NLP:
Clinical Natural Language Processing

Feb 25, 2020

Massachusetts Institute of Technology

Outline

- **Value of the data in clinical text**
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- UMLS resources
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

Bulk of Valuable Data are in Narrative Text

orange=demographics
blue=patient condition, diseases, etc.
brown=procedures, tests
magenta=results of measurements
purple=time

Mr. Blind is a 79-year-old white 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.

Selection of Rheumatoid Arthritis Cohort

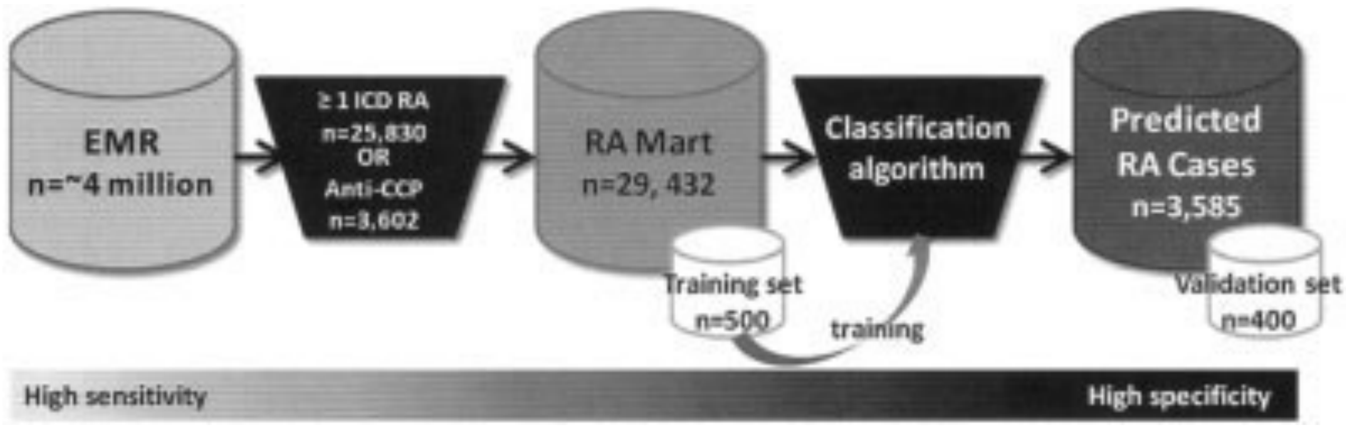
Table 4. Comparison of performance characteristics from validation of the complete classification algorithm (narrative and codified) with algorithms containing codified-only and narrative-only data*

Model	RA by algorithm or criteria, no.	PPV (95% CI), %	Sensitivity (95% CI), %	Difference in PPV (95% CI), %†
Algorithms				
Narrative and codified (complete)	3,585	94 (91–96)	63 (51–75)	Reference
Codified only	3,046	88 (84–92)	51 (42–60)	6 (2–9)‡
NLP only	3,341	89 (86–93)	56 (46–66)	5 (1–8)‡
Published administrative codified criteria				
≥3 ICD-9 RA codes	7,960	56 (47–64)	80 (72–88)	38 (29–47)‡
≥1 ICD-9 RA codes plus ≥1 DMARD	7,799	45 (37–53)	66 (57–76)	49 (40–57)‡

* The complete classification algorithm was also compared with criteria for RA used in published administrative database studies. RA = rheumatoid arthritis; PPV = positive predictive value; 95% CI = 95% confidence interval; NLP = natural language processing; ICD-9 = International Classification of Diseases, Ninth Revision; DMARD = disease-modifying antirheumatic drug.
† Difference in PPV = PPV of complete algorithm − comparison algorithm or criteria.
‡ Significant difference in PPV compared with the complete algorithm.

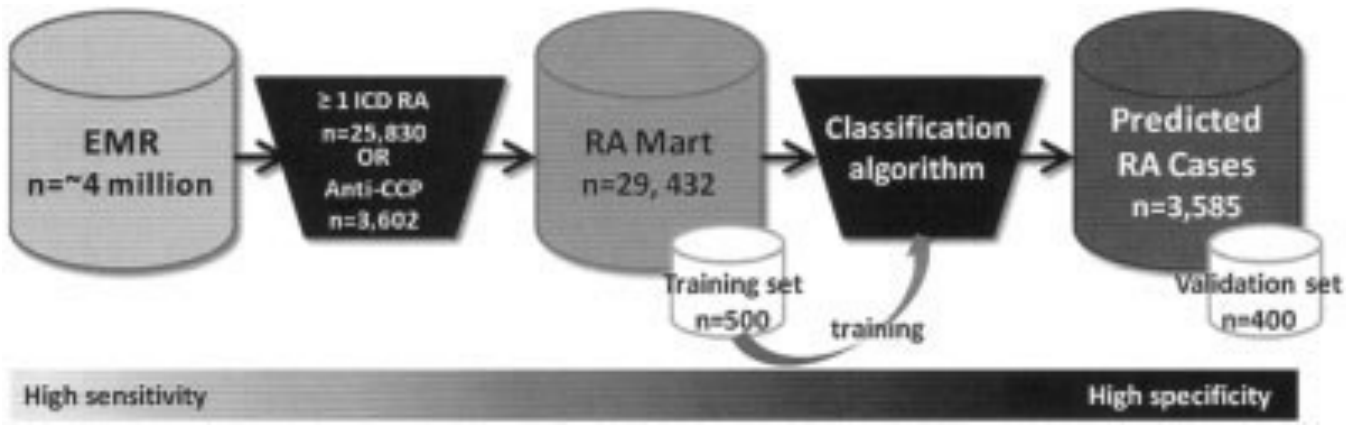
Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-Treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E., Plenge, R. (2010). Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research, 62(8), 1120–1127. http://doi.org/10.1002/acr.20184

Finding a Cohort of Rheumatoid Arthritis Cases



- Coded data:
  - ICD-9 codes, including RA and related diseases
    - ignore codes within 1 week of previous code
  - electronic prescriptions for
    - DMARDs: methotrexate, azathioprine, leflunomide, sulfasalazine, hydroxychloroquine, penicillamine, cyclosporine, and gold
    - Biologic agents: anti-TNF agents infliximab and etanercept, and abatacept, rituximab, anakinra, etc.
  - anti-cyclic citrullinated peptide (anti-CCP) & rheumatoid factor (RF) labs
  - total number of "facts" in the EMR

Finding a Cohort of Rheumatoid Arthritis Cases



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.

- Narrative text data (processed by HITEx)
  - From health care provider notes, radiology reports, pathology reports, discharge summaries, and operative reports
  - Extracted disease diagnoses (RA, SLE, PsA, and JRA)
  - medications (same as from prescriptions, with the addition of adalimumab)
  - laboratory data (RF, anti-CCP, and the term “seropositive”)
  - radiology findings of erosions on radiographs
- Hand-made lists of equivalent terms
- Negation detection, including special terms, e.g., “RF-”

Table 3. Variables selected for the complete algorithm (narrative and codified EMR data) from the logistic regression in order of predictive value*

Variable	Standardized regression coefficient	Standard error
Positive predictors		
NLP RA	1.11	0.48
NLP seropositive	0.74	0.26
ICD-9 RA normalized†	0.71	0.23
ICD-9 RA	0.66	0.44
NLP erosions	0.46	0.29
Codified RF negative	0.36	0.36
NLP methotrexate	0.3	0.34
Codified anti-TNF‡	0.29	0.3
NLP anti-CCP positive	0.27	0.25
NLP anti-TNF§	0.2	0.36
NLP other DMARDs	0.13	0.34
Negative predictors		
ICD-9 JRA	−0.98	0.9
ICD-9 SLE	−0.57	1.09
NLP PsA	−0.51	0.74

* EMR = electronic medical record; NLP = natural language processing; RA = rheumatoid arthritis; ICD-9 = International Classification of Diseases, Ninth Revision; RF = rheumatoid factor; anti-TNF = anti–tumor necrosis factor; anti-CCP = anti–cyclic citrullinated peptide; DMARDs = disease-modifying antirheumatic drugs; JRA = juvenile rheumatoid arthritis; SLE = systemic lupus erythematosus; PsA = psoriatic arthritis.

† ICD-9 RA normalized = ln (no. of ICD-9 RA codes per subject ≥1 week apart).

