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[] 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 success.

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)

From Table 4, when doing twice modifications, the method can cover 174.22% correc-

tions. 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 in 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 exists 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 apply two filters to get the candidates with highest mark, and the candidates with highest or second highest mark.

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: Neighbourhood search with different K (times of modification)

4 Evaluation

5 Application

6 Conclusions

References

Martin Kay. 1986. Parsing in functional unification grammar. In K. Spark Jones B. J. Grosz and B. L. Webber, editors, *Readings in Natural Language Processing*, pages 125–138. Morgan Kaufmann Publishers, Los Altos.

Fredrick Mosteller and David Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts.

K. Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.