Clinical Natural Language Technology for Health Care: Past, Present & Future Approaches

Medical records first started to make the transition from paper to electronic in the early 1990’s. Initially, electronic health records (EHRs) were used as a complement to paper records. With increasingly cheaper and more advanced technologies and electronics, medical records have become more electronic and used for not just physician’s notes, but orders, test results, medication lists, diagnoses and so forth. EHRs are now developed and used widely for a variety of reasons: clinical decision support (CDS), improved healthcare processes, clinical patient identification and more (Evans). New technologies-specifically EHRs- are also essential to fulfilling The Quintuple Aim (Nundy et al.; Pagliaroli). Interoperability and standards advancements only solidify EHR’s essential position.

EHRs are composed of both structured and unstructured data. Structured data are easy to implement in automated pipelines. Unstructured data, such as clinical notes, are difficult to use, but essential, if the full potential of EHRs are to be seen. Ford et al. discuss how coded data, an example of structured data, can be an important and effective, yet insufficient, tool in creating disease registries if coded data is missing and diagnostic information is included as free-text (Ford et al.). This is the case with urinary incontinence, a patient-centered outcome post prostate cancer treatment which is rarely coded as an ICD 9 code and rather reported as free text (Hernandez-Boussard et al.). Another paper shares how unstructured clinical narratives are essential to “solving 59% of the CLL trial criteria and 77% of the prostate cancer trial criteria” for patient recruitment in a clinical trial (Raghavan et al.). Two other papers investigated medication discrepancies and found variation between the formal medication list and progress notes/free text narratives (Ashfaq et al.; Walsh et al.). In short, the clinical narratives found in EHRs can provide a vast and useful resource for not just filling in the holes that the structured data provides, but perhaps can aid in finding errors, making diagnoses, classify patients into DRGs, aid in clinical studies and so forth.

Natural language processing (NLP) is the application of computer programs to analyze text. With the increase of EHR usage, standards and interoperability, NLP can be used to utilize EHRs to their fullest extent. Information retrieval, classification, text extraction, text summarization, text generation are some of the ways NLP can be used in a general setting, as well as in a clinical setting. NLP techniques are applied generally through rules-based and machine learning based approaches (Linna and Kahn).

Rules based approaches have been the predominant approach to applying NLP, which is the process of using pre-defined rules such as regular expressions. There are several libraries algorithms that assist in applying rules-based NLP approaches: NLTK and spaCy (open-source libraries for tokenizing, POS tagging, vectorizations, providing several corpuses, etc.), NegEx for negated findings (Chapman et al.) and others.

Classical machine learning (ML) approaches are effective but usually require training data. Popular ML models used in NLP applications include support vector machines (SVM), random forests, generalized linear models (GLM), latent Dirichlet allocation (LDA) and other Bayesian networks, k-nearest neighbors.

Other statistical techniques include word vectorizations, such as term frequency-inverse document frequency (TF-IDF). Usually, several techniques or models will be used in a pipeline as each method performs a specific function. For example, usually a tokenizer is used on some text, then TF-IDF can be performed on the text to vectorize it, which in turn is fed into a linear regression model that predicts the response.

Deep learning models, a subset of machine learning utilizing neural networks, have also been developed to perform various functions. Some of these popular models include recurrent neural networks (RNN), long term short term memory (LSTM), bidirectional encoder representations from transformers (BERT), gated recurrent unit (GRU), Word2vec.

Kreimeyer et al. provides a systematic review of NLP systems used in clinical practice in 2017 (Kreimeyer et al.). 71 systems were identified and information on these systems was explored. Most of these systems followed a rule-based or rule-based/ML hybrid approach. In recent years, deep learning has become more popular and “demonstrated superior performance to classic machine learning” (Linna and Kahn). I think this upward trend for deep learning will only continue to grow, especially as the popularity of large language models (LLMs) increases. As stated previously, machine learning requires large amounts of data for different training, testing and validation phases. To train, test and further tweak an LLM from scratch is a very intensive project to undergo. There are LLM providers (both proprietary and open source) that can be used. Some proprietary LLM providers include: OpenAI, co:here, AI21 Labs,and open-source providers: BigScience, Meta AI, Google. These options allow you to use LLM without having to train and develop your own model. These models are trained on an extensive amount of text and can perform very well in situations without much or any further training. For example, Kung et al. evaluated the performance of ChatGPT, OpenAI’s flagship model, on the United States Medical Licensing Exam (USMLE) without any extra training and “performed at or near the passing threshold” for the exam (Kung et al.). This usage of an LLM on a specific problem without further questioning or training is called zero-shot. If further performance is desired, the whole model or the top few layers of the model can be further trained on a specialized corpus of text, such as EHRs. Other methods to improve the specific accuracy of a LLM in a specific application include few-shot and prompt engineering: guiding the LLM to learn a specific pattern or behavior based on some examples. Few-shot can be very beneficial as you don’t need an extensive dataset to train the model on.