‡ Codified anti-TNF = etanercept and infliximab (adalimumab was not available in our EMR).

§ NLP anti-TNF = adalimumab, etanercept, and infliximab.

Algorithm for RA was Portable (!)

- Study replicated at Vanderbilt and Northwestern

	Partners	Northwestern	Vanderbilt
EHR	Local	Epic (inpatient) Cerner (outpatient)	Local
# Patients	4M	2.2M	1.7M
Meds	Structured meds entries (in- and outpatient) and text queries	Structured outpatient meds entries and in- and outpatient text queries	NLP (MedEx) for outpatient medications and structured inpatient records
NLP Queries	Custom RegEx	Custom RegEx from Partners	Generic UMLS concepts, derived from KnowledgeMap web interface

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

Table 3 Model performance

Algorithm	Testing set: Partners PPV	Sensitivity	AUC	Northwestern PPV	Sensitivity	AUC	Vanderbilt PPV	Sensitivity	AUC	Average 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.

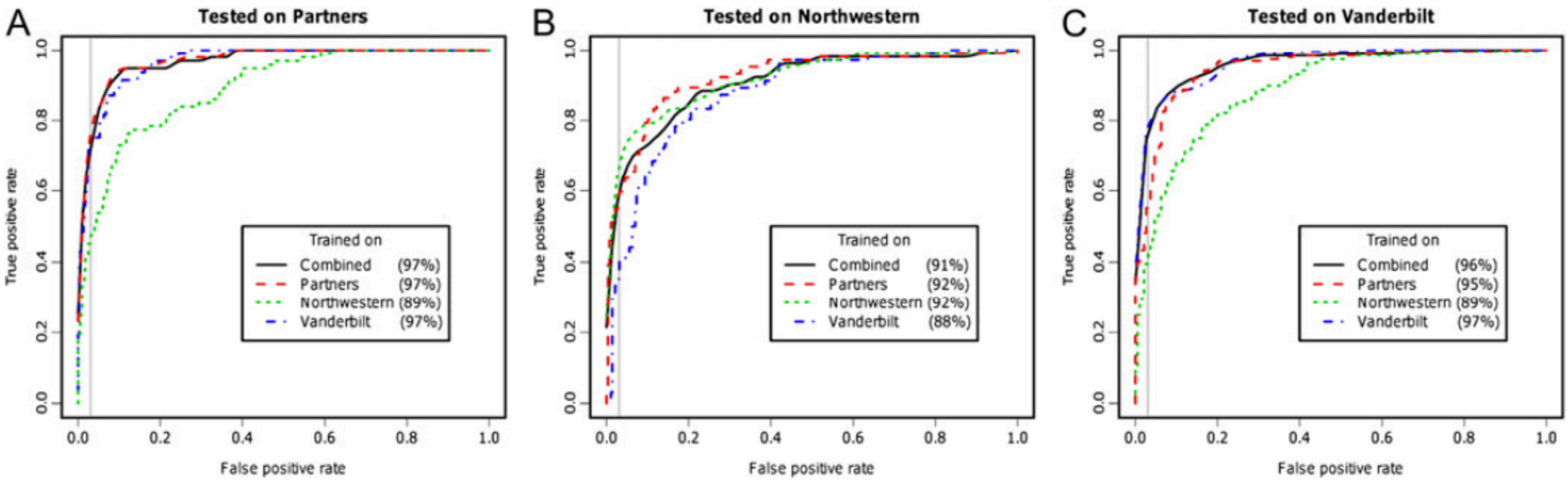


Figure 3 Receiver operating characteristic curves for each test set. The vertical line represents the 97% specificity cut-off used in this study. The test performance at Partners, Northwestern, and Vanderbilt are found in (a), (b), and (c), respectively.

Warning: Telegraphic Language

3/11/98 IPN	
SOB & DOE ↓	
VSS, AF	
CXR ⊕ LLL ASD no Δ	
WBC 11K	
S/B Cx ⊕ GPC c/w PC, no GNR	
D/C Cef →PCN IV	

Barrows, R. C., Jr, Busuioc, M., & Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proceedings / AMIA Annual Symposium AMIA Symposium, 51–55.

Telegraphic Language

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.

Typical Goals of MNLP

- for any word or phrase, assign it a meaning (or null) from some taxonomy/ontology/terminology;
  - e.g., “rheumatoid arthritis” ==> 714.0 (ICD9)
- for any word or phrase, determine whether it represents protected health information;
  - e.g., “Mr. Huntington suffers from Huntington’s Disease”
- determine aspects of each entity: time, location, certainty, ...
- having identified two meaningful phrases in a sentence, determine the relationship (or null) between them;
  - e.g., precedes, causes, treats, prevents, indicates, ...
  - note: we also need a taxonomy of relationships
- in a larger document, identify the sentences or fragments most relevant to answering a specific medical question;
  - e.g., where is the patient’s exercise regimen discussed?
- summarization
  - as data sets balloon in size, how to provide a meaningful overview

Two Types of Tasks

- Every word counts
  - De-identification
  - Extraction of all
    - entities
    - time
    - certainty
    - causation and association

- Aggregate judgment
  - E.g., “smoking” challenge
    - Most text may be irrelevant to specific result
  - Cohort selection—does a patient satisfy some set of inclusion and exclusion criteria
    - Often definite presence of a disease, complication, …

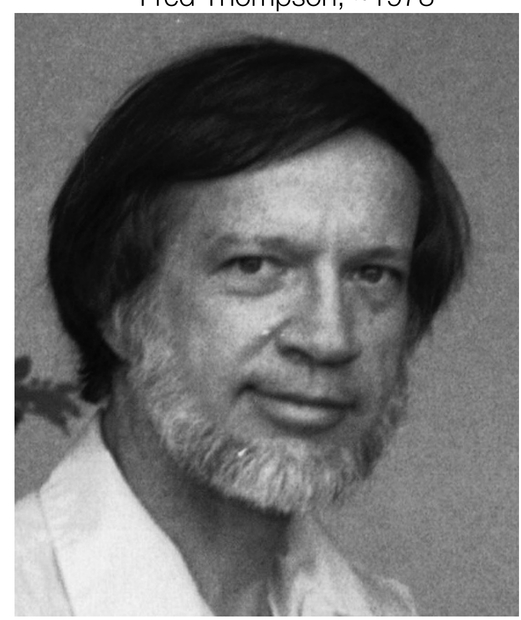
Outline

- Value of the data in clinical text
- **Hyper-simplified linguistics**
- Term spotting + handling negation, uncertainty
- UMLS resources
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

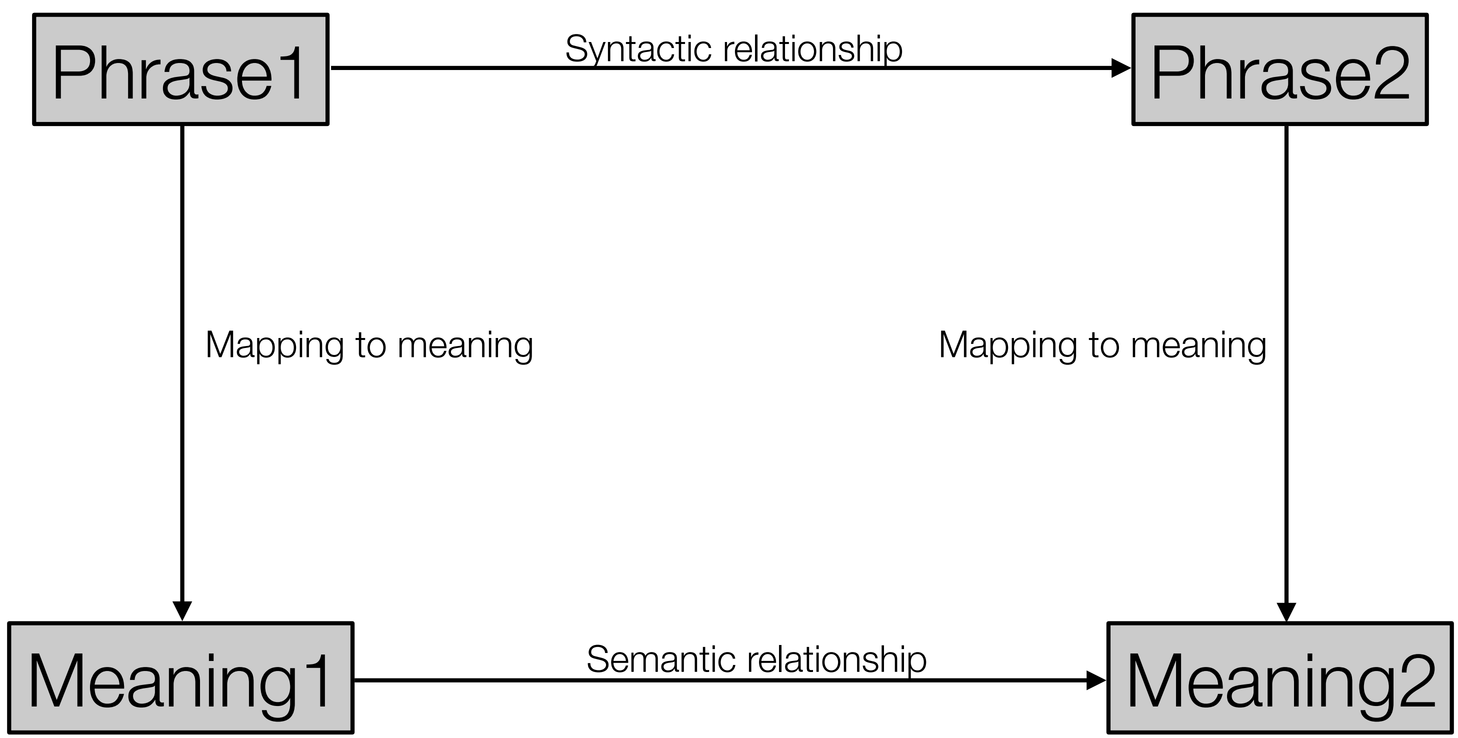
Historical Thought ...

