Published in final edited form as:

Stud Health Technol Inform. 2015 ; 216: 1033–1034.

A Frequency-based Strategy of Obtaining Sentences from Clinical Data Repository for Crowdsourcing

Dingcheng Li[a], **Majid Rastegar Mojarad**[a], **Yanpeng Li**[a], **Sunghwan Sohn**[a], **Saeed Mehrabi**[a], **Ravikumar Komandur Elayavilli**[a], **Yue Yu**[a,b], and **Hongfang Liu**[a]

[a]Mayo Clinic, Rochester, MN, USA

[b]Department of Biomedical Informatics, University of Jilin, Jilin, China

Abstract

In clinical NLP, one major barrier to adopting crowdsourcing for NLP annotation is the issue of confidentiality for protected health information (PHI) in clinical narratives. In this paper, we investigated the use of a frequency-based approach to extract sentences without PHI. Our approach is based on the assumption that sentences appearing frequently tend to contain no PHI. Both manual and automatic evaluations on 500 sentences out of the 7.9 million sentences of frequencies higher than one show that no PHI can be found among them. The promising results provide potentials of releasing those sentences for obtaining sentence-level NLP annotations via crowdsourcing.

Keywords

patient health information; clinical notes; high-frequent sentences; bigram filtering; crowdsourcing; de-identification

Introduction

Crowdsourcing has emerged as a popular method to generate training data for machine learning, natural language processing (NLP) and related fields[1,2]. However, in clinical NLP, no one has taken advantage of crowdsourcing since it may reveal the protected health information (PHI). Until now, no de-identification tools could guarantee 100% PHI removal from clinical narratives. Consequently, it remains infeasible to employ crowdsourcing for generating training data for clinical NLP. In this work, we propose a simple frequency-based approach to extract sentences containing no PHI under the assumption that sentences appearing frequently tend to contain no PHI. Based on this approach, it could be possible to generate training data through crowdsourcing for various sentence-level clinical NLP tasks such as concept identification or relation extraction.

Materials and Methods

In this work, we used all of electronic medical records (EMR) extracted from enterprise data trust (EDT) of Mayo Clinic. The EDT stores structured data and unstructured texts from a comprehensive snapshot of Mayo Clinic's service areas. It includes clinical notes, hospital summary, post-procedure notes, procedure note, progress note, tertiary trauma and transfer note. It is composed of 39.7 million records and 1.7 million tokens. We collected high-frequent sentences using following steps: building bigram repository; constructing sentence repository; bigram filtering; sentence frequency filtering and segmentation; dictionary-lookup filtering and sentence distribution analysis.

Experiments

System Results

The running of the above-described workflow yielded about 7.9 million unique sentences with frequencies higher than one (the total number was 276.7 million with an average frequency of 35). We then divided the whole sentence repository into 10 intervals based on log2 frequency. Although the numbers of sentences for each interval were different, we randomly sampled 50 sentences at each interval for evaluation. Meanwhile, we also sampled 200 more sentences the same way from sentences which had passed bigram filtering but only appeared once (the total number of such sentences is 109.2 million) for comparison purpose. Accordingly, the total number of sentences for evaluation is 700.

Manual Evaluation

The 700 sentences were assigned to four experienced reviewers to assess whether PHI could be found among them. In the event that any such information was found, reviewers were required to fill name, profession, location, age, date, contact, ids and comments for other sensitive information into a spreadsheet. The evaluation results (the union of the four reviewers' results), showed that, as expected, most PHI occurred in the first interval. The most frequent PHI element was name – appearing 7, 3, and 2 times in the first, second, and third interval respectively. Similar patterns were seen for profession as well. For age and date, 4 and 13 times were in the first interval and 1 and 3 times in the second interval. No contact was found for all intervals and ids showed one time in the first three intervals.

Conclusion

In summary, we proposed a method based on frequencies to extract sentences containing no PHI from a clinical data repository. We experimented in utilizing Mayo's EDT clinical notes. About 7 million unique sentences which appear more than two times were extracted. The final evaluation on 700 sampled sentences show that nearly no PHI can be found from sentences with higher frequencies (500 of them have frequencies higher than one). As follow-up steps, we will explore ways to make the crowdsourcing for clinical notes realistic and develop corresponding systems.

Acknowledgments

This work was made possible by joint funding from National Institute of Health R01GM102282, R01LM11369, R01LM11829, and R01 LM011934.

References

1. Safire W. On Language. New York Times Magazine. 2009 Feb 5.

2. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. Journal of medical Internet research. 2013; 15