6.S897/HST.956 Machine Learning for Healthcare

Lecture 8: Clinical Text, Part 2

Instructors: David Sontag, Peter Szolovits

1 NLP for Healthcare Data

Much of the work in clinical NLP is dependent on identifying important phrases as features and searching for them in large datasets. We elaborate on several studies which have made use of this technique.

1.1 Electronic Medical Record Phenotyping using Anchor and Learn Framework [PNI+18]

Overall goal: Predict patient phenotypes from clinical notes.

Current method: Doctors and domain experts provide words linked to certain disease phenotypes; NLP methods are used to parse large datasets to find matches.

Problem: It's a lot of work to get labels from domain experts, and it would require regular manual up-dates. There are thousands of phenotypes of interest and the entire corpus of the English language could be used as labels.

Idea: Use a small number of labels from domain experts, along with a large dataset to augment the feature set.

Methods:

- Data
  - 273,174 patient records
  - 2008-2013 emergency department (ED) patient records
- Anchor words
  - Features identified by doctors, most strongly identified with each phenotype
  - Identified with high positive predictive value, but not necessarily sensitivity
- Model
  - L2-regularized Logistic Regression model to predict if an anchor word is present in the clinical notes of each patient
  - Binning continuous variables using breaks found on a decision tree
  - Narratives represented as bag-of-words + significant bigrams after negation detection
  - Strategy: censor text within 3 words of anchor to avoid dependence on surrounding context

Results: By training on anchors, we can learn new terms which function as predictive features. The authors published a publicly accessible phenotype library, and the features are much faster and easier to build than manual models.

1.2 De-identification

Overall goal: De-identify patient data from corpus of clinical data.

Current method: At the time, most de-identification was done by hand.

Problem: The rise of EHRs and ML have brought large natural language datasets, which need to be de-identified for distribution and widespread use.

Idea: Use support vectors and local context to de-identify medical discharge summaries.

Methods:

- Data
  - 5 corpora from Partners Healthcare facilities
  - 1000 patient discharge summaries
- Model
  - Use SVM as a classifier for PHI
  - Lexical and Orthographic Features: Target, Lexical Bigrams, Capitalization, Punctuation, Num-bers, Word Length
  - Syntactic Features: Part of Speech, Syntactic Bigrams
  - Semantic Features: censor text within 3 words of anchor to avoid dependence on surrounding context
  -
- Feature Importance
  - Target words
  - Syntactic bigrams
  - Lexical bigrams
  - Part of Speech
  - Dictionary-based features
  - MeSH Features
  - Orthographic features

Results: The model achieved a 97% F-score in de-identifying patient data compared to a rules-based method (85% F-score) on five representative corpora. [USLS08]

1.3 Predicting early psychiatric readmission with natural language processing of narrative discharge summaries [RGN+16]

Overall goal: Predict whether or not a patient will be readmitted to a psychiactric ward within 30 days.

Problem: Patient readmission is quite difficult to predict; even psychiatrists are barely better than random chance when trying to predict 30-day readmission.

Idea: Train an SVM using text-derived features.

Methods:

- Data
  - 4687 patient records from one psychiatric ward, 1994-2012
  - All patients were originally admitted for major depression
  - Only 1240 patients were not readmitted within 30 days
  - Most readmitted patients were later not admitted for major depression
- Model
  - High-Dimensional SVM
  - Baseline features: age, gender, use of public health insurance, Charlson comorbidity index
  - Latent Dirichlet Allocation (LDA) on clinical notes to get 75 common topics; top n words were used as features

Results: AUC of predictive model on the order of 0.7, which is typically not considered very high. However, it still provides useful information, especially when compared to existing standard.

1.4 Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing [MCR$^+$16]

Overall goal: Understand what features of a discharge note narrative contribute to likelihood of suicide or other accidental death.

Problem: Suicide is the tenth leading cause of death in the US across all age groups; being able to pre-dict a patient's actions after being discharged, and intervene appropriately, could save lives.

Idea: Machine learning model using clinical and text-based features

Methods:

- Data
  - All hospital discharges from Mass. General and Brigham and Women's Hospital between 2005 and 2013
  - Patient records and narrative notes for each patient
  - Comparable to previous study, but with many more patients (over 800,000) and less rich data
  - Imbalanced classes: 235 suicides
- NLP
  - Used an open-source Python tool
  - Features drawn from positive and negative valance of words in narrative

Results: Greater positive valance was associated with substantially diminished patient risk of suicide after discharge.

2 Foundations of NLP

Since the development of early text-identification based NLP methods, more sophisticated methods have emerged.

2.1 Tensor Factorization for Unsupervised Exploitation of Text

In this paper by Luo et al. [LXH+15], the authors aim to identify patients with subtypes of lymphoma by analyzing their pathology notes.

