Extracting Everyday Health Interests from Online Communities

Albert Park, MS¹, Andrea Hartzler, PhD², Jina Huh, PhD¹, David McDonald, PhD³, Wanda Pratt, PhD²,³
¹Biomedical Informatics & Medical Education; ²Global Health; ³Information School, University of Washington, Seattle, WA

Abstract

Increasing numbers of patients are using online communities to seek support by posting personal stories and requesting information. To leverage this trend, we investigated Natural Language Processing (NLP) tools for extracting users’ health interests from online communities. We identified relevant semantic types of the Unified Medical Language System (UMLS) for online communities. The results suggest that we can repurpose NLP tools for online communities without losing critical information.

Introduction

Patients contribute a wide variety of narrative text to online health communities that can require thorough reading to gain insight. Although NLP tools hold promise for extracting health terms from this vast amount of text, existing tools are built for formally structured biomedical text. Thus, when they are applied to the informal online text generated by patients, they create as much noise as insight. In this work, we are repurposing and assessing whether NLP tools built for biomedical literature can be used to extract users’ health interests from online communities. In particular, we investigate effectiveness of processing online text with selected UMLS semantic types.

Methods

We repurpose an NLP tool by first selecting semantic types and then performing a relevance assessment. The semantic type selection process started with the authors independently rating the relevance of each of 133 available semantic types based on the description in the UMLS Metathesaurus. We then processed a small random sample online text from an online cancer community using MetaMap to extract terms as examples of semantic types. Based on the examples and previous initial ratings, we worked as a group to judge whether each semantic type was appropriate for identifying patients’ health interests.

Next, we conducted a relevance assessment to evaluate the effectiveness of processing online community text with the selected semantic types. Although using less semantic types would reduce the number of extracted terms, we wanted to retain high recall of relevant terms. To assess this cost of processing with selected semantic types, we randomly sampled 100 terms: 50 terms from positive-predicted terms (i.e., judged relevant by MetaMap) and 50 terms from negative-predicted terms (i.e., judged irrelevant by MetaMap) when applying semantic type selection. We blindly evaluated the relevance of each term in the sample, and then designated this human assessment as the gold standard. Finally, we calculated the precision and recall of semantic selection against the human assessment.

Results and Conclusion

We selected 25 of 133 available UMLS semantic types as most relevant to the everyday health interests of online community members. Processing community text with selected semantic types reduced the total number of extracted terms from 910 to 225. The findings from the relevance assessment in Figure 1 correspond with precision and recall rates of 0.83 and 0.90, respectively. On average, we reduced 140 of NL-processed terms for each sample while retaining most of the biomedical concepts related to users’ health interests. Validating results of NL-processed terms with users who contributed the online community text will further strengthen these findings. In conclusion, recalibrating semantic types of biomedical NLP tools is a promising approach to narrow the complexity of processing the everyday linguistic structure of online community text. We can leverage a valuable approach toward helping patients navigate the vast amount of online text, and pick out specific information that is most relevant to their personal interests.

Acknowledgment: This work was supported by NSF SHB 1117187.