- Frederick B. Thompson, "English for the Computer." *Proceedings of the Fall Joint Computer Conference* (1966) pp. 349-356
- Grammar defined by context-sensitive production rules + transformations
- Semantics defined by mappings:
  - Each grammar rule matches a semantic function
  - Terminal symbols are *referents* or *functions*
  - An environment is (in modern terms) a semantic network of complex interrelationships
  - Meaning is compositional, in terms of the semantic functions
- *Minor* 😆 remaining question: how to represent the "real world"?

Fred Thompson, ~1973



Proposed relationship between syntax and semantics



Formal language semantics

- SRI's DIAMOND/DIAGRAM system (~1980)
- each passage is expressed as a proposition or a conjunction of propositions:
    - a particular procedure for the prevention of hepatitis B could have associated with it the proposition "immunize(GAMMA-GLOBULIN,HEPATITIS-B)"
    - a passage concerned with the etiology of the disease could have the proposition "transmit(TRANSFUSION,HEPATITIS-B)"
    - synonym and hyponym relations
    - … *a language of primitives for the domain*
- French Remède system
    - "medical documentary language using current medical terms and few syntactic rules"
    - taught to doctors to write notes
    - *… not popular*

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.

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

Outline

- Value of the data in clinical text
- Hyper-simplified linguistics
- **Term spotting + handling negation, uncertainty**
- UMLS resources
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

Term Spotting

- Traditionally, lists of coded items, narrative terms and patterns hand-crafted by researcher
- Negation and uncertainty handled by somewhat ad-hoc methods
  - NegEx is widely used, ∃ many more sophisticated variants
- Generalize terms
  - Manually or automatically identify high-certainty “anchors”
  - Learn related terms to augment the set of terms
    - From knowledge bases such as UMLS
    - From co-occurrence in EMR data
    - From co-occurrence in publications
  - C.f. Sontag’s lecture

Negation

- “Identifying pertinent negatives, then, involves identifying a proposition ascribing a clinical condition to a person and determining whether the proposition is denied or negated in the text.”
- Simpler than general problem of negation in NLP because negation applies mostly to noun phrases indicating diseases, tests, drugs, findings, …
- NegEx
    - Find all UMLS terms in each sentence of a discharge summary
        - “The patient denied experiencing chest pain on exertion” ⇒

          “The patient denied experiencing S1459038 on exertion”
    - Find patterns
        - <negation phrase> *{0,5} <UMLS term>
            - "no signs of", "ruled out unlikely", "absence of", "not demonstrated", "denies", "no sign of", "no evidence of", "no", "denied", "without", "negative for", "not", "doubt", "versus"
        - <UMLS term> *{0,5} <negation phrase>
            - “declined”, “unlikely”
    - Pseudo-negation: "no further", "not able to be", "not certain if", "not certain whether", “not necessarily", "not rule out", "without any further", "without difficulty", "without further”, "gram negative"

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.

NegEx results

- Baseline:
  - <negation phrase> * <UMLS term>
    - "no", "denies", "not", "without", "*n't", "ruled out", "denied"

	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

- Extremely simplistic schemes (kind of) work

Generalize Terms

- Use synonymous terms as well as the starting ones
- Take advantage of others related terms
    - hypo- or hypernyms
    - other associated terms
        - e.g., common symptoms or treatments of a disease
- Recursive ML problem: learn how best to identify cases associated with a term
    - “phenotyping”

Outline

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- **UMLS resources**
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- Neural methods

Available Classification Thesauri
Most Available through UMLS

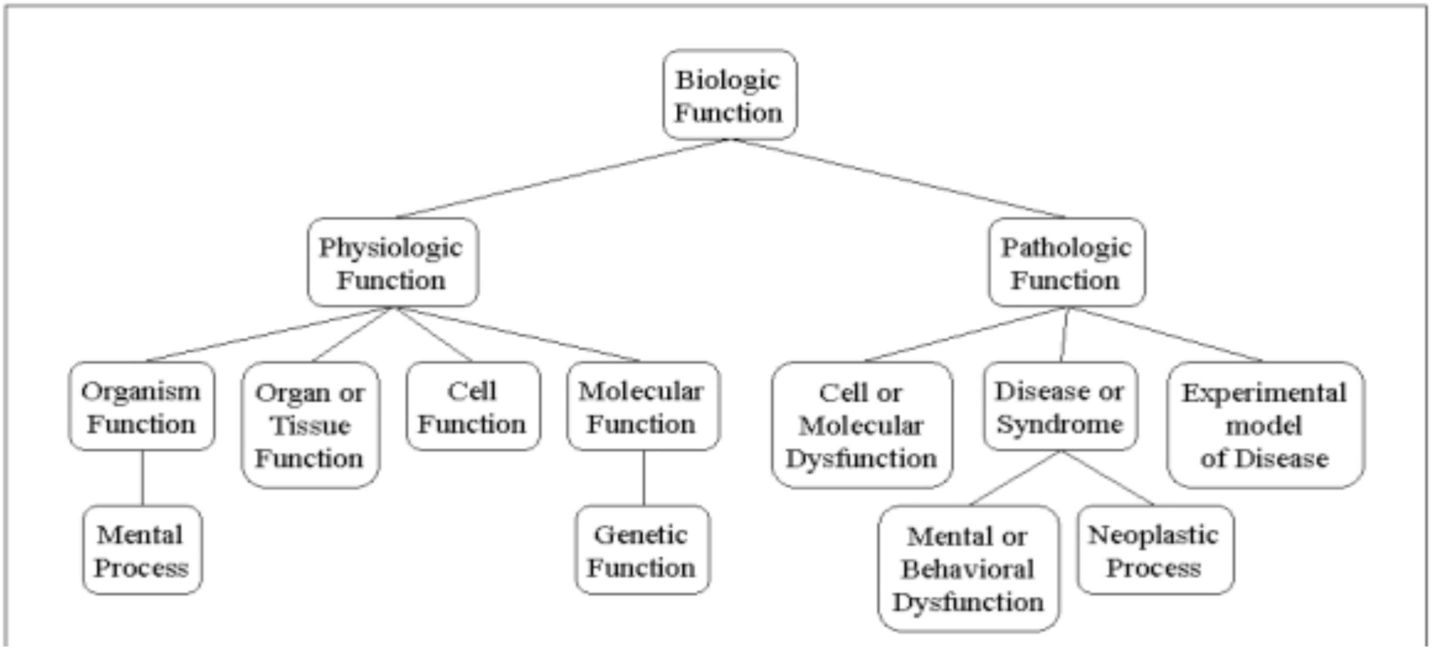
- Unified Medical Language Systems project of NLM; since ~1985
- *Metathesaurus* now (2019ab version) includes 211 source vocabularies
  - MeSH, SNOMED, ICD-9, ICD-10, LOINC, RxNORM, CPT, GO, DXPLAIN, OMIM, ...
- Synonym mappings across vocabularies;
  - e.g., “heart attack” = “acute myocardial infarct” = “myocardial infarction” ...
  - 4,209,309 distinct concepts, represented by concept unique identifier (CUI)
- Jumbled compendium of every hierarchy drawn from every source
- What granularity?
- *Semantic Network*
  - Hierarchy of
    - 54 relations
    - 127 types
  - Every CUI assigned ≥1 semantic type

