Spelling Correction and Dictionary Enhancement

Hengfeng Li
henli@student.unimelb.edu.au
Department of Computer Science and Software Engineering
The University of Melbourne
Parkville
Melbourne, Australia 3052

Abstract

In this paper, two different approximate matching techniques are introduced including N-gram and Editex. And the comparison of effectiveness between these techniques is given by using the pooling strategy. Besides, a selection mechanism is developed, which is based on the noisy channel model through setting a threshold of the probability of the best match to decide that this query is a misspelling or a new word which should be added into the dictionary. And the evaluation of this selection mechanism is also described in this paper.

1 Introduction

The main two tasks of this project are, given a corpus and a dictionary, to both identify misspellings in the corpus and mine the corpus to identify new words that could be added to the dictionary. This report describes the N-gram and Editex methods which are two different approximating matching techniques applied and compares them through pooling strategy. Moreover, a selection mechanism based on noisy channel model and a evaluation its effectiveness are given in this paper.

2 Approximate Matching Techniques

2.1 N-gram

A N-gram is the N-length substrings of a string, e.g., the 2-gram of string 'lended' is le, en, nd, de, ed. The n-gram distance of two strings s and t is

$$|G_n(s)| + |G_n(t)| - 2 \times |G_n(s) \cap G_n(t)| \quad (1)$$

In this project, its implementation adds a character '#' at the begin and end of strings in order to improving the accuracy of similarity. For example, from equation (1), s = '#lended#', t = '#deaded#', the n-gram distance is $6 + 6 - 2 \times 3 = 6$. It represents the distance between two strings and the smaller the n-gram distance, the closer the match.

2.2 Editex

Editex, developed by Zobel and Dart (1996),is a phonetic matching technique which improves the edit-distance algorithm by using the phonetic method. However, instead of setting a fixed weight for insertion, deletion and replacement in normal edit-distance algorithm, it uses the phonetic method to divide the letters into groups whose sound are similar. And there are two penalties:

- High for letters that are never similar, such as "d" and "m".
- Low for letters that can give rise to similar sound, such as "m"-"n" and "c"-"k".

2.3 Effectiveness

For the evaluation purpose, a normal edit-distance technique is also implemented in this project. The results of table 1 are computed at rank 12, including precision, recall and average precision.

From table 1, it is obvious that the edit-distance is superior to other two techniques, which is different from the original expectation that the Editex is the best. The main reason is that the Editex is a phonetic matching technique which are based on the sound of words. It can be better applied at the searching for a name or word when the exact spelling is unknown. However, the test query strings are all the words with a few characters' misspelling and thereof the edit-distance are better suitable for that situation.

3 Selection mechanism

From the articles of Ahmad and Kondrak (2005) and Norvig (2007), a noisy channel model is applied to compute that the probability of a candidate is a misspelling word and the highest probability is chosen to be the best match.

Table 1: Performance results for different approximate techniques

Method	A			B			C		
	precision	recall	AP	precision	recall	AP	precision	recall	AP
N-gram	0.47	0.67	0.52	0.47	0.68	0.50	0.46	0.63	0.45
Editex	0.49	0.70	0.56	0.48	0.67	0.52	0.49	0.68	0.56
Edit distance	0.57	0.79	0.70	0.57	0.78	0.69	0.54	0.76	0.68

For the problem of deciding whether the unknown word is just a misspelling word or a new word, the lower probability of the best match, the higher chance that this query is a new word. If an appropriate threshold T is found, it can be used to distinguish a misspelling word and a new word.

3.1 Noisy channel model

It considers the word w to be a misspelling word of the correction c. In Ahmad and Kondrak (2005) and Norvig (2007), they find the best candidate c is to maximize the follow probability:

$$argmax_c P(w|c)P(c) \quad (2)$$

- $P(c)$ is the probability that a correction c appears. This is called language model.
- $P(w|c)$ is the probability that the word w is typed when a desired word is c. This i called error model.

As described in Ahmad and Kondrak (2005), the error model $P(w|c)$ to be proportional to the number of edit distance which turns c into w.

$$log[P(w|c)] = -ED(w, c) \quad (3)$$

where $ED(w, c)$ is the edit distance between w and c.

