Automatic Identification of Lifestyle and Environmental Factors from Social History in Clinical Text

Meliha Yetisgen, PhD[1,2], Elena Pellicer[3], David R. Crosslin, PhD[1], Lucy Vanderwende, PhD[4,1]
[1]Biomedical and Health Informatics, [2]Department of Linguistics, [3]School of Medicine, University of Washington, Seattle, WA
[4]Microsoft Research, Redmond, WA

Abstract

Lifestyle and environmental factors play a significant role both in clinical research as well as clinical care. In clinical research, it has been established[1] that 5-10% of cancers can be attributed to hereditary factors, while 90-95% have been found correlated with lifestyle and environmental factors such as smoking, diet and exercise. For clinical care, it has long been practice to record social history during clinical care as this history impacts not only diagnosis but also treatment options[2]. We therefore propose in this work to automatically identify those lifestyle and environmental factors that clinical caregivers have documented. We extended Milton et. al.'s analysis of social and behavior information[3] and Uzuner et. al.'s information on smoking in discharge summaries[4].

Dataset

We created a corpus from MTSamples website (http://www.mtsamples.com/). The website provides a large collection of publicly available transcribed medical records. We scraped 516 history and physical notes since these reports contain very rich social history information. We applied our in-house statistical section chunker (http://depts.washington.edu/bionlp/index.html?software) and identified 342 sections tagged as social history in 516 reports for annotation.

Annotation Process

We created a detailed annotation guideline to annotate the following lifestyle and environment factors: (1) substance abuse (smoking, alcohol and drug use), (2) occupation, (3) marital status, (4) family information, (5) residence, (6) living situation, (7) environmental exposures, (8) physical activity, (9) weight management, (10) sexual history, and (11) infectious disease history. We then defined 9 different dimensions that might apply to each type of factor; i.a., for substance abuse (1), annotations are made regarding *status* (possible values: past, current, none, unknown), *time frame* (e.g. since 2010), *method* (e.g. drink, inhale, inject), *type* (e.g. cigarettes, wine, cocaine), *amount* (e.g. # of cigrettes|drinks), *frequency* (e.g. daily, socially, rarely), and *history* (e.g. after 10 years of smoking), while for occupation (2), *location* and *extent* (e.g. part-time, night-shift) dimensions are annotated.

Using the BRAT rapid annotation tool, two annotators each annotated 20 social history sections. In the first round, inter-rater agreement was 0.59 F1 for the 11 lifestyle and environmental factors and their 9 dimensions. The annotators met and resolved all the conflicts, and the annotation guideline was updated. A single annotator is in the process of annotating the rest of the dataset. Annotation of 120 social history sections has been completed.

Conclusion

The social history section in clinical text indeed contains a wealth of information regarding a patient's lifestyle and environmental factors, which can be used in both clinical care and in clinical research. We are in the process of building automated extractors based on the annotated set. We will release both the annotated corpus and the extractors to the research community. Our research goal is to apply these extractors to EMRs to facilitate robust correlation studies between these factors and disease outcomes.

Acknowledgements

This work was supported by University of Washington Institute of Translational Health Sciences UL1TR000423.

References

[1] Anand P, Kunnumakara AB, Sundaram C, et al. Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research*. 2008; 25(9):2097-2116.
[2] Srivastava R. Complicated Lives – Taking the Social History. NEJM 2011; 265:7: pp. 587-589.
[3] Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. AMIA Annu Symp Proc. 2012;2012:625-34.
[4] Uzuner Ö., Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. *2008; 15(1)15-24.*