Unsupervised Approach: This model is entirely unsupervised. Lymphoma types are clustered according to commonalities in the feature space, and the core clusters are later associated with known lymphoma types. This approach was chosen because of inconsistencies in the field regarding definitions of different types of lymphomas.

2.2 Language modeling

What is language modeling?
The question of how to model language can be broken down into many smaller questions, including how to model syntax or semantics. Current models focus on the following question: *Given a sequence of tokens, what is the next token?*
Most models make the assumption that language tokens can be modeled as nodes in a Markov chain; that is, the value of the next token is only based on the values of the previous n tokens.

Perplexity
When discussing language models, we measure perplexity, an aggregate measure of the complexity of a cor-pus. The entropy of the probability distribution is given by $2^{H(p))}$.
This measure is usually indicative of the necessary complexity of natural language processing models. Medical notes dictated by doctors, for example, use a restricted vocabulary and syntax, and have a perplexity of about 9. In comparison, the perplexity of everyday conversation in English is about 73.

Zipf's Law
Zipf's Law states that the frequency of a word is inversely proportional to its rank when compared to other words. The top ten words in the English language account for 23 percent of the corpus of the English language.

N-Gram Models
Unigram models predict the next token using only the current token. Bigram models use the most recent two tokens; trigram models use the most recent three tokens, etc.
The performance of prediction tasks has been shown to improve significantly as more previous terms are used; this has to do with the fact that some sequences of words are rarely seen. Experiments have been done with a corpus of Shakespeare text, as well as a 1TB large corpus of text published by Google.
Sequence generation is typically attributed to Claude Shannon, and tokens are generated based on the pre-vious n tokens until a stop marker is reached.

2.3 Clinical Applications

2.3.1 Feature Extraction for Phenotpying from Semantic and Knowledge Resources (SEDFE)

Overall Goal: Fully automated and robust unsupervised feature selection method that leverages **only pub-licly available medical knowledge sources**, instead of EHR data. [NCB+19]

Problem: PHI from EHRs and other sensitive data is difficult to obtain, and there is usually not as much available data as in public sources

2.3.2 ANN Model for De-Identification

Overall Goal: Improve existing ML patient de-identification methods using ANN models. [LSDU16]

Problem: HIPAA requirements and PHI advocates require that patient data be de-identified before being distributed for general use. In large datasets, it is not possible for a human to de-identify the data

3 Cutting Edge in NLP

Stepping away from the healthcare context, there are a few trends in NLP that truly define the cutting edge. Three such methods are defined below.

3.1 ELMo: Embeddings from Language Models

ELMo models both complex characteristics of word use and how those uses vary across linguistic polyseminy by using deep bidirectional language models (biLM) pre-trained on large text corpus [PNI+18]. These representations can be added to existing models and significantly improve the state of the art across:

- QuestionAnswering
- Textual Entailment
- Sentiment Analysis

An example of how this method compares to Global Vectors for Word Representations (GloVe) models shows the biLM model properly recognizing context of the word "play" in the below example:

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Figure 1: Comparing sentiment from a GloVe model to that of a biLM model.

3.2 BERT: Bidirectional Encoder Representations from Transformers

Another recent method that is bringing context recognition to new levels is BERT. [PNI+18].

> "Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers."

Because the pre-training in BERT looks at contexts on either side of a word, it has a significantly higher performance on eleven NLP tasks, some by significant margins (BERT improved the GLUE benchmark from **72.8%** to **80.4%**). The model is conceptually simple, pre-training using a transformer, as opposed to OpenAPI GPT and ELMo architectures (as seen in Figure 2).

BERT increases robustness and accuracy by masking some words at each level, or in some cases, inject-ing additional words in to the context. Masking some terms also helps to prevent the model from overfitting, and Peters et al. found that masking 15% of tokens was effective at preventing overfitting. Researchers also used BERT to predict next sentences based on context.

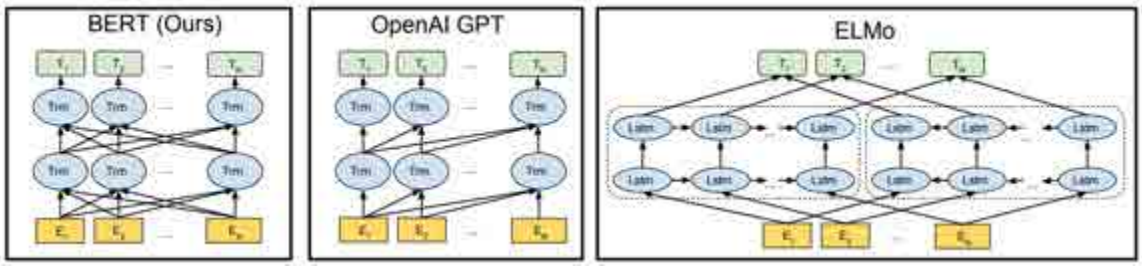

Figure 2: Architecture of BERT compared to OpenAPI GPT (left-to-right transformer) and ELMo (con-catenation of independently trained LSTM)

