Incorporating Bigram Statistics into Spelling Correction Tools

Bridget Thomson McInnes[b], Serguei V. Pakhomov[b], Ted Pedersen[a], and Christopher G. Chute[b]

[a]Department of Computer Science, University of Minnesota Duluth, MN, USA
[b]*Department of Medical Informatics, Mayo Clinic, Rochester, MN, USA*

Abstract

The study presented here describes an algorithm designed to automatically determine the correct spelling of misspelled words found in clinical notes. Based on the list of suggested spelling correction from a spelling suggestion tool, the correct suggestion is determined by incorporating the content words surrounding the misspelled word. The content words and the suggested spelling corrections form bigrams in which measures of association can be performed to determine the likelihood of a suggested word being the correct spelling.

Introduction

Spelling correction is an important segment of text normalization of clinical notes. The Electronic Medical Record at the Mayo Clinic consists of 16 million notes to date and is growing at the rate of 50-60,000 notes per week. Approximately 18% of these notes contain spelling errors. These errors may negatively effect Information Retrieval and Data Mining applications. The study presented here describes an algorithm designed to automatically determine the correct spelling of a misspelled word based on the list of suggested spelling corrections from the spelling suggestion tool, *Gspell*.

Bigram Model

The bigram model performs statistical analysis of the first content word prior and after a misspelled word in conjunction with a list of possible suggestions, created by *Gspell*, to automatically determine the proper correction of a spelling error. The bigram model determines the associations for each possible suggestion and content word. These two scores are combined using a weighted average to account for the importance of seeing both content words with the suggested word in our training set which was created from 2001/2002 clinical notes.

Results

The experiment was conducted on a test set compiled by a human annotator from 3,500 clinical notes consisting of 9,224 words and 322 misspellings. We found that measures of association, that include expected values in their calculation, such as Mutual Information, result in a lower precision while those that do not take the sample size into consideration perform better. Thus, we believe that there is a justification for using measures that do not include their expected values in their calculation.

Discussion

Examination of Log Likelihood scores showed bigrams seen a few times had similar scores to those seen often. Analysis showed that due to the occurrence of large expected values, as the actual frequency of the bigram deviated from the expected value in either direction the Log Likelihood score increased.

Conclusion

We found that context sensitive re-ranking of spelling suggestions produced by a minimum edit distance algorithm offer an improvement in terms of precision/recall; however, room for improvement still exists and can be diminished by using larger dictionaries and lemmatization approaches. We also found that large corpus size negatively affects association measures such as Log Likelihood.

Bigram model Results

Measure of Association	Precision	Recall	F Measure
Gspell	.33	.52	.40
Frequency	.35	.55	.42
Mutual Information	.33	.52	.40
Log Likelihood	.32	.50	.40
Dice Coefficient	.38	.59	.46
Phi Coefficient	.38	.59	.46

References

1. Church, K.W. and Hanks, P. Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Conference of the ACL, pages 76 – 82, 1989.

2. Church, K.W. and Gale, W.A. Probability Scoring for Spelling Correction. Statistics and Computing, Vol 1. pages 93-103. 1991.