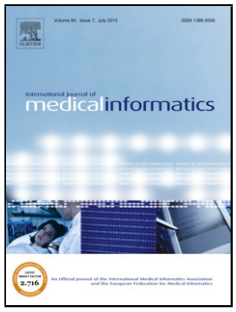

Title: Towards automated clinical coding

Author: Finneas Catling Georgios P. Spithourakis Sebastian Riedel

PII: S1386-5056(18)30403-9
DOI: https://doi.org/doi:10.1016/j.ijmedinf.2018.09.021
Reference: IJB 3757

To appear in: *International Journal of Medical Informatics*

Received date: 15-5-2018
Revised date: 21-8-2018
Accepted date: 30-9-2018

Please cite this article as: Finneas Catling, Georgios P. Spithourakis, Sebastian Riedel, Towards automated clinical coding, *<![CDATA[International Journal of Medical Informatics]]>* (2018), https://doi.org/10.1016/j.ijmedinf.2018.09.021

Towards automated clinical coding

Finneas Catling[a,b,*], Georgios P. Spithourakis[a], Sebastian Riedel[a]

[a] *University College London, Gower Street, London, UK. WC1E 6BT*
[b] *28 Davidson Gardens, London, UK. SW8 2XB (present address)*

Abstract

Background. Patients' encounters with healthcare services must undergo clinical coding. These codes are typically derived from free-text notes. Manual clinical coding is expensive, time-consuming and prone to error. Automated clinical coding systems have great potential to save resources, and realtime availability of codes would improve oversight of patient care and accelerate research. Automated coding is made challenging by the idiosyncrasies of clinical text, the large number of disease codes and their unbalanced distribution.

Methods. We explore methods for representing clinical text and the labels in hierarchical clinical coding ontologies. Text is represented as term frequency-inverse document frequency counts and then as word embeddings, which we use as input to recurrent neural networks. Labels are represented atomically, and then by learning representations of each node in a coding ontology and composing a representation for each label from its respective node path. We consider different strategies for initialisation of the node representations. We evaluate our methods using the publicly-available Medical Information Mart for Intensive Care III dataset: we extract the history of presenting illness section from each discharge summary in the dataset, then predicting the International Classification of Diseases, ninth revision, Clinical Modification codes associated with these.

Results. Composing the label representations from the clinical-coding-ontology nodes increased weighted F1 for prediction of the 17561 disease labels to 0.264-0.281 from 0.232-0.249 for atomic representations. Recurrent neural network text representation improved weighted F1 for prediction of the 19 disease-category labels to 0.682-0.701 from 0.662-0.682 using term frequency-inverse document frequency. However, term frequency-inverse document frequency outperformed recurrent neural networks for prediction of the 17561 disease labels.

*Corresponding author
URL: f.catling@ucl.ac.uk (Finneas Catling), g.spithourakis@cs.ucl.ac.uk (Georgios P. Spithourakis), s.riedel@cs.ucl.ac.uk (Sebastian Riedel)