The practicality of approximate match methods

Anonymous

1 Introduction

Spellings are not always correct, so the users need suggestions. Only the users know what the correct words they want should be and what computer systems can do is to guess the possible candidates. An appropriate start for this problem is to make the assumption that a wrong spelling might be similar in some way to the correct one. This report analyses several methods which can be used to test the similarity between two words. Using these methods, a list of candidates can be produced from a dictionary. Evaluations of these methods of their effectiveness and efficiency are explained to measure their practical value to use to recommend corrections.

2 Dataset

The dataset includes a list of misspelled words and a list of its corresponding expected corrections with 719 items. A dictionary from Urban-Dictionary[1] which has 393954 items is provided to help the correction process. This is a real-world dataset, so the expected corrections are not guaranteed. A test directly match the items in misspelled list and corrected list with the dictionary has been done to check how many cases in the misspelled list are possibly solvable using the dictionary.

Dataset	Measure	Value
corrections	accuracy	594/716 (82.96%)
misspells	recall	6/716 (0.84%)
misspells	precision	6/175 (3.43%)

Table 1: Direct matching of the dataset

From the table, it can be seen that the accuracy of the corrections is 594/716, which means the 122 correct spellings do not exist in the dictionary. For these cases, correction suggestions from the dictionary can never succeed.

There is also another assumption that if a word in the misspelled list exists in the dictionary, it is not misspelled. It is because the dictionary cannot produce a better match than a perfect match and providing corrections for a probably right word seems to be strange. However, from the test result above, 175 words are considered right spellings while only 3.43% of them are correct indeed. This assumption can solve 0.84% of the problem and other 23.6% of the problem are wrong speculations.

More data sources such as frequency statistics are meaningful. These data can provide a wider cover of the solutions and can support decision makings when the system do not know whether a word is misspelled or not.

3 Methods Overview

3.1 Neighbourhood Search

The neighbourhood search method is to enumerate possible variants from a given spelling and then verify them. The variants are generated with one or more modification. In this report, the method adopts insertion, deletion and replacement. More times of modification will produce much larger candidate sets and its recall might be improved. K is the times of modification.

K	1	2
Positives	6642	143508
Avg Positives	9	200
True Positives	291	507
Recall	40.64%	70.81%
Precision	4.38%	0.35%
F1 Score	0.079	0.007
Time/ms	25.12	7658.25

Table 2: Neighbourhood search with different K (times of modification)

[1] http://urbandictionary.com

From Table 2, when doing twice modifications, the method can cover 174.22% corrections. However, it costs 30485.66% more time and gets only 7.99% precision. It produces much more candidates, and when K=1, users can get 9 candidates on average which has been enough. As the K increases, this problem will be more obvious. Therefore, K=1 is appropriate for this method.

3.2 N-Gram Distance

The n-gram distance method focuses more on evaluating whether words have similar combinations of components. Here the parameter N in the test represents the n of the method.

N	2	3	4
Positives	1469	1528	8095
Avg Positives	2	2	11
True Positives	149	97	59
Recall	20.81%	13.55%	8.24%
Precision	10.14%	6.35%	0.73%
F1 Score	0.136	0.086	0.013
Time/s	42.00	37.93	35.14

Table 3: N-gram distance ranking with different N

When N=2 or N=3, this method shows a good result. They all produce stable numbers of candidates and precisions are acceptable. For this dataset, N=2 turns to be the best choice and generally it is better then N=3. N=4 fails in most cases as the n-gram method flavours shorter words when it is difficult to find common components. Many words with two or three characters or even the alphabet exist in the dictionary, so they have more chances to be matched.

3.3 Global Edit Distance

The global edit distance is another method calculates similarity using a distance measure. It follows a similar idea with the neighbourhood search method, which is to test how to modify the spelling to get the correct one, but it is likely to run in that method's reverse path. For this method, the test applies two filters to get the candidates with highest mark, and the candidates with highest or second highest mark.

