Improving Spelling Correction with Consumer Health Terminology

Chris J. Lu, PhD[1, 2] and Dina Demner-Fushman, MD, PhD [1]
[1]National Library of Medicine, Bethesda, MD [2]Medical Science & Computing, LLC, Rockville, MD

Introduction

The demand of consumer health informatics (CHI) has grown rapidly due to the increase in Internet usage [1]. Consumers (patients, families, caregivers, and the general public) seek health information online every day. Automatic and semi-automatic consumer NLP systems are developed to reduce the cost for CHI. NLM launched the Consumer Health Information and Question Answering (CHIQA) project to help consumers find reliable health information. A consumer spelling tool (CSpell) was developed for spelling error correction in the NLP pipeline due to the high spelling error rate in consumer questions. The SPECIALIST Lexicon was used as the default dictionary in CSpell because all lexical entries are manually verified by linguists. A systematic approach was developed to retrieve consumer health terminology (CHT) from the UMLS Metathesaurus and MEDLINE. These terms were added to the dictionary in CSpell to improve system performance.

Methods

First, 32 semantic types (STs) were collected empirically from interventions, drugs, problems, anatomy and populations that are commonly used in CHI [2]. 1.85M English preferred terms, not abbreviations or acronyms, that match the above STs, were retrieved from UMLS-2017AB. The retrieved terms were then lowercased and tokenized resulting in 154,992 unique unigrams. These unigrams were further processed by excluding punctuation, digits, numbers, units, measurements, possessives and Lexicon entries to generate 39,042 words. These words form consumer health data (CHD). A 3,596 subset of CHD was found in the MEDLINE n-gram set (CHDM) [3]. CHD and CHDM respectively were added to the CSpell dictionary to test the effect of CHT on spelling correction.

Test, Results, Conclusion and Future Work

A development set for CSpell was established from the training and testing sets of the ensemble method of spelling correction [4]. This set consists of 471 consumer health questions with 24,837 tokens and 774 instances of annotated spelling error corrections. The results of the ensemble method and CSpell with different dictionaries (rows 3-5) are shown in Table 1. We observed: 1) CSpell with default dictionary has a slightly better performance than the ensemble method and 2) an increase in the precision, recall, and F1 with the addition of CHD and CHDM (best-case). This confirms that the CHT is indeed an important resource for consumer NLP systems.

In addition, CHDM and the associated terms were used as lexicon entry candidates and added to the Lexicon. We plan further acquisition of CHT by establishing consumer health corpora from various NIH websites and using them as additional CHT in CSpell dictionary for better performance, as well as using them as resources for broader coverage of the Lexicon.

Table 1. Test result: Lexicon with CHD and CHDM.

Dictionary	TP*	FP*	Precision	Recall	F1
Ensemble method	553	272	67.03%	71.45%	0.6917
Lexicon (default)	565	264	68.15%	73.00%	0.7049
Lexicon + CHD	591	187	75.96%	76.36%	0.7616
Lexicon + CHDM	592	185	76.19%	76.49%	0.7634

*T: True, F: False, P: Positive

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We thank Dr. H. Kilicoglu, Dr. L. McCreedy and W. Rogers for their valuable discussions.

References

1. K Roberts, D Demner-Fushman, Interactive use of online health resources: a comparison of consumer and professional questions, J Am Med Inform Assoc. 2016; 23(4):802-811.
2. https://lsg3.nlm.nih.gov/LexSysGroup/Projects/lexicon/current/docs/designDoc/UDF/CHD/index.html.
3. CJ Lu, D Tormey, L McCreedy, AC Browne, Generating A Distilled N-Gram Set: Effective Lexical Multiword Building in the SPECIALIST Lexicon, The 10th International Joint Conference on Biomedical Engineering Systems and Technologies, HEALTHINF 2017, PORTO, Portugal, Feb. 21-23, 2017; Vol (5):77-87.
4. H Kilicoglu, M Fiszman, K Roberts, D Demner-Fushman, An Ensemble Method for Spelling Correction in Consumer Health Questions, AMIA 2015 Annual Symposium, San Fran., CA, Nov. 14-18, 2015; p. 727-736.