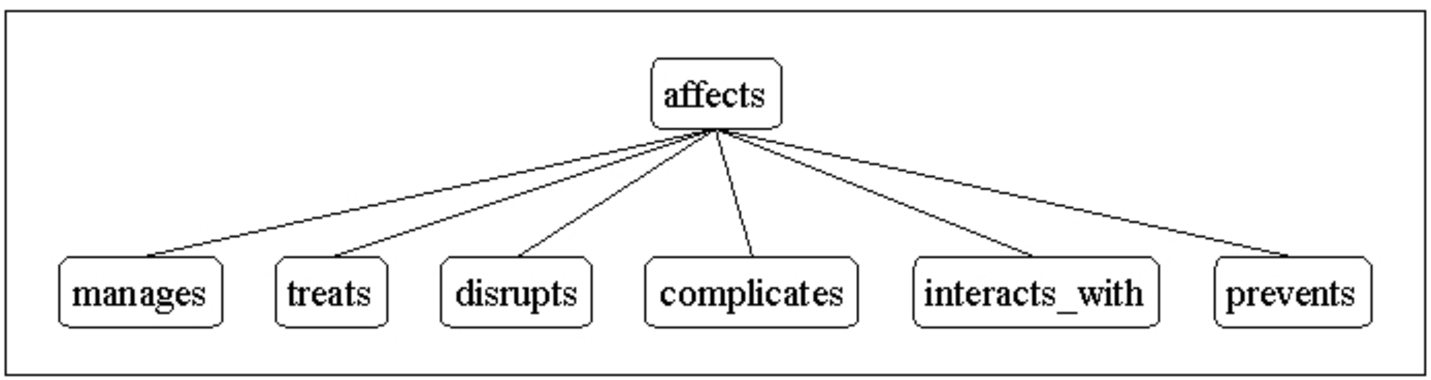
Wealth of UMLS Concepts of Various Types

```
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                         | 157901 |
| 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 |
| T004 | Fungus                                   |  47291 |
| T060 | Diagnostic Procedure                     |  46106 |
| T037 | Injury or Poisoning                      |  43924 |
| T191 | Neoplastic Process                       |  33539 |
| T044 | Molecular Function                       |  31369 |
| T126 | Enzyme                                   |  25766 |
| T129 | Immunologic Factor                       |  25025 |
| T059 | Laboratory Procedure                     |  24511 |
| T058 | Health Care Activity                     |  19552 |
| T029 | Body Location or Region                  |  16470 |
| T013 | Fish                                     |  16059 |
| T046 | Pathologic Function                      |  13562 |
| T184 | Sign or Symptom                          |  13299 |
| T130 | Indicator, Reagent, or Diagnostic Aid    |  12809 |
| T170 | Intellectual Product                     |  12544 |
| T118 | Carbohydrate                             |  10722 |
| T110 | Steroid                                  |  10363 |
| T012 | Bird                                     |   9908 |
| T043 | Cell Function                            |   9758 |
...
```

```
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                         |
| C0001126 | Renal tubular acidosis                   |
| C0001127 | Acidosis, Respiratory                    |
| C0001139 | Acinetobacter Infections                 |
| C0001142 | Acladiosis                               |
| C0001144 | Acne Vulgaris                            |
| C0001145 | Acne Keloid                              |
| C0001163 | Vestibulocochlear Nerve Diseases         |
| C0001168 | Complete obstruction                     |
| C0001169 | Acquired coagulation factor deficiency NOS |
| C0001175 | Acquired Immunodeficiency Syndrome       |
| C0001197 | Acrodermatitis                           |
| C0001202 | Acrokeratosis                            |
| C0001206 | Acromegaly                               |
| C0001207 | Hypersomatotropic gigantism              |
| C0001231 | ACTH Syndrome, Ectopic                   |
| C0001247 | Actinobacillosis                         |
...
```

