Machine Learning for Healthcare 2022 – Clinical Abstract, Software, and Demo Track

Structured Understanding of Assessment and Plans in Clinical Documentation

Doron Stupp, MD-PhD, Ronnie Barequet, BSc, I-Ching Lee, BSc, Eyal Oren, PhD, Amir Feder, PhD, Ayelet Benjamini, PhD, Avinatan Hassidim, PhD, Yossi Matias, PhD, Eran Ofek, PhD and Alvin Rajkomar, MD

Google Research

Background.

Physicians record their thought-processes about diagnoses and treatments as unstructured text in a section of a clinical note called the assessment and plan (A&P). Extracting those core concepts as structured data can allow a clearer bird-eye view of, for example, when treatments were started and stopped. The A&P is commonly written as a loosely structured problem-oriented list of conditions (e.g. "rheumatoid arthritis"), their assessments (e.g. "a new flare"), and plans (e.g. "CT of the neck and initiate treatment"). The text itself is free-form and written in physician- speciality-, and organizationally-idiosyncratic ways, making it hard to algorithmically parse into structured data.

Methods.

We modeled the task of identifying the spans of active problem titles, descriptions and action items as a sequence tagging task. We used a 2-layer bidirectional long-short term memory network (LSTM) with a conditional random field (CRF) prediction head. For each token, the model predicted its span type (problem title, problem description and action item) and action item subtype (e.g. medication, lab test) if the span was tagged as an action type. Since human-labeled data is expensive to obtain in the clinical domain, we sought to assess if harnessing domain knowledge could lower data requirements and provide a boost in performance. Specifically, we focused on weak supervision using a clinically-driven heuristic and data augmentations that leverage how assessments and plans are written. The heuristic was implemented using regular expressions to capture a bulleted list of active problems followed by a bulleted list action items. Data augmentations were implemented by algorithmically restructuring labeled assessment and plan sections to contain multiple action items on the same line, action items interleaved with descriptions and mixed bulleting.

Results.

We collected over 30,000 clinician annotations for the A&P sections of 579 notes from the publicly available, de-identified, MIMIC-III ICU dataset. Inter-rater agreement of the human raters measured as the micro average Jaccard similarity across classes was 0.77 (95% bootstrapped confidence intervals 0.75-0.79) for span type and 0.62 (0.6-0.64) for action item type on the held out test set of 48 notes. Evaluation results on the test set as micro-average F1 score for the span type/action item type were 0.91 (0.9-0.92) / 0.84 (0.82-0.86) for the median rater and 0.88 (0.86-0.9) / 0.77 (0.74-0.8) for the best performing model, using both data augmentations and weak supervision. The heuristic achieved an F1 of 0.67 (0.61-0.71) for span type. Training using both weak supervision on clinically-driven heuristics and data-augmentations with 25 notes achieved a micro-average F1 score of 0.85 for the span type, similar to a model trained solely on human rated data with the full training set of 481 notes. These additions also improved model generalizability between medical and surgical ICU wards when compared with models trained on human rated data alone, improving the F1 score from 0.82 to 0.87 for generalizing from medical to surgical and 0.72 to 0.83 from surgical to medical.

Conclusion.

We demonstrate a method to scalably and accurately parse unstructured clinician assessment and plan sections into structured data-elements. We present models and an associated clinician-annotated dataset for structuring assessment and plan sections in clinician notes. We show that excellent performance can be achieved with a limited number of human-labeled notes, with performance maintained across departments, by incorporation of domain expertise in the form of weak supervision with clinically-driven heuristics and curated domain-specific data augmentations.
Our full paper is available at: https://www.medrxiv.org/content/10.1101/2022.04.13.22273438. The annotated dataset is available at: https://zenodo.org/record/6413405.