Appendix: Content Summaries of Best Papers for the 2018 IMIA Yearbook, Section Clinical Natural Language Processing

Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, Jacobson RS

Automated annotation and classification of BI-RADS assessment from radiology reports

J Biomed Inform 2017 May;69:177-87

This paper presents a system for automatically extracting and classifying mammography information from radiology reports written in English. This is a well conducted study comparing rule-based and machine learning methods for BI-RADS (Breast Imaging Reporting and Data System) categories extraction from breast radiology reports (Conditional Random Field, F=0.95), together with their laterality (Partial decision trees, F=0.91-0.93). It can be noted that the study addresses types of clinical texts and entities that are less challenging than others. However, the corpus offers high text variety with reports from 18 hospitals. While the authors are neither experts on rule-based or machine learning methods, they present an excellent report of their work from the application point of view: describing the caveats of reproducing or adapting previous work, and using toolkits at their disposal towards the targeted application goal.

Pérez A, Weegar R, Casillas A, Gojenola K, Oronoz M, Dalianis H

Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora

J Biomed Inform 2017 Jul;71:16-30

This paper addresses named entity extraction from clinical corpora in Swedish and Spanish. It studies the influence of two types of unsupervised word representations on clinical information extraction performance: word embeddings as obtained by the word2vec method, clustered using K-means, and Brown clusters. The authors go beyond reporting experiments on two languages other than English and also offer methodological insight on the contribution of different classifiers and unsupervised features when little training data is available. The study is original in comparing the same set of configurations based on ensembles of word representations in both corpora, although different types of entities are annotated in the two languages (Drug/Diseases for Spanish, and Body Part/Disorder/Finding for Swedish).

Tapi Nzali MD, Bringay S, Lavergne C, Mollevi C, Opitz T

What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer

JMIR Med Inform 2017 Jul 31;5(3):e23

This study reports on social media analysis relying on two corpora on the topic of breast cancer in French. This paper presents a strong use case and good explanations of why the work is significant from multiple points of view: social media, quality of life in cancer patients, and clinical questionnaire development. The authors use Latent Dirichlet Analysis (LDA) to detect the topics in breast cancer messages from Facebook groups and in forum posts. After a balanced analysis of the LDA results, the automatically identified topics are aligned with internationally standardized self-administered questionnaires used in cancer clinical trials in order to validate the results and identify gaps in the questionnaires. The study draws conclusions that may bring an impact on the maintenance of the international questionnaires.