Hierarchy of UMLS Semantic Network Types and Relations





from http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls&part=ch05

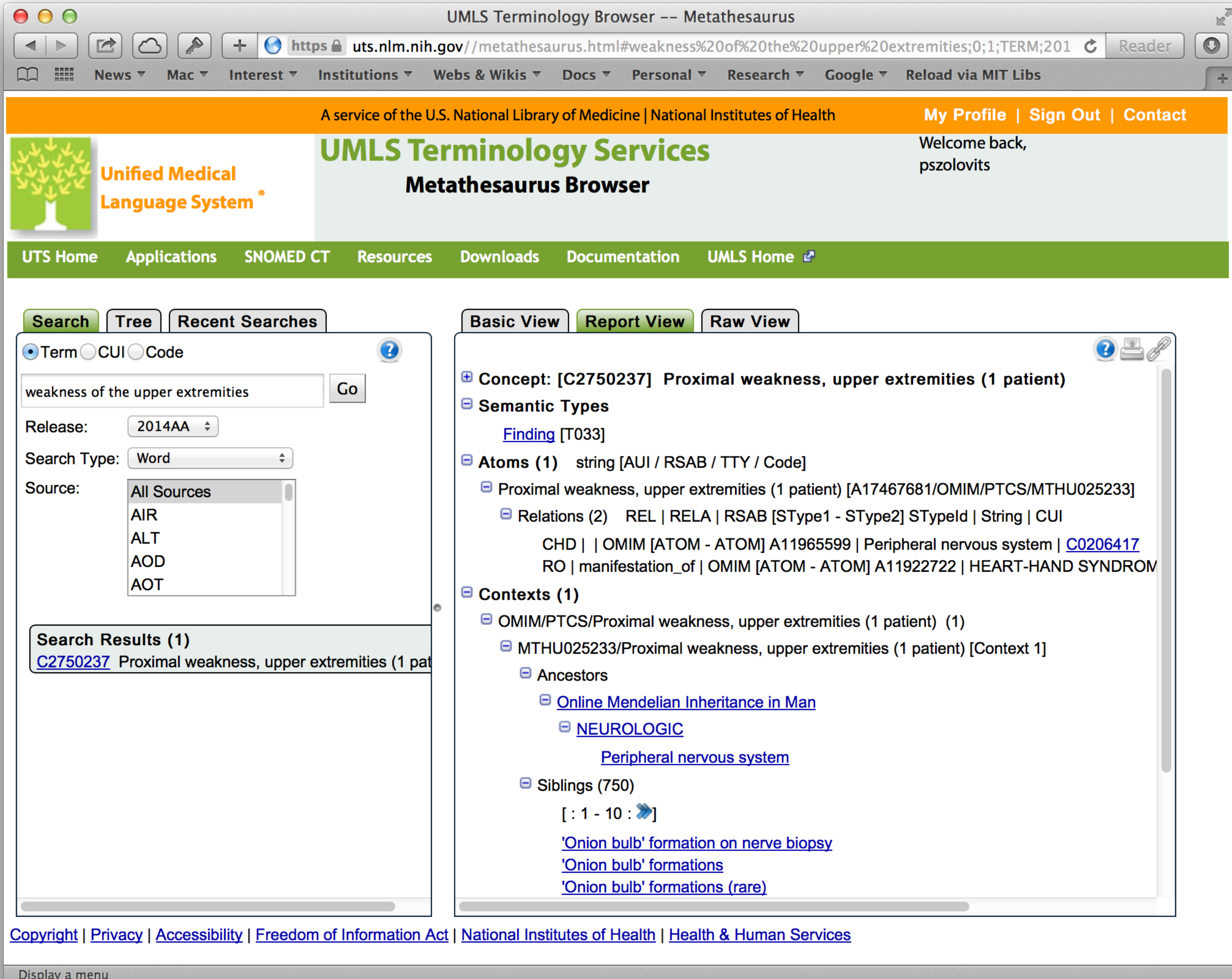
Lexical Variant Generation (LVG) Tools

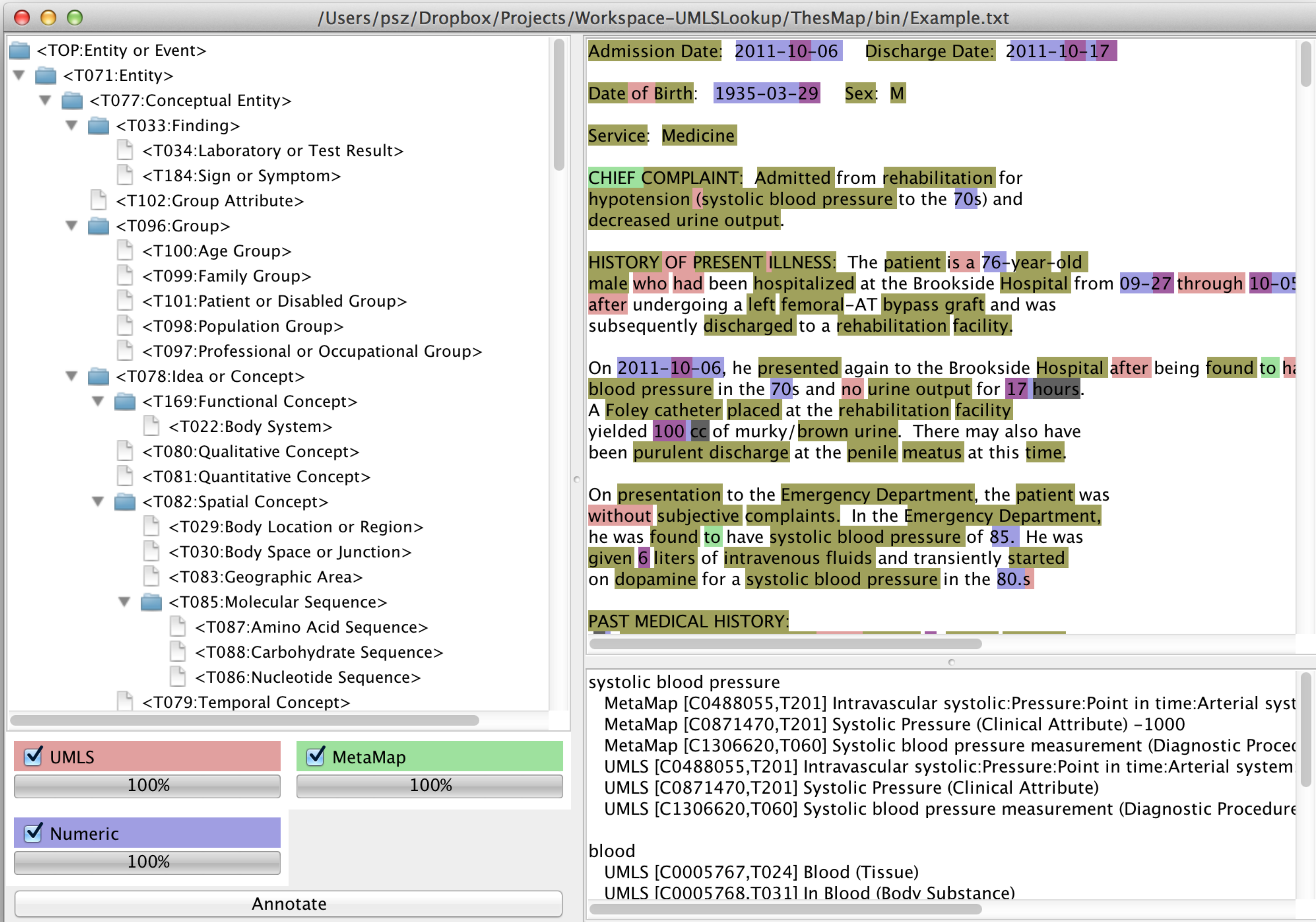
(from National Library of Medicine)

- Normalized words and phrases used as index to UMLS
- Lemmatization of words
  - stripping typical prefixes, suffixes
    - plurals, in-word negation, gerunds
- Discarding “noise” words, punctuation
- Lower-casing
- Alphabetic order of all remaining words

```
Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.
Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admit be chest hospital huntington huntington march
memorial mr pain
Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admit chest hospital huntington huntington march
memorial mr pain was
Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admitted be chest hospital huntington huntington march
memorial mr pain
Mr. Huntington was admitted to Huntington Memorial Hospital for acute chest pain in March.|acute admitted chest hospital huntington huntington march
memorial mr pain was

Weakness of the upper extremities
Weakness of the upper extremities|extremity upper weakness
```


UMLS Terminology Browser -- Metathesaurus
https uts.nlm.nih.gov//metathesaurus.html#weakness%20of%20the%20upper%20extremities;0;1;TERM;201
Reader
News Mac Interest Institutions Webs & Wikis Docs Personal Research Google Reload via MIT Libs
A service of the U.S. National Library of Medicine | National Institutes of Health
My Profile | Sign Out | Contact
UMLS Terminology Services
Metathesaurus Browser
Welcome back, pszolovits
Unified Medical Language System
UTS Home Applications SNOMED CT Resources Downloads Documentation UMLS Home
Search Tree Recent Searches
Term CUI Code
weakness of the upper extremities
Go
Release: 2014AA
Search Type: Word
Source: All Sources AIR ALT AOD AOT
Search Results (1)
C2750237 Proximal weakness, upper extremities (1 pat
Basic View Report View Raw View
Concept: [C2750237] Proximal weakness, upper extremities (1 patient)
Semantic Types
Finding [T033]
Atoms (1) string [AUI / RSAB / TTY / Code]
Proximal weakness, upper extremities (1 patient) [A17467681/OMIM/PTCS/MTHU025233]
Relations (2) REL | RELA | RSAB [SType1 - SType2] STypeId | String | CUI
CHD | | OMIM [ATOM - ATOM] A11965599 | Peripheral nervous system | C0206417
RO | manifestation_of | OMIM [ATOM - ATOM] A11922722 | HEART-HAND SYNDROM
Contexts (1)
OMIM/PTCS/Proximal weakness, upper extremities (1 patient) (1)
MTHU025233/Proximal weakness, upper extremities (1 patient) [Context 1]
Ancestors
Online Mendelian Inheritance in Man
NEUROLOGIC
Peripheral nervous system
Siblings (750)
[: 1 - 10 :]
'Onion bulb' formation on nerve biopsy
'Onion bulb' formations
'Onion bulb' formations (rare)
Copyright | Privacy | Accessibility | Freedom of Information Act | National Institutes of Health | Health & Human Services
Display a menu


/Users/psz/Dropbox/Projects/Workspace-UMLSLookup/ThesMap/bin/Example.txt
<TOP:Entity or Event>
<T071:Entity>
<T077:Conceptual Entity>
<T033:Finding>
<T034:Laboratory or Test Result>
<T184:Sign or Symptom>
<T102:Group Attribute>
<T096:Group>
<T100:Age Group>
<T099:Family Group>
<T101:Patient or Disabled Group>
<T098:Population Group>
<T097:Professional or Occupational Group>
<T078:Idea or Concept>
<T169:Functional Concept>
<T022:Body System>
<T080:Qualitative Concept>
<T081:Quantitative Concept>
<T082:Spatial Concept>
<T029:Body Location or Region>
<T030:Body Space or Junction>
<T083:Geographic Area>
<T085:Molecular Sequence>
<T087:Amino Acid Sequence>
<T088:Carbohydrate Sequence>
<T086:Nucleotide Sequence>
<T079:Temporal Concept>
UMLS
100%
MetaMap
100%
Numeric
100%
Annotate
Admission Date: 2011-10-06 Discharge Date: 2011-10-17
Date of Birth: 1935-03-29 Sex: M
Service: Medicine
CHIEF COMPLAINT: Admitted from rehabilitation for
hypotension (systolic blood pressure to the 70s) and
decreased urine output.
HISTORY OF PRESENT ILLNESS: The patient is a 76-year-old
male who had been hospitalized at the Brookside Hospital from 09-27 through 10-0
after undergoing a left femoral-AT bypass graft and was
subsequently discharged to a rehabilitation facility.
On 2011-10-06, he presented again to the Brookside Hospital after being found to h
blood pressure in the 70s and no urine output for 17 hours.
A Foley catheter placed at the rehabilitation facility
yielded 100 cc of murky/brown urine. There may also have
been purulent discharge at the penile meatus at this time.
On presentation to the Emergency Department, the patient was
without subjective complaints. In the Emergency Department,
he was found to have systolic blood pressure of 85. He was
given 6 liters of intravenous fluids and transiently started
on dopamine for a systolic blood pressure in the 80.s
PAST MEDICAL HISTORY:
systolic blood pressure
MetaMap [C0488055,T201] Intravascular systolic:Pressure:Point in time:Arterial syst
MetaMap [C0871470,T201] Systolic Pressure (Clinical Attribute) -1000
MetaMap [C1306620,T060] Systolic blood pressure measurement (Diagnostic Procec
UMLS [C0488055,T201] Intravascular systolic:Pressure:Point in time:Arterial system
UMLS [C0871470,T201] Systolic Pressure (Clinical Attribute)
UMLS [C1306620,T060] Systolic blood pressure measurement (Diagnostic Procedure
blood
UMLS [C0005767,T024] Blood (Tissue)
UMLS [C0005768,T031] In Blood (Body Substance)

