitate a subject that has α = 0.9, as with these many errors inserted, the proposed system would be tried against enough true positive cases, in contrast to just 10 cases if, had I chosen to replace 1 in 1000 words, as in the actual T20 test-set.The reason to chose T20 over other two which are considered to be closer to real world scenario is that the language model is being built over a smaller corpus (1 million words) as a result the model would underperform badly on **MAL** and **T62** making it impossible to infer performance boost while experimenting.

4 Building the Intuition

According to (Mays et al., 1991) , α can be considered to be a prior belief that the observed input is correct.They state that it is a constant and does not change with the word being examined.

Lets consider a scenario where a subject is made to write a statement during which he/she is expected to introduce an error and then he/she is shown each word along with the words succeeding and preceding it.It is highly likely that his/her confidence about the displayed word will fluctuate depending upon how good the word fits into context.Thus pointing to the fact that α should not remain a constant but rather change with statistical information related to the word being examined.

(Mays et al., 1991) also distributed the remaining probability mass (1-α) among all the words that exist in the confusion set w_c.

This means that a word in w_c that is more likely to occur given a previous word in the sentence being examined would have the same probability of being replaced as compared to a word that is highly unlikely.The same argument applies when considering a word that follows.Hence it seems logical to not distribute the (1-α) equally but distribute it on the basis of some statistically information.

One might think about replacing the whole system with a simple tri-gram or bi-gram based probabilities but doing so would make the system to favour only those statements that the langauge model supports.α plays a critical role in preventing the language model from picking up only those words that the later can support.In other words it gives a weight to the subject's word choice over model's preference for word replacement.

I proposes a bi-gram based model for dynamic alpha and unequal distribution of probability mass among confusion set each, which is expected to improve the overall performance of the baseline.I later combine these 2 model to check how the 2 concepts work together.

5 The Dynamic Alpha Model

I expect to counter the static α assumption by introducing a new factor called the fluctuation factor β.This factor determines the percentage of probability mass of α that might change with the context, making α a dynamic term α_d.At any point of time the α_d of a word being examined for replacement is

$$\alpha_d = min(\alpha + (\alpha\beta(P(w|w_{i-1}) - 0.5)), 1) \quad (2)$$

The remaining probability mass .ie (1-α) is equally distributed among all the words in the confusion set as done in the originally in (Mays et al., 1991)

$$P(x|w) = \begin{cases} \alpha_d & \text{if } x = w \\ (1 - \alpha_d)/|w_c| & \text{if } x! = w \end{cases} \quad (3)$$

In this model, it might sometimes happen that there is no probability mass left to be distributed among the confusion set. Such situation would happen when bi-gram of preceding word and the word being examined is highly likely.In this way the model supports the subject's word when it too, is highly

confident about it.Therefore β needs to be selected carefully such that the α_d does not touch 1 even for small bi-gram probabilities.
The performance for detection and correction for various values of β are provided in **Table 1** and **Table 2** respectively.

Model	Precision	Recall	F1
β=0.05	0.77	0.67	0.72
β=0.10	0.77	0.68	0.72
β=0.20	0.75	0.74	0.74
Baseline	0.77	0.66	0.71

Table 1: Evaluation of Detection for Dynamic Alpha Model

Model	Precision	Recall	F1
β=0.05	0.61	0.61	0.61
β=0.10	0.60	0.63	0.62
β=0.20	0.57	0.69	0.62
Baseline	0.62	0.61	0.61

Table 2: Evaluation of Correction for Dynamic Alpha Model

6 The Unequal Probability Mass Model

The remaining probability mass ie.(1-α) can be distributed to words in confusion set on the basis of corresponding fraction of bi-gram probabilities f_p.So, for a particular word w_c^i in the confusion set, the fraction of bi-gram probabilities is

$$f_p(w_c^i) = \frac{p(w_c^i|w_{i-1})}{\sum_{j=1}^{|w_c|} p(w_c^j|w_{i-1})} \tag{4}$$

The probability that a word x from w_c might be chosen for replacement of w is

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ (1-\alpha) * f_p(x) & \text{if } x! = w \end{cases} \tag{5}$$

The performance for detection and correction, for comparison with baseline are provided in **Table 3** and **Table 4** respectively.

Model	Precision	Recall	F1
UPMM	0.75	0.77	0.76
Baseline	0.77	0.66	0.71

Table 3: Evaluation of Detection for Unequal Probability Distribution Model(UPMM)

Model	Precision	Recall	F1
UPMM	0.54	0.70	0.61
Baseline	0.62	0.61	0.61

Table 4: Evaluation of Correction for Unequal Probability Mass Model (UPMM)

7 The Combined Model

In this model I have combined the dynamic alpha model along with the unequal probability mass model, to find out how the two models would work. The following are the equations that run the combined model.

$$\alpha_d = min(\alpha + (\alpha\beta(P(w|w_{i-1}) - 0.5)), 1) \tag{6}$$

$$f_p(w_c^i) = \frac{p(w_c^i|w_{i-1})}{\sum_{j=1}^{|w_c|} p(w_c^j|w_{i-1})} \tag{7}$$

$$P(x|w) = \begin{cases} \alpha_d & \text{if } x = w \\ (1-\alpha_d) * f_p(w_c) & \text{if } x! = w \end{cases} \tag{8}$$

The performance for detection and correction for various values of β are provided in **Table 5** and **Table 6** respectively.

Model	Precision	Recall	F1
β=0.05	0.74	0.80	0.77
β=0.10	0.72	0.83	0.77
β=0.20	0.72	0.87	0.79
Baseline	0.77	0.66	0.71

Table 5: Evaluation of Detection for Combined Model

8 Conclusion

In terms of performance of correction abilities of the proposed models, all the models show an increased

Model	Precision	Recall	F1
β=0.05	0.53	0.75	0.62
β=0.10	0.52	0.77	0.62
β=0.20	0.5	0.82	0.62
Baseline	0.62	0.61	0.61

Table 6: Evaluation of Correction for Combined Model

recall but a decreased precision, such that the F-score remains comparable to that of baseline.This points to the fact that the false negative counts are less and false positive counts are more as compared to that of baseline, which also means that model is now able to wisely choose which sentences really require to be changed, however when it comes to selecting a correct word from the confusion set, it fails.

The performance of the detection abilities support the previous inference as there is significant increase in the recall as compared to the drop in precision, leading to higher F-score.

My intuition is that if the language model is trained using a larger set like the WSJ corpus then the combined model might get good at selecting a correct word from the confusion set.Also with the larger corpus,one has the privilege of building a larger vocabulary which would reduce the counts under any n-gram containing UNK tags.Thus leading to an overall increase in precision.

9 Future Work

One can work on increasing the precision by creating a language model based on a larger data-set in order to find out how the performance of these models varies and later test it on MAL and T62 test-sets.

Apart from this, one can also try a different type of n-gram like tri-grams or $P_{continuation}$ to build a new model that incorporates the information that can be obtained from succeeding words.

A new model can also be created where distribution of remaining probability mass is done on the basis of which pairs of letters are commonly replaced when a subject commits a mistake.This model can be built by drawing some intuition from the noisy channel model built by (Kernighan et al., 1990)

References

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 27(5):517–552.

Wilcox O'Hearn, Graeme Hirst, and Alexander Budanitksy. 2008. Real-word spelling correction with trigrams: A reconsideration of the mays, damerau, and mercer model.