6.S897/HST.956 Machine Learning for Healthcare

Lecture 7: Natural Language Processing (NLP)

Instructors: David Sontag, Peter Szolovits

1 Outline

This lecture and the next covers the role of Natural Language Processing in machine learning in healthcare. The two lectures in succession first cover methods, which are not based on neural networks representations and then discusses techniques which employ neural network architectures.

We begin by first motivating why we care about clinical text. Later, we discuss some conceptually very appealing, but practically infeasible methods that involve analyzing the narrative texts as linguistic entities in a way that a linguistic might approach them. Next, we discuss what is actually often done e.g. a term spotting approach that says that we might not be able to understand everything that goes on in the narratives, but we can identify certain words/phrases that are highly indicative of whether a certain patient has a certain disease, or a symptom or a medical procedure that was done to them. This is the bread and butter of how clinical research is done nowadays.

2 Value of the data in clinical text

Let's see an example of a discharge summary from MIMIC dataset. The text has been de-identified in the dataset. We know that in MIMIC dataset, we see astericks in places of names, dates, locations etc. Here those entities have been replaced with synthetics names, dates, locations etc. to make it look like a piece that reads like a real text. We want to take advantage of these clinical notes because they carry important information about what happened to the patient over the course of their stay at the hospital.

> *Mr. Blind is a 79-year-old white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center. The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient's hematocrit was stable and he was given no further intervention.*
>
> *The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.*
>
> *On the morning of the 22nd the patient developed tachypnea with a chest X-ray showing a question of congestive heart failure. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given intravenous Lasix.*

Note: orange=demographics; blue=patient condition, diseases, etc.; red=procedures, tests; magenta=results of measurements; yellow=time

In fact to give you a more quantitative version of this, Professor Solovitz and Dr. Katherine worked on a project in 2010 in which they tried to understand what are the genetic correlates of rheumatoid arthritis (RA). In order to do this, they went to Research Patient Data Repository (RPDR) of Massachusetts General and Brigham Partners Healthcare and tried to find the patients who had been billed for rheumatoid arthritis. Naturally, there were thousands of those patients who had been billed for RA. So, they selected a random subset of those patients and gave their records to dermatologists to find out which of those patients actually had rheumatoid arthritis. They found out that the positive predictive value of having a billing code for RA

in this dataset turned out to be really low ($\sim 19\%$)! There is a systematic reason for this as the billing codes were not created to specify what was actually wrong with the patient; instead the billing codes were meant to tell insurance companies/medicare that how much of the payment is reserved for the doctors taking care of them. So the billing codes are very imperfect versions of reality as the billing codes for a patient who is diagnosed with the disease eventually has the same billing codes as the ones for a patient who doesn't eventually get diagnosed with the same disease as the diagnostic procedure is the same! Next, they insisted that instead of just a single billing code for RA, they selected patients from a pool of patients who had three billing codes for RA. This raised the positive predictive value to about 27%. This was again surprising. The reason for such a low PPV even with 3 billing codes was because you can have multiple billing codes during the same visit e.g. one billing code for x-ray, another billing code for blood test for anti-CCP titre. It is entirely possible that all of this is negative and the patient doesn't even have the disease. These aspects are important to consider in clinical data. In order to understand the relation between genetics and the disease, they required a PPV of more than 95% to get a very pure sample of people with RA to get some meaningful results.

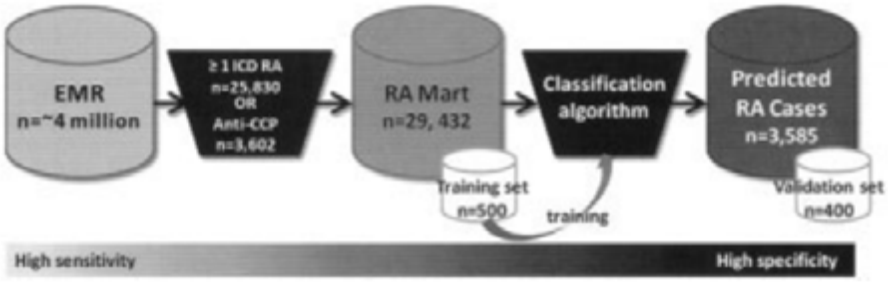


Figure 1: Finding a Cohort of Rheumatoid Arthritis Cases.

