MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Incorporating Bigram Statistics to Spelling Correction Tools

Bridget Thomson McInnes[a], 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*

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 the 2000/2001 clinical notes contain spelling errors. 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 uses statistical analysis of the words surrounding a misspelled word in conjunction with a list of possible suggestions to automatically determine the proper correction of the spelling error. The bigram model uses a list of spelling suggestions created by *Gspell* along with the first content word prior to and after the misspelled word. 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. The experiment using the raw frequency counts was conducted to determine whether the mere occurrence of a bigram is more important than the association of the words in the bigram establishing a base case. We found that measures of association, that include expected values in their calculation, result in a lower precision, seen in the Mutual Information results. Measures that do not take the sample size into consideration perform better, seen in the Phi and Dice Coefficient results. 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 that were seen a few times had similar scores to those seen often. Analysis showed that due to the occurrence of large expected values compared to actual values, as the actual frequency of the bigram deviated from the expected value in either direction the Log Likelihood score increased.

Out-of-vocabulary words were seen approximately 300 times out of the 652 words tagged as misspelled, accounting for most of the losses in precision/recall. We believe the results can be improved by a lemmatization approach such as the Lexical Variant Generator to normalize morphological forms of the words.

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 Meaure
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.

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

[*] http://umlslex.nlm.nih.gov/lvg