COMP90049 Report

Abstract

This paper investages several spell checking methods. The main goal is to compare and analysis the performance of spelling correction methods, on a peculiar data set: a number of headwords taken from UrbanDictionary1 that have been automatically identified as being misspelled ()

1 Introduction

Spelling correction is a basic task in natural language processing. The method of spelling correction have been very mature. Some neural method also been presented in recent years. In this paper, we investigate some non-neural network method to deal with it. Including soundex, n gram, edit distance and editex.

2 Method

In this section, we simplily introduce the algorithms we evaluated in our paper.
Soundex
N-Gram
Edit distance
Editex

3 Experiment

3.1 Dataset

Type	Words number
Testset size	716
Dictionary size	393954
Misspelled in Dictionary	175
Correct not in Dictionary	122

Table 1: Dataset

3.2 Settings

Soundex Calculate the soundex code for every word and matched with global edit distance.
N-Gram We evaluate the N-Gram algorithm for n in range 1 to 9. For a particular n, we first pad (n-1) # in the front and end of every word. This gurantee the building of n-gram set. For example, 5-gram set for word "he" defined as:

$$\{\#\#\#\#h, \#\#\#he, \#\#he\#, \#he\#\#, he\#\#\#, e\#\#\#\#\} \quad (1)$$

N	Predicted	Right	Precision	Recall
1	7150	183	2.56	25.56
2	1484	151	10.19	21.09
3	1429	149	10.43	20.81
4	1426	148	10.38	20.67

Table 2: N-gram algorithm results

Edit Distance There are two kinds of edit distance algorithm, local edit distance and global edit distance. In this paper, we implement both of them and evaluate them. For global edit distance, we have two distance calculate scheme: 1) (+1) for indel and mismatch and (-1) for match; 2) (+1) for indel and mismatch and do nothing for match. The different between them will discuss in Section **??**. For local distance algorithm, we use (-1) for indel and mismatch and (+1) for match, and always assign 0 if 0 is better.
Editex We calculate the editex follows () settings.

4 Results

5 Conclusions

Scheme	Predicted	Right	Precision	Recall
GED-1	5528	253	4.57	32.12
GED-2				
LED	727774	133	0.02	18.58

Table 3: Edit distance algorithm results

Method	Predicted	Right	Precision	Recall
Soundex	495146	436	0.09	60.89
Local Edit Distance	727774	133	0.02	18.58
Global Edit Distance				
N-Gram (N=2)	1484	151	10.18	21.09
Editex	2830	230	8.13	32.12

Table 4: All method results

Method	Misspelled	Correct	Matched set
Soundex	accually ahain	actually again	akal, axile, azalea, asylees, auxiliar, acculturational, ... awin, annoy, ani, aam, aani, aoyama, anne, ayme, anay, ...
Local Edit Distance	accually ahain	actually again	actually, tactually, unactually, contactually, ... disenchain, rechain, toolchain, toolchains, ...
Global Edit Distance	accually ahain	actually again	actually chain, amain, arain, again, ghain, alain, hain
N-Gram (N=2)	accually ahain	actually again	actually, ally ain
Editex	actually ahain	actually again	usually, actually, annually, facially, casually, chally, ... amain, attain, arain, again, alain, hain

Table 5: Demostrate of different algorithm's spelling correction result.