3.3 GPT-2: General Pre-Training 2

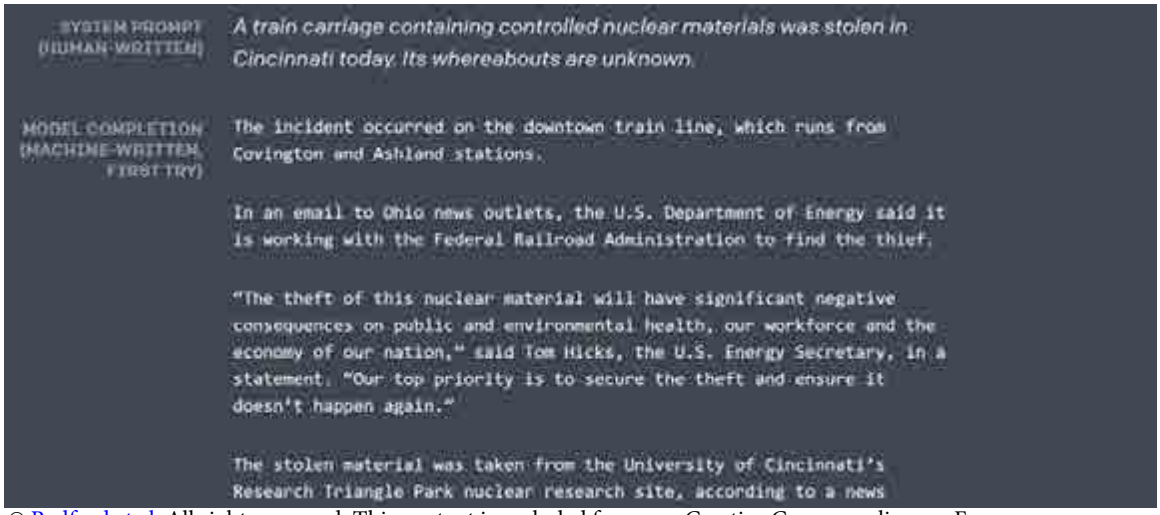

Figure 3: An example human-written (fictional) prompt, and contextual answer given by GPT-2. More examples can be accessed at https://blog.openai.com/better-language-models/

A third exciting model on the cutting edge was developed by OpenAI as a successor to their GPT model.

> "GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset of 8 million web pages.[Ope19]

Given a short human-written system prompt, GPT-2 is able to give a coherent and contextual response. Various responses are given on the OpenAI website. GPT-2 is a unified transformer-based architecture for man tasks. After the GPT model, which was trained on large datasets to complete specific tasks, GPT-2 was created to build a better model by trying to solve many tasks simultaneously, instead of one at a time. Tasks themselves are given as a sequence of tokens:

- *Example:* 'Translate to French', 'English text', 'French Text'
- *Example:* 'Answer the Question', 'document', 'question', 'answer'

References

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.

[HCSH16] Yoni Halpern, Youngduck Choi, David Sontag, and Steven Horng. Electronic medical record phenotyping using the anchor and learn framework. Journal of the American Medical Informatics Association, 23(4):731–740, 04 2016.

[LSDU16] Ji Young Lee, Peter Szolovits, Franck Dernoncourt, and Ozlem Uzuner. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association, 24(3):596–606, 12 2016.

[LXH+15] Yuan Luo, Yu Xin, Ephraim Hochberg, Rohit Joshi, Ozlem Uzuner, and Peter Szolovits. Subgraph augmented non-negative tensor factorization (santf) for modeling clinical narrative text. Journal of the American Medical Informatics Association, 22(5):1009–1019, 2015.

[MCR+16] Thomas H McCoy, Victor M Castro, Ashlee M Roberson, Leslie A Snapper, and Roy H Perlis. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA psychiatry, 73(10):1064–1071, 2016.

[NCB+19] Wenxin Ning, Stephanie Chan, Andrew Beam, Ming Yu, Alon Geva, Katherine Liao, Mary Mullen, Kenneth D. Mandl, Isaac Kohane, Tianxi Cai, and Sheng Yu. Feature extraction for phe-notyping from semantic and knowledge resources. Journal of Biomedical Informatics, 91:103122, 2019.

[Ope19] OpenAI. Better language models and their implications, Feb 2019.

[PNI+18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. CoRR, abs/1802.05365, 2018.

[RGN+16] A Rumshisky, M Ghassemi, T Naumann, P Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational psychiatry, 6(10):e921, 2016.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017.

[USLS08] zlem Uzuner, Tawanda C. Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. Arti*fi*cial Intelligence in Medicine, 42(1):13 – 35, 2008.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019