Article published online: 2021-09-03

Content Summaries of Best Papers for the Natural Language Processing Section of the 2021 IMIA Yearbook

Jin Q, Tan C, Chen M, Liu X, Huang S

Predicting clinical trial results by implicit evidence integration

Proc of Empirical Methods in NLP; 2020

The clinical trial result prediction (CTRP) task is based on medical literature containing PICO (how the Intervention group compares with the Comparison group in terms of the measured Outcomes in the studied Population). The authors proposed an EBM-Net model which is a transformer model that uses unstructured sentences as implicit evidences and a fine-tuning approach. They compared their fine-tuned model w.r.t. the BioBERT model and other approaches (MeSH ontology, bag-of-words, …etc.) and achieved better results on the COVID-19 clinical trials' dataset (22 clinical trials from the CORD-19 dataset).

Poerner N, Waltinger U, Schütze H

Inexpensive domain adaptation of pretrained language models: case studies on biomedical NER and Covid-19 QA

Proc of Empirical Methods in NLP; 2020

The authors highlight the expensive cost of domain adaptation while training a model on target-domain text. This cost is expressed in terms of hardware requirement, high running time and negative impact on the CO2 footprint. The authors investigated a solution they called GreenBioBERT that relies first on a Word2vec training stage on target-domain texts (PubMed, PMC, CORD-19), and second on the alignment of word vectors with the vectors from BioBERT. They applied their model on two issues: 8 biomedical NER tasks in English, and question-answering (QA) on COVID-19 issue. The authors achieved competitive results using BioBERT and a better precision on a few tasks; on the COVID-19 QA task, their model achieved better results than the SQuADBERT model (designed for QA). In this paper, the authors proposed a useful method to use existing pretrained language models in order to adapt them to new datasets, new tasks, new languages, etc.

Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, Cardinal R, Roberts A, Stewart R, Velupillai S

Generation and evaluation of artificial mental health records for Natural Language Processing

NPJ Digit Med 2020;3:69

The main problem for biomedical NLP is the difficult access to clinical documents and the inherent complexity to completely de-identify documents. The solution proposed by the authors consists in generating artificial discharge summaries. In this paper, the authors produced artificial summaries in mental health, based on the MIMIC-III data. Then, they used their artificial texts in order to train models using the Keras toolkit for classification tasks. The authors observed that models trained on their synthetic data perform as well as models trained on real data. Since the synthetic discharge summaries have been produced taking as input the MIMIC-III data, the authors cannot share their resources. Nevertheless, the method is reproducible.