Article published online: 2022-12-04

Content Summaries of Best Papers for the Natural Language Processing Section of the 2022 IMIA Yearbook

Flamholz ZN, Crane-Droesch A, Ungar LH, Weissman GE

Word embeddings trained on published case reports are lightweight, effective for clinical tasks, and free of protected health information

J Biomed Inform 2022 Jan;125:103971

The authors highlight two main points that may impact NLP research: (i) the importance of clinical data to train NLP systems, although accessing those data may be complex out of hospitals, and (ii) the interest of word embeddings to represent the structure of the language. In order to prevent from using potentially identifying data, the authors trained word embeddings on several datasets (PMC clinical cases, English Wikipedia, MIMIC-III corpus, and clinical notes from the Pennsylvania University where they come from). Then, they used those embeddings on distinct NLP tasks (mortality prediction, de-identification). They conclude that using clinical cases allows them to achieve the best consistency in their experiments despite a lower number of tokens in comparison with other corpora. In addition, as expressed in the title of this paper, they remind that clinical cases are still de-identified.

Majewska O, Collins C, Baker S, Björne J, Brown SW, Korhonen A, Palmer M

BioVerbNet: a large semantic-syntactic classification of verbs in biomedicine

J Biomed Semantics 2021 Jul 15;12(1):12

This paper concerns verbal reasoning, a linguistic feature that may improve the performances of NLP models in several tasks. Nevertheless, the manual construction of lexicons is still challenging and time-consuming. In order to tackle this point, the authors produced BioVerbNet, a resource for the English language that proposes semantic-syntactic verb classes for 639 verbs. They produced this network using neural classification and expert annotations. BioVerbNet was then used for automatic classification at document and sentence levels. To this end, the experiments made rely on the merging of verbs belonging to the same class (identified through BioVerbNet) in existing word embeddings, thanks to a vector retrofitting approach.

Ding X, Mower J, Subramanian D, Cohen T

Augmenting aer2vec: Enriching distributed representations of adverse event report data with orthographic and lexical information

J Biomed Inform 2021 Jul;119:103833

This paper deals with post-marketing drug event surveillance, or pharmacovigilance. In a previous work, the authors presented aer2vec, a method to produce distributed a representation of drugs and adverse events from reports. In this work, they added subword embeddings and applied vector retrofitting NLP techniques. By combining subword embeddings and vector retrofitting, they improved the identification of relationships between drugs and adverse events. In addition, the authors highlighted that this approach does not require an extensive manual preprocessing stage.

De Angeli K, Gao S, Danciu I, Durbin EB, Wu XC, Stroup A, Doherty J, Schwartz S, Wiggins C, Damesyn M, Coyle L, Penberthy L, Tourassi GD, Yoon HJ

Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types

J Biomed Inform 2022 Jan;125:103957

This paper concerns the classification of cancer pathology reports, using convolutional neural networks (TextCNN). The authors highlight that CNNs are dependent on the distribution of data. Due to the prevalence of cancer types, the authors observed an imbalance between classes, which has a negative impact on system performances. In order to tackle this issue, they proposed a novel implementation of ensemble learning (using multiple models) which is based on class weight. To compute those weights, they defined an undersampling threshold and defined a significant weight for majority classes. This new ensemble learning outperforms other class imbalance techniques. The authors also provide several recommendations to solve this issue.