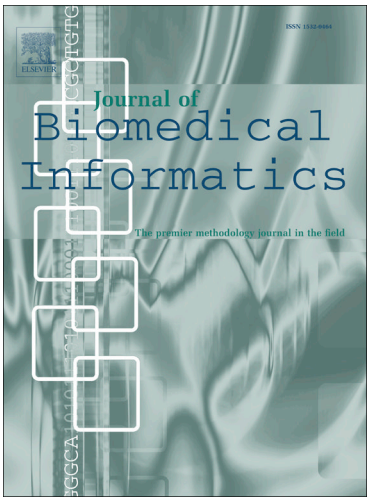

De-identification of medical records using conditional random fields and long short-term memory networks

Zhipeng Jiang, Chao Zhao, Bin He, Yi Guan, Jingchi Jiang

PII: S1532-0464(17)30222-8
DOI: https://doi.org/10.1016/j.jbi.2017.10.003
Reference: YJBIN 2868

To appear in: *Journal of Biomedical Informatics*

Received Date: 31 January 2017
Revised Date: 30 September 2017
Accepted Date: 3 October 2017

Please cite this article as: Jiang, Z., Zhao, C., He, B., Guan, Y., Jiang, J., De-identification of medical records using conditional random fields and long short-term memory networks, *Journal of Biomedical Informatics* (2017), doi: https://doi.org/10.1016/j.jbi.2017.10.003

De-identification of medical records using conditional random fields and long short-term memory networks

Zhipeng Jiang* HIT.JIANG@HOTMAIL.COM
Chao Zhao* ZHAOCHAOCS@GMAIL.COM
Bin He HEBIN_HIT@HOTMAIL.COM
Yi Guan† GUANYI@HIT.EDU.CN
Jingchi Jiang JIANGJINGCHI0118@163.COM
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, Heilongjiang, 150001, CHN

Abstract

The CEGS N-GRID 2016 Shared Task 1 in Clinical Natural Language Processing focuses on the de-identification of psychiatric evaluation records. This paper describes two participating systems of our team, based on conditional random fields (CRFs) and long short-term memory networks (LSTMs). A pre-processing module was introduced for sentence detection and tokenization before de-identification. For CRFs, manually extracted rich features were utilized to train the model. For LSTMs, a character-level bi-directional LSTM network was applied to represent tokens and classify tags for each token, following which a decoding layer was stacked to decode the most probable protected health information (PHI) terms. The LSTM-based system attained an i2b2 strict micro-F_1 measure of 0.8986, which was higher than that of the CRF-based system.

Keywords: Protected health information, de-identification, conditional random fields, long short-term memory networks

1. Introduction

The Electronic Health Record (EHR) is the systematized collection of electronically stored health information of patients in digital format [1]. It consists of a large amount of medical knowledge, and is a novel and rich resource for clinical research. A limitation of the large-scale use of EHR is the privacy of information contained in the text. To protect the privacy of patients and medical institutions, the US Congress passed the Health Insurance Portability and Accountability Act (HIPAA) in 1996. HIPAA defines 18 kinds of protected health information (PHI) that must be removed before the EHR can be reused, such as names, all geographic subdivisions smaller than a State, and so on[1]. The

∗. These authors contributed equally to this work.
†. corresponding author
1. https://www.hipaa.com/hipaa-protected-health-information-what-does-phi-include/ lists all PHIs in the HIPAA