Outline

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- UMLS resources
- ML to expand terms
- **pre-NN ML to identify entities and relations**
- language models
- Neural methods

The Importance of Context

- "Mr. Huntington was treated for Huntington's Disease at Huntington Hospital, located on Huntington Avenue."
  - Huntington
  - Huntington's Disease
  - Mr. Huntington's Disease
- "Atenalol was administered to Mr. Huntington."
  - vs. "Atenalol was considered for control of heart rate."
  - vs. "Atenalol was ineffective and therefore discontinued."

Building Models

- Features of text from which models can be built
  - words, parts of speech, capitalization, punctuation
  - document section, conventional document structures
  - identified patterns and thesaurus terms
  - lexical context
    ➡ all of the above, for n-tuples of words surrounding target
  - syntactic context
    ➡ all of the above, for words syntactically related to target
    - E.g., “The lasix, started yesterday, reduced ascites ...”

```
    +---------------------------------Xp----------------------------------+
    |          +-----------------Ss-----------------+                      |
    |          +----MXsp----+-------Xc-------+      |                      |
    +----Wd----+     +--Xd--+---MVpn---+     |      +-----Os-----+        |
    |          |     |      |          |     |      |            |        |
LEFT-WALL lasix[?].n , started.v-d yesterday , reduced.v-d ascites[?].n .
```

(Output from Link Grammar Parser, w/o special medical dictionary)

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42(1), 13–35. http://doi.org/10.1016/j.artmed.2007.10.001

Parsing Can be Ambiguous

- Prepositional phrase attachment
- Part of speech
  - e.g., white.n vs. white.a
- Hope that there is enough redundancy to overcome such limitations

```
Found 111 linkages (24 with no P.P. violations)
  Linkage 1, cost vector = (UNUSED=0 DIS=0 AND=0 LEN=22)

    +------------------------------------------------Xp-----------------------------------------------+
    +------Wd------+     +--------Ost--------+                                                        |
    |       +---G--+     |  +-------Dsu------+             +---Jp---+      +--------Jp--------+       |
    |       +Xi+   +--Ss-+  |      +----Ah---+---Ma--+-MVp-+  +-Dsu-+--Mp--+      +-----AN----+       |
    |       |  |   |     |  |      |         |       |     |  |     |      |      |           |       |
LEFT-WALL Mr.x . Blind is.v a 79-year-old white.n male.a with a history.n of diabetes.n mellitus[?].n .

Constituent tree:

(S (NP Mr . Blind)
   (VP is
       (NP a 79-year-old white
           (ADJP male
                 (PP with
                     (NP (NP a history)
                         (PP of
                             (NP diabetes mellitus)))))))
   .)
```

Stanford Parser

nlp.stanford.edu:8080/parser/index.jsp

News Mac Interest Institutions Webs & Wikis Docs Personal Research Google

Stanford Parser

Please enter a sentence to be parsed:

The patient experienced weakness of the upper extremities.

Language: English   Sample Sentence   Parse

Your query

The patient experienced weakness of the upper extremities.

Tagging

```
The/DT  patient/NN  experienced/VBD  weakness/NN  of/IN  the/DT  upper/JJ  extremities/NNS  ./.
```

Parse

```
(ROOT
  (S
    (NP (DT The) (NN patient))
    (VP (VBD experienced)
      (NP
        (NP (NN weakness))
        (PP (IN of)
          (NP (DT the) (JJ upper) (NNS extremities)))))
    (. .)))
```

Typed dependencies

```
det(patient-2, The-1)
nsubj(experienced-3, patient-2)
root(ROOT-0, experienced-3)
dobj(experienced-3, weakness-4)
prep(weakness-4, of-5)
det(extremities-8, the-6)
amod(extremities-8, upper-7)
pobj(of-5, extremities-8)
```

Display a menu

Example of Features Available for Model

Mr. Blind is a 79-year-old white white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum

263 266 "Mr."
TUI: T060,T083,T047,T048,T116,T192,T081,T028,T078,T077; SP-POS: noun; SEM: _modifier,_disease,_procparam;
CUI: C0024487,C0024943,C0025235,C0025362,C0026266,C0066563,C0311284,C0475209,C1384671,
C1413973,C1417835,C1996908,C2347167,C2349188; lptok: 6;
MeSH: C07.465.466,C10.292.300.800,C10.597.606.643,C14.280.484.461,C23.888.592.604.646,D12.776.826.750.530,
D12.776.930.682.530,E05.196.867.519,F01.700.687,F03.550.600,Z01.058.290.190.520;
267 468 "Blind is a 79-year-old white white...hsandpot Center." sent: nil;
267 272 "Blind"
TUI: T062,T047,T170; SP-POS: verb,adj,noun; SEM: _disease; CUI: C0150108,C0456909,C1561605,C1561606;
lptok: 1; MeSH: C10.597.751.941.162,C11.966.075,C23.888.592.763.941.162;
273 277 "is a" TUI: T185,T169,T078; SEM: _modifier; CUI: C1278569,C1292718,C1705423;
273 275 "is" SP-POS: aux,noun,adj; lptok: 2;
276 277 "a" SP-POS: det,noun,adj; lptok: 3;
278 289 "79-year-old" lptok: 4;
290 295 "white" TUI: T098,T080; SP-POS: noun,adj; SEM: _modifier; CUI: C0007457,C0043157,C0220938; lptok: 5;
296 301 "white" TUI: T098,T080; SP-POS: noun,adj; SEM: _modifier; CUI: C0007457,C0043157,C0220938; lptok: 6;
302 306 "male"
TUI: T032,T098,T080; SP-POS: adj,noun; SEM: _modifier,_bodyparam;
CUI: C0024554,C0086582,C1706180,C1706428,C1706429; lptok: 7;
307 311 "with" SP-POS: prep,conj; lptok: 8;
312 313 "a" SP-POS: det,noun,adj; lptok: 9;
314 342 "history of diabetes mellitus" TUI: T033; SEM: _finding; CUI: C0455488;

Learning Models

- Given a target classification, build a machine learning model predicting that class
  - support vector machines (SVM)
  - classification trees
  - naive Bayes or Bayesian networks
  - artificial neural networks
  - ...
- class(word) = function (feature$_1$, feature$_2$, feature$_3$, ...)
  - sometimes, astronomically large (binary) feature set; SVM can deal with it
    - f_1 ... $f_{100,000}$: whether the word is “a”, “aback”, “abacus”, ..., “zymotic”
    - $f_{100,001}$ ...: whether word’s POS is “noun”, “verb”, “adj”, ...
    - $f_{100,100}$ ...: whether the word maps to CUI “C0000001”, “C0000002”, ...
    - $f_{3,000,000}$ ...: same as above, but for 1st, 2nd, 3rd word to right/left
    - $f_{6,000,000}$ ...: {lp-link, word} for 1st, 2nd, 3rd link in parse to right/left
    - ...