Some limitations to LLMs include what model to choose from, including the data they were trained on, hallucinations – a behavior that LLMs exhibit that when given a question they don’t know the answer to they will form a reasonable-enough guess, even if completely wrong. Proprietary vs. open-source models really come down to a tradeoff between cost and performance. And hallucinations may only improve with time or with better prompt engineering. Perhaps they can be trained in doing better in not hallucinating.

References

Ashfaq, Hamza A., et al. “Medication Accuracy in Electronic Health Records for Microbial Keratitis.” *JAMA Ophthalmology*, vol. 137, no. 8, JAMA Ophthalmol, Aug. 2019, pp. 929–31, doi:10.1001/JAMAOPHTHALMOL.2019.1444.

Chapman, Wendy W., et al. “A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries.” *Journal of Biomedical Informatics*, vol. 34, no. 5, Academic Press, Oct. 2001, pp. 301–10, doi:10.1006/JBIN.2001.1029.

Evans, R. S. *Electronic Health Records: Then, Now, and in the Future*. 2016, doi:10.15265/IYS-2016-s006.

Ford, Elizabeth, et al. “Optimising the Use of Electronic Health Records to Estimate the Incidence of Rheumatoid Arthritis in Primary Care: What Information Is Hidden in Free Text?” *BMC Medical Research Methodology*, vol. 13, no. 1, BioMed Central, 2013, p. 105, doi:10.1186/1471-2288-13-105.

Hernandez-Boussard, Tina, et al. “New Paradigms for Patient-Centered Outcomes Research in Electronic Medical Records: An Example of Detecting Urinary Incontinence Following Prostatectomy.” *EGEMs*, vol. 4, no. 3, Ubiquity Press, May 2016, p. 1, doi:10.13063/2327-9214.1231.

Kreimeyer, Kory, et al. “Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review.” *Journal of Biomedical Informatics*, vol. 73, Academic Press, Sept. 2017, pp. 14–29, doi:10.1016/J.JBI.2017.07.012.

Kung, Tiffany H., et al. “Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.” *PLOS Digital Health*, vol. 2, no. 2, PLOS, Feb. 2023, p. e0000198, doi:10.1371/JOURNAL.PDIG.0000198.

Linna, Nathaniel, and Charles E. Kahn. “Applications of Natural Language Processing in Radiology: A Systematic Review.” *International Journal of Medical Informatics*, vol. 163, Elsevier, July 2022, p. 104779, doi:10.1016/J.IJMEDINF.2022.104779.

Nundy, Shantanu, et al. “The Quintuple Aim for Health Care Improvement: A New Imperative to Advance Health Equity.” *JAMA*, vol. 327, no. 6, American Medical Association, Feb. 2022, pp. 521–22, doi:10.1001/JAMA.2021.25181.

Pagliaroli, Sonia. *How Can Technology Help Achieve the Quintuple Aim?* https://www.cerner.com/ca/en/news/how-can-technology-help-achieve-the-quintuple-aim. Accessed 23 May 2023.

Raghavan, Preethi, et al. “How Essential Are Unstructured Clinical Narratives and Information Fusion to Clinical Trial Recruitment?” *AMIA Summits on Translational Science Proceedings*, vol. 2014, American Medical Informatics Association, 2014, p. 218, /pmc/articles/PMC4333685/.

Walsh, Kathleen E., et al. “Accuracy of the Medication List in the Electronic Health Record-Implications for Care, Research, and Improvement.” *Journal of the American Medical Informatics Association : JAMIA*, vol. 25, no. 7, J Am Med Inform Assoc, July 2018, pp. 909–12, doi:10.1093/JAMIA/OCY027.