The methodology followed by Professor Solovitz is shown in Fig. 1. This was done by first taking 4 million patients in EMR and selecting $\sim 29,000$ patients who had atleast one ICD-9 code for RA, or they had an anti-CCP titre. They selected 500 cases on which they got gold standard readings from rheumatologists and then trained an algorithm to predict whether the patient really had RA or not and that predicted $3,585$ cases out of the $29,432$ had RA. Then, they sampled a validation set of 400 out of those and got those evaluated by rheumatologists to give them gold standard on those! Note that they removed people with ICD-9 codes that fell under the general category of rheumatoid diseases because those people were not appropriate for the data sample they required. They dealt with multiple coding for the same visit by ignoring codes that occurred within a week of each other. They looked for electronic prescriptions of various sorts and lab tests. Counting the number of facts in the database for a patient served as a good proxy for how sick the patient was.

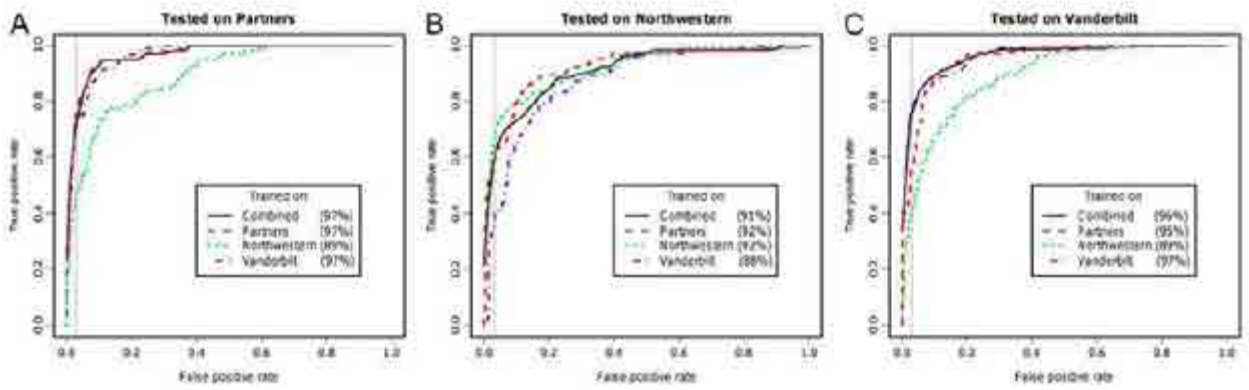
For the narrative text, they used a system called HITEx that extracted entities from narrative text [1]. This was done from health care provider notes, radiology reports, pathology reports, discharge summaries, and operative reports. They also used diagnoses notes, medications, laboratory data and radiology findings. This list from the system was augmented with a hand-curated list of alternative ways of saying the same thing to expand the text. They also ensured that they dealt with negation. The model used was logistic regression. It was interesting to see the positive predictors were a mixture of those dependent on codified data and others dependent on NLP. This work built a compelling reason to show there is real value in narrative text. Using codified data (e.g. lab values, demographics) only to predict whether a patient has rheumatoid arthritis lead to a PPV of 88%. On the other hand, using natural language processing on clinical text (nursing notes, discharge summaries etc.) gave a PPV of 89%. Not surprisingly, a combination of both codified data and NLP gave a PPV of 94% [2].

Another interesting study tried to replicate the results at Vanderbilt and Northwestern [3]. Even though you couldn't run exactly the same methodology based on the fact that all three had different systems.

	Testing set											
	Partners			Northwestern			Vanderbilt			Average		
Algorithm	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC	PPV	Sensitivity	AUC
Published algorithm	88%*	79%*	97%*	87%	60%	92%	95%	57%	95%	90%	65%	95%
Retrained with												
Northwestern	79%	47%	89%	87%	73%	92%	93%	43%	89%	86%	54%	90%
Vanderbilt	85%	74%	97%	82%	40%	88%	97%	81%	97%	88%	65%	94%
Combined	86%	71%	97%	86%	65%	91%	97%	82%	96%	90%	72%	95%
ICD-9 only†												
≥1 RA code	22%	97%	N/A	26%	100%	N/A	49%	100%	N/A	33%	99%	N/A
≥3 RA code	55%	81%	N/A	42%	87%	N/A	73%	98%	N/A	57%	89%	N/A
97% Specificity	80%	49%	88%	80%	36%	84%	93%	43%	93%	84%	43%	88%
Code count for 97% specificity	53			29			48			43.3		

The PPV and sensitivity values reported represent model performance with a specificity set at 97% for logistic regression models.
*These results are from a fivefold cross-validation on the Partners training set. The PPV and sensitivity as published in Liao *et al* was calculated from a separate Partners validation set (PPV 94%, sensitivity 63%).
†ICD-9 cut-off used the count of 714.* codes, excluding codes for juvenile RA (714.3*).
AUC, area under the receiver operating characteristic curve; ICD-9, International Classification of Diseases, version 9 CM; PPV, positive predictive value; RA, rheumatoid arthritis.