And the language model P(c) is the frequency of the word $F(c)$ divided by the total number of query tokens N in the corpus.

$$P(c) = \frac{F(c)}{N} \quad (4)$$

If a correction c has never been queried, it means that its probability is zero which would lead to a bad computation result. So an assumption is proposed that the original frequency of each word begin from one and is increased when a query happens. This general process is called smoothing.

Table 2: Precision of selection mechanism

Techniques	Threshold	Precision
N-gram	1.79e-6	34%
Editex	3.58e-6	52%
Edit distance	3.58e-6	44%

3.2 Evaluation

The table 2 is the precision of selection mechanism applied to different approximate techniques. The Editex is based on the edit distance algorithm the number of which is proportional to the $P(w|c)$, the error model. N-gram also can use this selection mechanism because it is similar to the edit distance algorithm in this experiment which the lower number, the closer match. The result is that the n-gram gets a lowest score and the Editex gets a higher score.

However, the precision of this selection mechanism is not good enough. The major two factors are that first, the corpus is too small to show the actual frequency of each word in the language model. Second, the most of unknown words are the names which are hard to identify whether is a new word.

4 The evaluation scheme

4.1 Evaluation scheme for approximate techniques

The evaluation scheme for the three approximate techniques is pooling. First, a dictionary file "words.txt" is use which contains 115326 words and three sets of query strings that each one includes 50 items are proposed.

And then using the pooling strategy: take the output top-30 answers of each method, find out a sample of overlap, and assume that these answers have been judged correct

Finally, find the correct matches in top-12 answers of each method(see table 3) and use them to compute precision, recall and average precision. Using these three merits is more convictive to decide which method has the greater effectiveness.

Table 3: Answers of a query *embaras*

Rank	N-gram	Editex	ED
1.	embarrass	embarrass	embarks
2.	embarks	embarks	embalms
3.	embarrasses	embalms	embanks
4.	embark	embers	embargo
5.	eras	embolus	embark
6.	arras	umbras	embarrass
7.	embalms	ambages	embers
8.	embanks	ambary	umbras
9.	embargo	cembalos	ambages
10.	embargoes	embanks	ambary
11.	embarrassed	embargo	baas
12.	embassies	embargoes	balas

4.2 Evaluation scheme for selection mechanism

The evaluation scheme for the selection mechanism uses the human assessment of the correctness of the results. 50 words are chosen from the corpus. And then the results of these words are applied to judge the correctness of the selection mechanism. The correctness are judged by the human checking each result.

5 Consideration for scalability

Table 4: Search time for different approximate techniques(running on specific machine)

Method	1 query	1415 queries
N-gram	221ms	219080ms
Editex	423ms	496242ms
Edit distance	245ms	282459ms

The table 4 is the search time for three different approximate techniques. It shows that the Editex almost spend twice time to complete one query than other two techniques. And the time needed for a query is similar between N-gram and edit-distance.

When the size of corpus or dictionary size grows, the problems of time and memory space may happens. First, the data structure of dictionary is a trie which need to spend a lot of memory space and the increasing size of corpus or dictionary is leading to the crash of system.

Second, the method of search answers is the exhausting search whose complexity is $O(n)$. A better method to find answers reasonably quickly is mentioned in Zobel and Dart (1996), which indexes short substrings of the words in the databases, sets of matches can be identified in a small fraction of a second.

On the other hand, with the increasing size of corpus, the precision of the selection mechanism should be improved.

6 Conclusions

Through testing three different approximate string techniques by pooling strategy, the evaluation shows that the edit distance has a greater effectiveness in spelling correction than N-gram and Editex. In this paper, a selection mechanism based on noisy channel model is developed to decide whether the word is a misspelling word. Although its precision does perform good enough, it can be improved by getting more accurate language model and category the name from other words.

References

Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a spelling error model from search query logs. *Proceedings of EMNLP 2005*, pages 955–962.

Peter Norvig. 2007. How to write a spelling corrector@ONLINE. *http://norvig.com/spell-correct.html*.

J. Zobel and P. Dart. 1996. Phonetic string matching: lessons from information retrieval. *In Proceedings of the 19th ACM International Conference on Information Retrieval (SIGIR'96)*, pages 166–172.