A String Similarity Measure Based on Orthographic and Phonetic Similarity for Spelling Correction

Do-Gil Lee[1*], Ilhwan Kim[1] and Seok Kee Lee[2]
[1] Research Institute of Korean Studies, Korea University,
Anam-dong 5-1, Seoul, 136-701, Korea
[2] Department of Industrial & Management Engineering, Hansung University,
Samseongyoro-16gil, Seoul, 136-792, Korea
{motdg, haigh}@korea.ac.kr and seelee@hansung.ac.kr

Abstract. The most commonly used string similarity measure for spelling correction is minimum edit distance (MED), which is based solely on the orthographic similarity between two strings. In order to overcome this shortcoming, this paper presents a more sophisticated similarity measure that considers both the orthographic and phonetic similarity between two strings. To demonstrate the effectiveness of the proposed measure, we implement and test a spelling correction system and apply it to two languages, namely English and Korean. We investigate the useful features for each language through various experiments and achieve 10-best accuracies of 95.2 for English and 97.4 for Korean.

Keywords: String similarity, Spelling correction, Edit distance, Phonetic algorithm, *n*-gram indexing

1 Introduction

Spelling correction is a function which automatically corrects a string containing spell errors. More generally, if a string not in a dictionary is found, similar words with the string listed in the dictionary will be suggested automatically. Spelling correction can be a very convenient function when users do not know the exact spelling of a word or make typing mistakes.

The core technique for spelling correction is an approximate string matching algorithm. The most commonly used string similarity measure for approximate string matching is minimum edit distance (MED) [1]. MED between two strings is defined as the minimum number of editing operations, e.g., insertion, deletion and substitution, required to transform one string into another.

MED is effective for differentiating orthographic similarity, but not for phonetic similarity. Therefore, MED is not suitable for catching phonetic or cognitive mistakes. This paper presents a string similarity measure that combines orthographic and phonetic features.

Recent approaches to spelling correction based on probability models [2], [3] require a large training set of spelling errors paired with the correct spelling of the

[*] Do-Gil Lee is a corresponding author.

word. Such data should be collected from real texts and manually corrected, which is time-consuming. The method proposed herein, with an 'unsupervised' manner, does not require massive data.

2 String Similarity Measure

The devised scoring function considers the following four features for string similarity: common letter feature, ***n***-gram overlap feature, edit distance feature, and phonetic feature.

The following four assumptions for the scoring function are made. First, two strings with more letters in common are more similar. Second, two strings with a higher value of ranking score function are more similar. Third, two strings with shorter edit distance are more similar. Fourth, two strings with shorter edit distance in phonetic code are more similar.

Various types of function can reflect these assumptions. Each assumption can be expressed in more than one function. Among the various combinations of these functions, the optimal can be determined experimentally, and the final score calculated by multiplication of the functions for the features.

3 Experiments

To evaluate the system performance, *k*-best accuracy is used, which is the ratio of the number of candidates containing the correct answer to the total *k* candidates suggested by the system (in this study, 10-best accuracy).

In this work, we tested our spelling correction system for both English and Korean. An English dictionary containing 1,300,000 headwords was used to generate candidates of strings for the query and an error/correct answer data set of 998 pairs was used. A Korean dictionary containing 500,000 headwords was used and error/correct answer data set of 2,579 was used.

Metaphone is used as a phonetic algorithm for English. Two cases were tested with 1) the front 4 bytes of the generated code, and 2) the entire phonetic code. For Korean, Kodex [4] was used.

To investigate the effect of the parameters used in each feature on the performance of the system, various combinations of parameters were tested.

Table 1. Top 5 results for English

C	N	E	S	*n*-gram size	Phonetic code	10-best acc.
1	1	1	1	2-gram	Metaphone(4byte)	**95.19**
1	0	1	2	2-gram	Metaphone(4byte)	94.79
2	0	1	2	2-gram	Metaphone(4byte)	94.69
1	1	1	2	2-gram	Metaphone(4byte)	94.69
1	1	3	1	2-gram	Metaphone(4byte)	94.69

Table 2. Top 5 results for Korean

C	N	E	S	n-gram size	10-best acc.
1	0	1	2	2-gram	**97.40**
1	0	1	3	2-gram	97.29
2	0	1	3	2-gram	97.29
1	0	1	2	3-gram	97.29
1	0	1	2	4-gram	97.21

Tables 1 and 2 exhibit 10-best accuracies of top 5 results with combining features.

4 Conclusion

We have presented a more sophisticated similarity measure than MED by considering both orthographic and phonetic similarity, and implemented a spelling correction system based on the proposed similarity measure. The method is language independent and was demonstrated for both English and Korean. The experimental results demonstrate that the proposed similarity measure outperforms MED in both languages for spelling correction problems.

Acknowledgements. This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MEST) (KRF-2007-361-AL0013).

References

1. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady. 707--710 (1966)
2. Brill, E. and Moore, R. C.: An improved error model for noisy channel spelling correction. In: the 38th Annual Meeting on Association for Computational Linguistics. Annual Meeting of the ACL. (2000)
3. Kristina Toutanova and Robert C. Moore: Pronunciation modeling for improved spelling correction. In: the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 144--151 (2002)
4. Kang, B. and Choi, K.: Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. In: the Fifth international Workshop on information Retrieval with Asian Languages, (2000)