(a) Algorithm for RA was portable.



(b) Receiver Operating Characteristic curves for each test set. The vertical line represents the 97% specificity cut-off used in this study. The test performance at Partners, Northwestern, and Vanderbilt are found in (A), (B), and (C) respectively.

Figure 2:

Medication, for example, was extracted from their local EMR in different ways. It was expected that this replication might not produce similar results as it varies how people describe patients in different regions. However, surprisingly, the model performance, despite variation in data representation, was fairly similar as shown in Fig. 2a. Plotting the ROC curves for the three cohorts, as shown in Fig. 2b, demonstrated that training on NorthWestern data and testing on Partners and Vanderbilt was not so great. However, training on Partners and Vanderbilt and testing on Northwestern data turned out to be quite decent. This indicated that there was some generality to the algorithm!

Next, Professor Solovitz showed a sample of terms in a nursing note, which demonstrated the terms was unreadable as English because they were highly abbreviated and shortened. This is shown in Fig. 3. This is still an open research question how to deal with this effectively.

3 Goals of NLP

The typical goals of NLP are described below:

- Assign a meaning (or null) to any word or phrase from some taxonomy/ontology/terminology. For

example, "rheumatoid arthritis" gets codified to 714.0 (ICD-9).

- Determine whether any word or phrase represents protected health information. For example, de-identify "Mr. Huntington suffers from Huntington's Disease" without losing medical information.
- Determine aspects of each entity such as time, location, certainty etc.
- Identify relationships between two meaningful phrases in a sentence, for example precedence, causality, indication etc.
- Identify the sentences or fragments most relevant to answering a specific medical question. For instance, where is the patient's exercise regimen discussed?
- Summarize large corpus of medical text to provide a meaningful overview.

It is important to understand that there are two kinds of tasks. For instance, if you are performing de-identification, you need to look at each word in order to see if it is protective health information. In contrast, the second kind of task requires aggregate judgements where many of the words do not make any difference. For example, one of the challenges the healthcare community working in NLP ran in 2006 gave people medical records and gave them the task of predicting whether the patient is a smoker. In this context, there were obviously words such as "smoker", "tobacco user" that were helpful in making a prediction; however, even these terms were sometimes misleading as a researcher who was working in tobacco mosaic virus got predicted to be a smoker! Another interesting case said that the patient quit smoking two days ago, which is certainly impossible to correctly predict whether that patient was a smoker or not. Similarly, aggregate judgement in process of cohort selection doesn't require you to know everything about a patient but only whether they fit a certain inclusion criteria.

4 Hyper-simplified linguistics

Dr. Thompson, Professor Solovitz's PhD advisor, published an article "English for the Computer" in 1966 [4], which discussed a method to process english. It assumed that there was a grammar and any english text you come across is parsed according to this grammar and each parsing rule corresponds to some semantic function and the picture that emerges is shown in Fig. 4. If you have two phrases with some syntactic

3/11/98 IPN	(date of) Intern Progress Note,
SOB & DOE ↓	the patient's shortness of breath and dyspnea on exertion are decreased,
VSS, AF	the patient's vital signs are stable and the patient is afebrile,
CXR @ LLL ASD no Δ	a recent new chest xray shows a left lower lobe air space density that is unchanged from the previous radiograph,
WBC 11K	a recent new white blood cell count is 11,000 cells per cubic milliliter,
S/B Cx @ GPC c/w PC, no GNR	the patient's sputum and blood cultures are positive for gram positive cocci consistent with pneumococcus, no gram negative rods have grown,
D/C Cef →PCN IV	so the plan is to discontinue the cefazolin and then begin penicillin treatment intravenously.

Figure 3: Sample terms used in a nursing note.

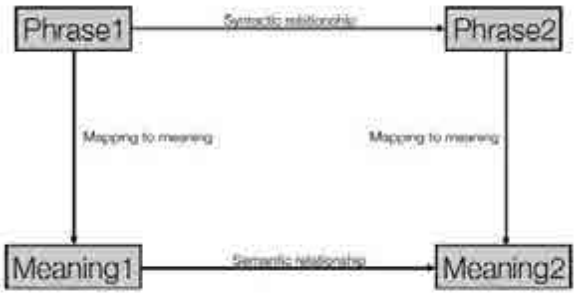


Figure 4: Proposed relationship between syntax and semantics.