Using this model for de-identification

Table 6 Evaluation on authentic discharge summaries

Method	Class	Precision (%)	Recall (%)	*F*-measure (%)
Stat De-id	PHI	98.46	95.24	**96.82**
IFinder	PHI	26.17	61.98	36.80*
H + D	PHI	82.67	87.30	84.92*
CRFD	PHI	91.16	84.75	87.83*
Stat De-id	Non-PHI	99.84	99.95	**99.90**
IFinder	Non-PHI	98.68	94.19	96.38*
H + D	Non-PHI	99.58	99.39	99.48*
CRFD	Non-PHI	99.62	99.86	99.74*

The *F*-measure differences from Stat De-id in PHI and in non-PHI are significant at $\alpha = 0.05$.

Table 7 Evaluation of SNoW and Stat De-id on authentic discharge summaries

Method	Class	Precision (%)	Recall (%)	*F*-measure (%)
Stat De-id	PHI	98.40	93.75	**96.02**
SNoW	PHI	96.36	91.03	93.62*
Stat De-id	Non-PHI	99.90	99.98	**99.94**
SNoW	Non-PHI	99.86	99.95	99.90*

The *F*-measure differences from Stat De-id in PHI and in non-PHI are significant at $\alpha = 0.05$.

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42(1), 13–35. http://doi.org/10.1016/j.artmed.2007.10.001

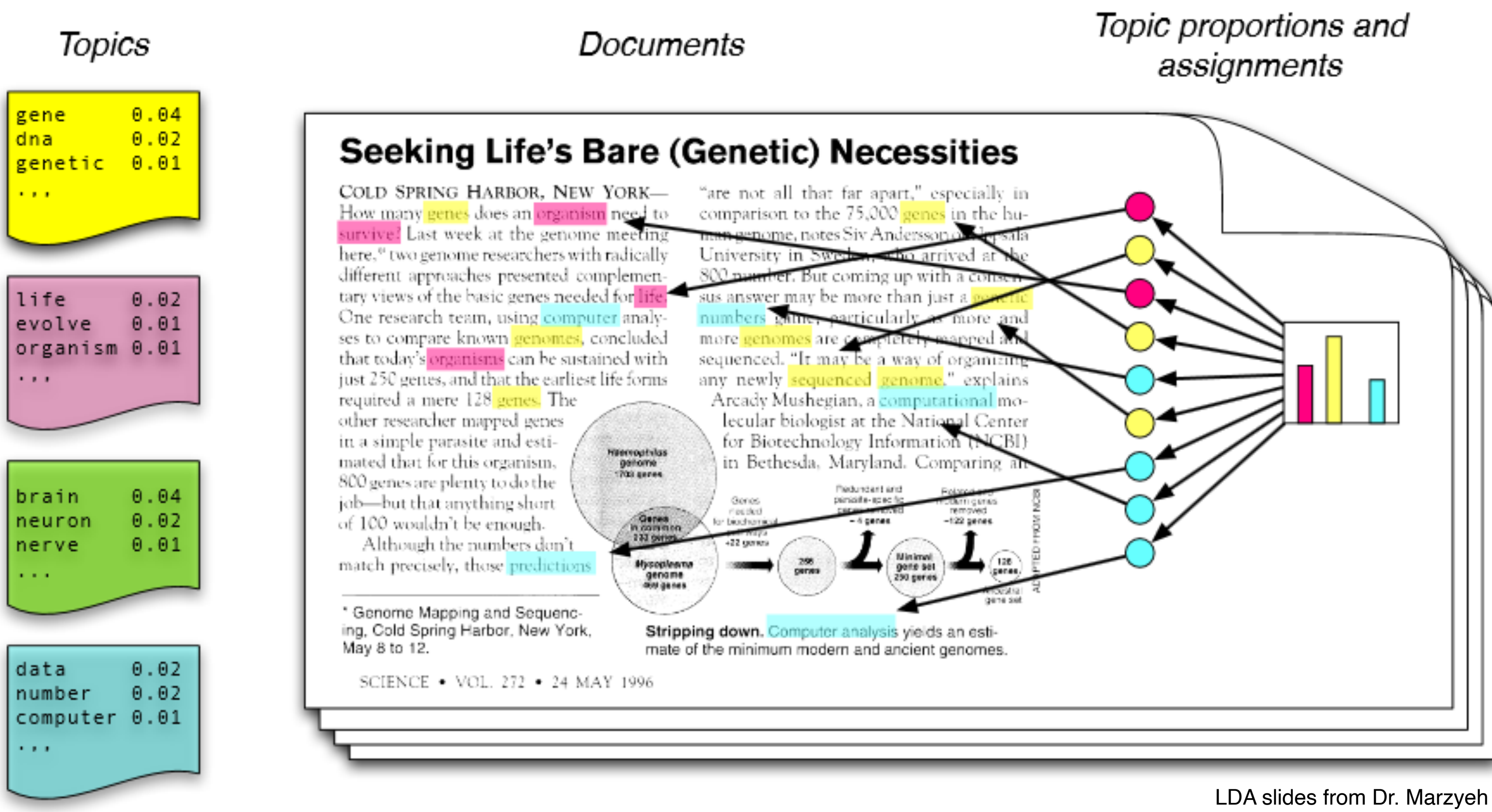
Predicting early psychiatric readmission by LDA

- Can we predict 30-day psych readmission?
- Cohort: patients admitted to a psych inpatient ward between 1994-2012 with a principal diagnosis of major depression
  - 470 of 4687 were readmitted within 30 days with a psych diagnosis; 2977 additionally were readmitted in 30 days with other diagnoses; 1240 not readmitted
- Compare predictive models built using SVM from
  - baseline clinical features
    - age, gender, public health insurance, Charlson comorbidity index
  - + common words from notes
    - 1–1000 most informative words per patient, by TF-IDF
    - top-1 used 3013 unique words, top-10 used 18 173, top-1000 use almost entire vocabulary (66 429/66 451 words)
  - + 75 topics from LDA on notes
- AUCs range from 0.62 to 0.78; difficult to reproduce on larger, more heterogeneous data sets

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational Psychiatry, 6(10), e921–5. http://doi.org/10.1038/tp.2015.182

Intuition: Documents are made of Topics

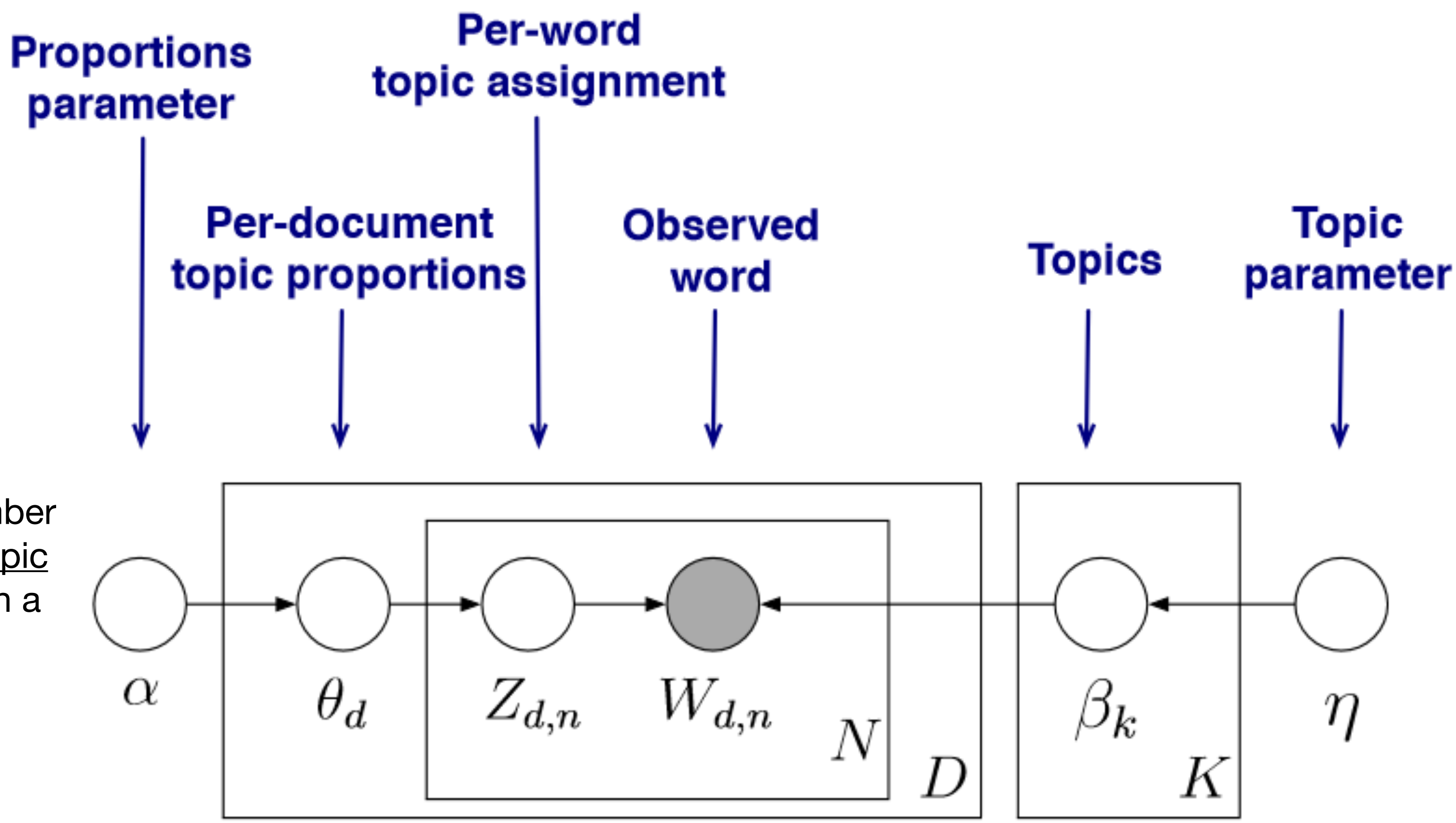
- Every <u>document</u> is a mixture of <u>topics</u>
- Every <u>topic</u> is a distribution over <u>words</u>
- Every <u>word</u> is a draw from a <u>topic</u>