With the result, the candidates with the highest mark have a higher F1 score. It is similar to former methods that the second options are always much more than the best matching candidates, but these candidates cannot provide a same-order growth of precision. In this case, the precision when accepting candidates with second highest mark drops from 4.64% to 0.37%, which is a bad trade-off. In addition, the method responds all correction with the filtered highest-mark candidates, so this level of candidates filter is enough for practice.

Mark	Highest	with 2nd Highest
Positives	5566	125113
Avg Positives	8	175
True Positives	258	457
Recall	36.03%	63.83%
Precision	4.64%	0.37%
F1 Score	0.082	0.007
Time/s	76.25	76.25

Table 4: Global edit distance method with top candidates

4 Evaluation

As explained in the former section, the statistics below will compare methods with their preferred parameters.

Method	NS(K=1)[2]	NG(N=2)[3]	GED-1[4]
Avg Positives	9	2	8
Responds	652	716	716
Recall	40.64%	13.55%	36.03%
Precision	4.38%	10.14%	4.64%
F1 Score	0.079	0.136	0.082
Time/s	0.03	42.01	76.26
Avg Time/ms	0.04	58.67	106.51

Table 5: Comparison of matching methods

The neighbourhood search method is the fastest among the three methods. Then, the n-gram method has the best F1-score, which means it is generally the possible method which can provide a satisfying output. The result of the global edit distance method looks similar to the neighbourhood search method, but it is the slowest. But this data does not mean the neighbourhood search method or the n-gram distance method is the best, as the results vary for different spellings. It responds 652 spellings and for the rest 64 spellings it cannot provide any

[2]Neighbourhood search method with 1 modification
[3]N-gram distance method with 2-gram
[4]Global edit distance method with highest mark candidates

option. The other two methods respond all the correction requests.

Method	NS(K=1)	NG(N=1)	GED-1
Give	adn		
Want	and		
Possible[5]	true		
PosNum[6]	28	1	1
Positives	pdf,dn abn...	addn	addn
Recall	0	0	0
Precision	0	0	0

Table 6: Comparison of matching methods

In this case, all the three methods fail to get the correct spelling and they output 28, 1 and 1 suggestions separately. When all the methods cannot get the right answer, the less wrong options given is better. So the neighbourhood search method is a bad choice for this correction.

Method	NS(K=1)	NG(N=1)	GED-1
Give	phsycotic		
Want	psychotic		
Possible	true		
PosNum	0	1	1
Positives		photic	psychotic
Recall	0	0	100%
Precision	0	0	100%

Table 7: Comparison of matching methods

For the spelling phsycotic, GED-1 is the only one gives the right answer. This spelling has two wrong characters with its corresponding correct spelling. So the neighbourhood search allowing one modification does not work. Also, the two wrong characters distributed in the different offsets of the word, therefore the n-gram method cannot give good marks to the right matching as they do not share enough number of components. This example tells that although the global edit distance is the slowest one. It can sometimes the right answer for specific tasks.

5 Conclusions

This report evaluates several approximate matching methods' performances in spelling corrections. Different methods have different advantages as their flavours are not the same. For use in reality, a combination of these methods can apply. The precision of the approximate matching might be improved if statistics of real-world word frequency is provided. It also shows that in the non-contextual environment, doing matchings are sometimes not reasonable as the spelling might conflict with the users' sentences.

[5]Possible means the correct spelling exists in the dictionary

[6]PosNum is the number of the positives or candidates

References

Zobel, Justin and Philip Dart *Phonetic String Matching: Lessons from Informa-tion Retrieval.* In Proceedings of the Eighteenth International ACM SIGIR Conference on Research and Development in Information Retrieval. Zu?rich, Switzerland. pp. 166?173.

Naomi Saphra and Adam Lopez *Evaluating Informal-Domain Word Representations with UrbanDictionary* In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany. pp. 94?98.

F. J. Damerau *A technique for computer detection and correction of spelling errors* Communications of the ACM, vol. 7, no. 3, pp. 171?176, 1964.

Navarro, G *A guided tour to approximate string matching* ACM computing surveys (CSUR), 33(1), 31-88.