relationship between them, then each phrase can be mapped to their meanings and the semantic relationship between those two meanings is determined by the syntactic relationship in the language. Dr. Thompson built computer systems that tried to follow this method and these systems were able to help researchers who worked in areas such as anthropology, where you don't have codified data and a lot of the information is in the form of narrative text. In 1980, Stanford Research Institute built a system called "DIAMOND/DIAGRAM", which intended to help people interact with computer systems when they didn't know command language. So, the people expressed something in english, which got translated to some semantic representation and that was used by the computer. This idea was applied by Walker and Hobbs to natural language access to medical text and they built a system that essentially translated english into some formal representation and process it [5]. The original "DIAMOND/DIAGRAM" system had a very rigid syntax and relied on adaptation of humans. The most radical version of this by the name "French Remede system" was implemented and tested in a medieval hospital in Paris, where an artificial language was developed to take notes about cardiac patients instead of writing them in French [6]. However, it was quickly discarded as doctors reported that the language was not expressive enough.

5 Term spotting and handling negation, uncertainty

Traditionally, term-spotting is done by hand-crafting a list of all the terms that might appear in the note that could be indicative of some condition by a medical practitioner and then the notes are searched through for those terms by the researcher. More sophisticated techniques would use algorithms such as NegEx, which is a negation expression detector, that gets rid of things which not true. This led to more sophisticated machine learning algorithms which aimed to automatically augment the hand-crafted list to create a more indicative list of terms.

For negation, Chapman described a simple algorithm to identify negated findings [7], which found all the UMLS (discussed in next section) terms in each sentence of discharge summary and then searched for two kind of patterns. First pattern looked for a negation phrase such as "no signs of", "ruled out unlikely", "absence of", "not demonstrated", "denies", "no sign of", etc. followed within 5 words by UMLS terms. The second pattern looked for post modifiers such as "declined", "unlikely" etc. Furthermore, they hacked up a bunch of exceptions such as "gram negative", "no further", "not able to be", "not certain if". This algorithm, despite being incredibly simple, does reasonably well as shown in Fig. 5. Comparing with the baseline which looks for negation phrases which are immediately followed by a UMLS term, NegEx significantly improves the specificity from 52.69% to 82.50%.

Generalization is done by taking advantage of related terms, for instance hypo- and hyper-. You could also employ associative reasoning; for example, if you see a lot of symptoms in the clinical text of a particular condition, then the disease is likely to be present as well. The recursive machine learning problem is how best to identify things associated with the term, which is known as "phenotyping".

	Baseline			NegEx		
	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences
n	500	500	1000	500	500	1000
Sensitivity	88.27	0.00	**88.27**	82.31	0.00	77.84
Specificity	52.69	100.00	85.27	82.50	100.00	**94.51**
PPV	68.42	–	68.42	84.49	–	**84.49**
NPV	79.46	96.99	**93.01**	80.21	96.99	91.73

Figure 5: NegEx Results.

6 Unified Medical Language System (UMLS)

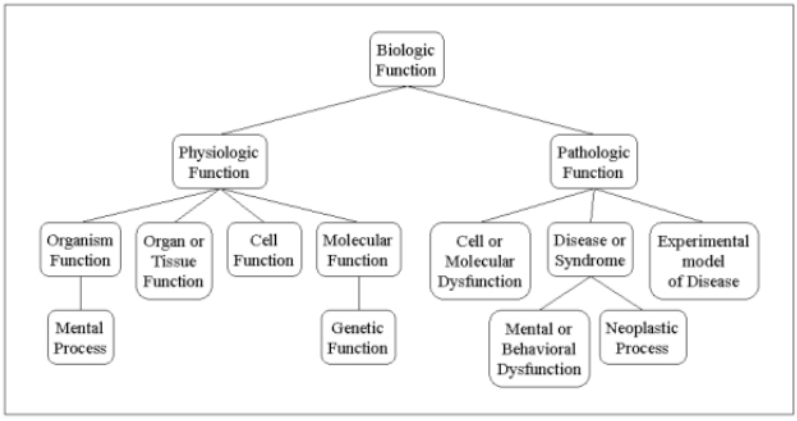
In 1985, National Library of Medicine made a huge effort to create UMLS; this was an attempt to take all of the terminologies that various medical societies had developed and unify them into what they termed as "meta-thesaurus". They also dedicated huge amount of human and machine resources to identify cases in which two different expressions from different terminologies meant the same thing. For instance, heart attack, myocardial infarction, and acute myocardial infarction mean the same. They used the resources to scour the databases and come up with a mapping of each of these terms to a single concept. This is an enormous help to normalize databases that come from different places and are described differently. It also gives you, in the context of natural language processing, a treasure trove of ways of expressing the same conceptual idea and it gives you ways to expand the kind of expressions that you are looking for. There are about 3.7 million concepts in this meta-thesaurus, each of which is assigned a concept unique identifier (CUI). There are also hierarchies and relationships that are imported from all of these different sources of terminology; though these are a jumbled mess. They also created a semantic network of 54 relations and 127 types. Every CUI is assigned at least one semantic type. Examples of UMLS concepts of various types are shown in Fig. 6. The types are hierarchically organized: an example is shown in Fig. 7

