

Pre-processing to Improve the Classification of Chief Complaint Data

Richard D. Boyce B.S., Bryant T. Karras M.D., William B. Lober, M.D.

Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle

Abstract

We implemented and evaluated an automated text-normalizing process for ED CCs. The process was found to significantly reduce the number of misspellings, abbreviations, and truncations in CC data.

Introduction

Reducing orthographic variation in the CCs may improve the quality of syndrome classification for Syndromic Surveillance¹. Natural Language Processing (NLP) techniques have been used to map CC data to UMLS concepts². Similar techniques should improve automated classification of CCs into syndromic categories. As a first step towards an automated text-normalizing process for ED CCs, we implemented a program that:

1. removes stop words
2. searches for and replaces abbreviations, and truncations, and acronyms with their full terms
3. automatically replaces misspellings with unambiguous replacements

Methods

Stop words and punctuation symbols were removed from nine subsets of 1000 and one of 558 CCs from a community ED. These were placed into two separate groups, Groups 1 & 2.

Group 1 was interactively spell-checked using the GNU *aspell* program (aspell.sourceforge.net). Abbreviations (e.g. GI) and truncations (e.g. Poss) were not counted as misspellings. A tally of all misspellings showed an average of 25 per 1000 CCs.

Group 1 was then checked for abbreviations and truncations using a modified *aspell* dictionary with common truncations removed. A count of all abbreviations and truncations showed an average of 30 per 1000 CCs.

Each distinct abbreviation and truncation was recorded in context. Co-authors BK and BL provided common expansions for abbreviations and truncations. We identified those abbreviations and truncations with only a single expansion (e.g. GI is always gastro-intestinal but APE can represent Acute Pulmonary Embolism or Acute Pulmonary Edema). Using Python (python.org), we developed a program that replaced all the identified unambiguous abbreviations and truncations with their expansions in each of the 10 subsets of Group 1. The remaining abbreviations and truncations in each subset were tallied.

We then conducted a 10-fold cross-validation evaluation of the performance of the automated spell-checker. Each un-processed subset in Group 2 was processed using a spell-checker trained on the union of distinct, manually spell-checked, Group 1 subsets.

A non-interactive spell-checker was invoked to automatically replace misspellings with un-ambiguous substitutions. An un-ambiguous substitution is defined as a string that:

1. Can be made equivalent to the misspelled words with at most two deletions, insertions, exchanges, or adjacent swaps
2. Is present in a corpus of words created from a distinct set of CCs
3. Is either unique in the corpus or occurs in the corpus with the same left and/or right neighbors more often than any other candidate substitution

Results

Pairwise T-tests showed a significant reduction in both abbreviations/truncations and misspelled words (Table 1). The spell checker incorrectly replaced misspellings 2.3% of the time, 64% of all misspellings were correctly replaced.

<i>Parameter</i>	<i>Avg Pre</i>	<i>Avg Post</i>	<i>p-value</i>
Abbreviations ⁺	31/1000	4/1000	<.01
Misspellings	25/1000	9/1000	<.01

Table 1: Normalized average variation per data set, pre & post processing (*Including truncations)

Conclusion

The algorithm significantly normalized the data and established performance characteristics for our data set. These performance characteristics will be compared with those of similar approaches on other data sets, and the impact of this pre-processing on a standard classification algorithm will be assessed.

Acknowledgments:

This work was supported in part by NLM grant T15LM07442 and the Foundation for Healthcare Quality, US Army Medical Research Acquisition Activity W23RYX-3263-N612.

References

- [1] Shapiro A. Taming the variability in free text: Application to health surveillance. *Morbidity and Mortality Weekly Report*, 53 (Supplement): 95-100, 2004.
- [2] Travers D. Validation of a new tool for extracting terms from clinical text: emergency medical text processor. In: Proceedings of Medinfo 2004 (CD), 2004:1884-5.