Published online: 2020-08-21

Content Summaries of Best Papers for the Natural Language Processing Section of the 2020 IMIA Yearbook

Guan J, Li R, Yu S, Zhang X

A Method for Generating Synthetic Electronic Medical Record Text

IEEE/ACM Transact on Comput Biology and Inform 2019

The main problem to perform Natural Language Processing in the biomedical domain is the access to clinical texts for non-medical staff, and more accurately for languages other than English. This paper presents a method based on neural networks (GAN + reinforce algorithm) to produce clinical documents in Chinese, for a given disease (either pneumonia or lung cancer). The authors used a corpus of 2,216 clinical notes written in Chinese, using the 'History of Present Illness' section as input and the 'Admission Diagnosis' section as tags. The authors report an accuracy of 0.7635 for generated data. They also defined three types of errors in their generated content: repetitions, inconsistent content ("temperature of 39.5°C; no fever"), and improper word matching ("body temperature paroxysmal cough").

Lee J, Yoon W, Kim S, Kim D, Kim S, Ho So C, Kang J

BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Bioinformatics 2019;36(4):1234-40

Current NLP methods rely on word representations to improve results, among which BERT is the most commonly used resource. Nevertheless, while general resources exist, a domain-specific language needs specific resources. This paper introduces BioBERT, a BERT model tuned for the biomedical domain. In order to produce this model, the authors used several corpora in English (Wikipedia, BooksCorpus, PubMed abstracts, and PMC full texts). They compared results achieved by the BioBERT model with the BERT general model on three tasks (named entity recognition, relation extraction, and question-answering). For each task, better results were achieved when using the BioBERT model.

Rosemblat G, Fiszman M, Shin D, Kılıçoğlu H

Towards a characterization of apparent contradictions in the biomedical literature using context analysis

J Biomed Inform 2019;98:103275

This paper aims at identifying contradictions in scientific papers. The authors defined five categories of contradictions: (a) internal to patient, such as comorbidities, (b) external to patient, such as dosage, (c) endogenous and exogenous, (d) known controversy, and (e) contradictions in literature. They used the SemRep tool to identify relationships between 20 common diseases and pathologies, or sign or symptoms. Then, they assessed the level of certainty based on the SemMedDB repository (from PubMed) which contains subject-relation-object predications. On 117,000 instances (from 62,000 abstracts), they identified 2,236 apparent contradictions, among which 58 contradictions were real ones.