There are also tools that deal with some simplistic linguistic problems. For example, "lead", "leads", and "leading" are the same concept. So, there are Lexical Variation Generation (LVG) tools that help you normalize this sort of problem. Similarly, there is a normalization function that helps you normalize sentences into lower-case alphabetized version of the text, e.g. "Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March" is normalized to "acute admit be chest hospital huntington huntington march memorial mr pain". Then, text can translated into other potential linguistic meanings of that text. There is also an online tool available through UMLS Terminology Services, where you can type

```
mysql> select tui,sty,count(*) c from mrsty group by sty
order by c desc;
+------+--------------------------------------------+--------+
| tui  | sty                                        | c      |
+------+--------------------------------------------+--------+
| T061 | Therapeutic or Preventive Procedure        | 260914 |
| T033 | Finding                                    | 233579 |
| T200 | Clinical Drug                              | 172069 |
| T109 | Organic Chemical                           | 167901 |
| T121 | Pharmacologic Substance                    | 124844 |
| T116 | Amino Acid, Peptide, or Protein            | 117508 |
| T009 | Invertebrate                               | 111044 |
| T007 | Bacterium                                  | 110065 |
| T002 | Plant                                      |  95017 |
| T047 | Disease or Syndrome                        |  79370 |
| T023 | Body Part, Organ, or Organ Component       |  73402 |
| T201 | Clinical Attribute                         |  60998 |
| T123 | Biologically Active Substance              |  55741 |
| T074 | Medical Device                             |  51708 |
| T028 | Gene or Genome                             |  49960 |
```

```
select c.cui,c.str from mrconso c join mrsty s on c.cui=s.cui
where c.ts='P' and c.stt='PF' and c.ispref='Y' and
c.lat='ENG' and s.tui='T047';
+----------+-----------------------------------------+
| cui      | str                                     |
+----------+-----------------------------------------+
| C0000744 | Abetalipoproteinemia                    |
| C0000774 | Gastrin secretion abnormality NOS       |
| C0000786 | Spontaneous abortion                    |
| C0000809 | Abortion, Habitual                      |
| C0000814 | Missed abortion                         |
| C0000821 | Threatened abortion                     |
| C0000822 | Abortion, Tubal                         |
| C0000823 | Abortion, Veterinary                    |
| C0000832 | Abruptio Placentae                      |
| C0000880 | Acanthamoeba Keratitis                  |
| C0000889 | Acanthosis Nigricans                    |
| C0001080 | Achondroplasia                          |
| C0001083 | Achromia parasitica                     |
| C0001125 | Acidosis, Lactic                        |
```

Figure 6: Wealth of UMLS Concepts of Various Types.



Courtesy of National Library of Medicine. Image is in the public domain.

Figure 7: Hierarchy of UMLS Semantic Network Types and Relations.

something to get the concept, semantic type etc.

In the next lecture, Professor Solovitz will discuss the advanced machine learning approaches for natural language processing, some of which are based on neural network representations.

References

[1] Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.

[2] Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-Treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E., Plenge, R. (2010). Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research, 62(8), 1120-1127. http://doi.org/10.1002/acr.20184

[3] Carroll, R. J., Thompson, W. K., Eyler, A. E., Mandelin, A. M., Cai, T., Zink, R. M., et al. (2012). Portability of an algorithm to identify rheumatoid arthritis in electronic health records. Journal of the American Medical Informatics Association, 19(e1), e162-9. http://doi.org/10.1136/amiajnl-2011-000583

[4] Frederick B. Thompson, "English for the Computer." Proceedings of the Fall Joint Computer Conference (1966) pp. 349-356

[5] Walker, D. E., Hobbs, J. R., 1981. Natural Language Access to Medical Text*. (pp. 269-273). Presented at the Proc Annu Symp Comput Appl Med Care.

[6] de Heaulme M, Tainturier C, Thomas D. [Computer treatment of medical reports: example of the "Remde" system (author's transl)]. Nouv Presse Med. 1979 Oct 22;8(40):3223-6. French. PubMed PMID: 534182

[7] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019