The Lancet Regional Health - Western Pacific 9 (2021) 100132

Contents lists available at ScienceDirect

The Lancet Regional Health - Western Pacific

journal homepage: www.elsevier.com/locate/lanwpc

Commentary

Applying natural language processing to electronic medical records for estimating healthy life expectancy

Rebecka Weegar

Department of Computer and Systems Sciences, Stockholm University, Sweden

ARTICLE INFO

Article history:
Received 4 March 2021
Accepted 4 March 2021
Available online 21 March 2021

Health-adjusted life expectancy (HALE), measures the overall health of a population by adjusting life expectancy for years lived with disability. HALE allows for monitoring changes of population health over time, by providing a measure of the number of healthy years expected for different age groups. HALE calculations require data on morbidity, but even though high quality data sources are crucial for estimations of population health, such sources are not available for all countries, specifically in lower income regions [1].

In *health-adjusted life expectancy (HALE) in Chongqing, China, 2017: an artificial intelligence and big data method estimating the burden of disease at city level*, Liang Xu and colleagues propose an alternative method for estimating prevalence of injury and disease, by leveraging information extracted from electronic medical records (EMRs) [2]. In their study, health records of close to a million patients were collected from a wide range of hospitals and health centres in Chongqing. These records included both structured data containing ICD-10 codes, and free text notes [2].

Using EMR data comes with a number of challenges requiring careful consideration, specifically for automatically analysing their free text content. Here, the work by Liang Xu and colleagues addresses two important natural language processing (NLP) tasks: to identify mentions of diseases in free text clinical notes, and to align these mentions with a standardized terminology.

Named entity recognition is a method for extracting entities from free text, where the entities in this case are mentions of disease. In their study, Liang Xu and colleagues trained a neural network model, known as a Bidirectional Long Short-Term Memory network (BiLSTM), to recognize these entities. BiLSTM is a robust method for named entity recognition [3], and similar neural architectures have previously been shown to be effective for extracting clinical entities from health records in Chinese [4].

To be able to summarize the contents of free text health records, one possible strategy is to align the entities extracted from the free text with a terminology or other structured representation, such as ICD-10. Free text, however, always allows for variation of form, where the same condition can be described in multiple ways, and it is rarely the case to find the exact wording of an ICD-10 code descriptor used in a health record.

To determine which extracted entities correspond to which ICD-10 codes, an NLP method for representing text known as word embeddings can be utilized. When a text is represented by word embeddings, each word in the text is mapped to a numerical vector of a fixed length. These word vectors are derived from large text corpora (large text collections). Such word vectors have many applications within natural language processing, as they can capture a semantic relationship between words [5]. Words that have a similar meaning will appear in similar contexts in the source corpora and therefore be represented by similar vectors. The semantic similarity between words thus becomes measurable, making it possible to find an alignment between entities extracted from the free text and ICD-10 code descriptors.

Liang Xu and colleagues set a threshold for this similarity measure; if an entity from the health records and an ICD-10 descriptor were found to have a similarity score over this threshold, they were considered as representing the same concept [2]. This method made it possible to create a joint representation of both the structured and the free text data in the included health records. Evaluation confirmed the importance of careful analysis of free text information in EMRs, as a considerable portion of the diseases mentioned in the free text notes were never recorded in the structured parts of the EMRs, and would be missed in the final analysis if only the structured data from the EMRs had been included [2].

All ICD-10 codes and descriptors extracted from the health records were next mapped to the disease categories of the Global Burden of Disease study [6], and Liang Xu and colleagues used Sul-

E-mail address: rebeckaw@dsv.su.se

https://doi.org/10.1016/j.lanwpc.2021.100132

livan method for the HALE calculations [2]. The main sources of ill health were found to be cancer, injuries—particularly for the male population—and cerebrovascular disease. The HALE estimations, 68•9 and 74•4 years for the male and female population respectively, were found to be in line with previous estimations of HALE for Chinese populations, showing the feasibility of the overall approach [2].

EMRs contain rich information, but there are still open questions regarding how to best integrate EMR data with other data sources. The completeness of EMR data can be difficult to validate and it is possible to question the overall representativeness of EMR data to a population, as there might, for example, be unequal access to health care. On the other hand, as Liang Xu and colleagues point out, using EMRs can allow for larger sample sizes compared to using surveys for estimating disease prevalence [2]. Another important advantage of EMR data is that it is continuously collected, meaning that with appropriate NLP methods for analysis, the data contained in health records can facilitate continuous monitoring of changes in population health over time, which can have positive impact on planning of health interventions, also at the regional level.

Declaration of Interests

Dr. Weegar has nothing to disclose.

References

[1] World Health Organization. World health statistics 2020: monitoring health for the SDGs, sustainable development goals. 2020.
[2] Ruan X, Li Y, Jin X, et al. Health-adjusted life expectancy (HALE) in Chongqing, China, 2017: an artificial intelligence and big data method estimating the burden of disease at city level. Lancet Region Health – West Pac 2021. doi:10.1016/j.lanwpc.2021.100110.
[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360, 2016.
[4] Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019;92:103133.
[5] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546, 2013.
[6] World Health Organization The global burden of disease: 2004 update. World Health Organization; 2008.