LDA – Latent Dirichlet Allocation

- We <u>observe words</u>, we <u>infer everything else,</u> with our <u>assumed structure</u>

- α is the number of times a <u>topic</u> is sampled in a <u>document</u> (prior)



- η is the number of times <u>words</u> are sampled from a <u>topic</u> (prior)

$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Table 2. Example topics for MDD patients readmitted with a psychiatric diagnosis within 30 days

Terms	_Topic annotation_
*patient _alcohol withdrawal_ depression _drinking_ end ativan _etoh drinks_ medications clinic inpatient diagnosis days hospital _<substance use treatment program name> use abuse_ problem number	Alcohol
*mg daily discharge _anxiety klonopin seroquel clonazepam_ admission wellbutrin given md lexapro date b signed night low admitted sustained hospitalization	Anxiety
*_ideation suicidal mood decreased_ hallucinations history _depressed depression thought_ psychiatric energy denied sleep auditory appetite homicidal symptoms increased speech _thoughts_	Suicidality
*_ect depression_ treatment treatments dr mg course _<ECT physician name>_ symptoms received medications prior improved decreased medication md trials tsh continued qhs	ECT
*_weight eating_ admission discharge hospital intake _loss_ date hospitalization day dr week physical months _prozac food_ increased md did _anorexia_	Anorexia
*_seizure seizures_ intact _eeg neurology_ normal _temporal dilantin_ head _bilaterally events activity_ weakness sensation disorder tongue _neurologist brain_ loss _tegretol_	Seizure
*_therapist mother_ program _father_ disorder age school parents brother _abuse_ treatment _relationship_ outpatient college behavior partial plan currently _group personality_	Psychotherapy
*psychiatry _suicide overdose attempt transferred_ depression _transfer_ level _tylenol_ hospital service _unit_ normal floor screen _tox_ room admission medical general	Overdose
*_baby delivery bleeding vaginal breast feeding cesarean_ weight ibuprofen _maternal newborn_ available p fever _pregnancy_ sex estimated _danger_ gp	Postpartum
*_psychotic thought_ features _paranoid psychosis paranoia_ symptoms psychiatric dose continued treatment mental cognitive memory _risperidone_ people th somewhat interview affect	Psychosis

Abbreviation: MDD, major depressive disorder; ECT, electroconvulsive therapy.

Table 3. Comparison of models with and without inclusion of LDA topics

Configuration	_AUC_	_Sensitivity_	_Specificity_
Baseline = age/gender/insurance/ Charlson	0.618	0.979	0.104
Baseline+top-1 words	0.654	—	—
Baseline+top-10 words	0.676	—	—
Baseline+top-100 words	0.682	—	—
Baseline+top-1000 words	0.682	0.213	0.945
Baseline+75 topics (no words)	0.784	0.752	0.634

Abbreviations: AUC, area under the curve; LDA, Latent Dirichlet Allocation.

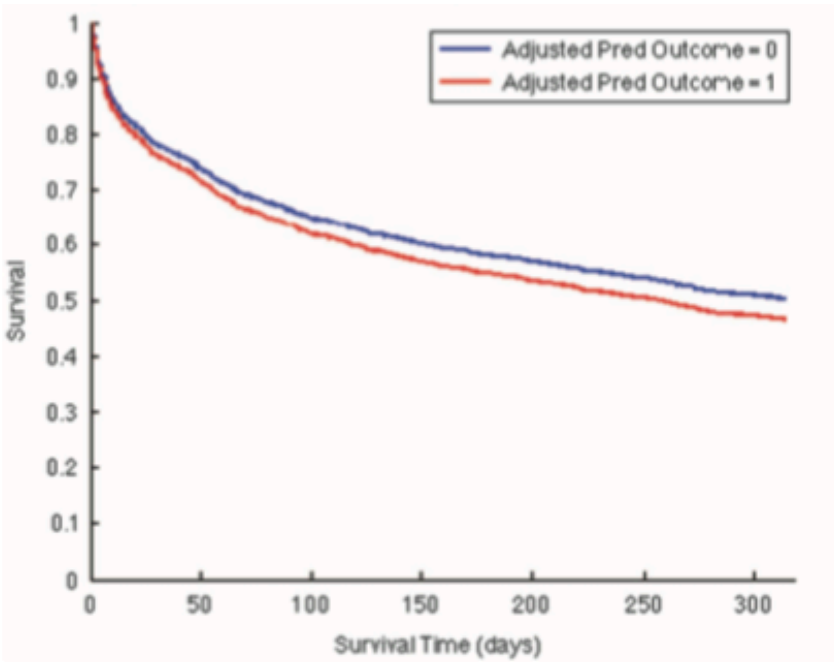


Figure 1. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using baseline sociodemographic and clinical variables only. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

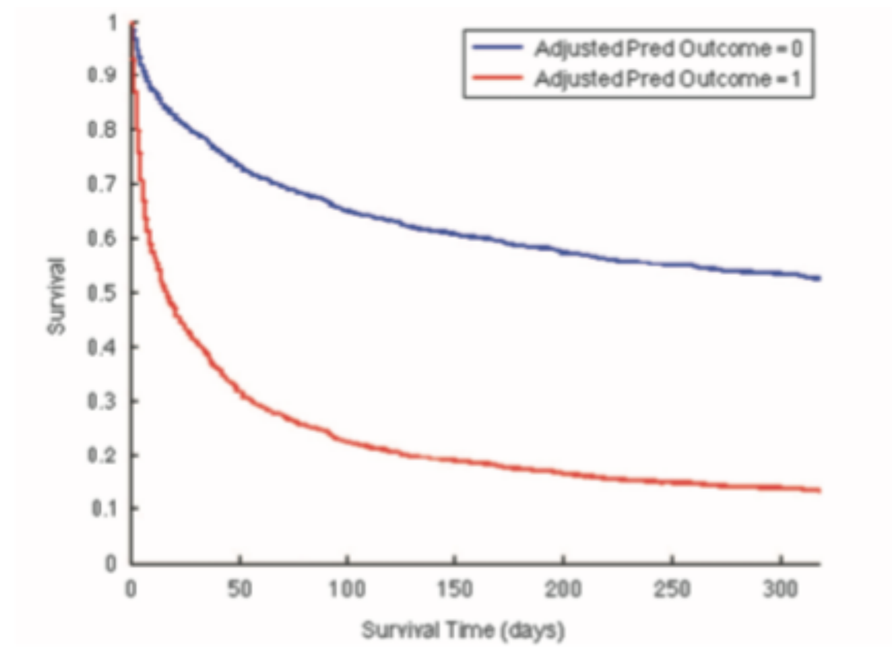


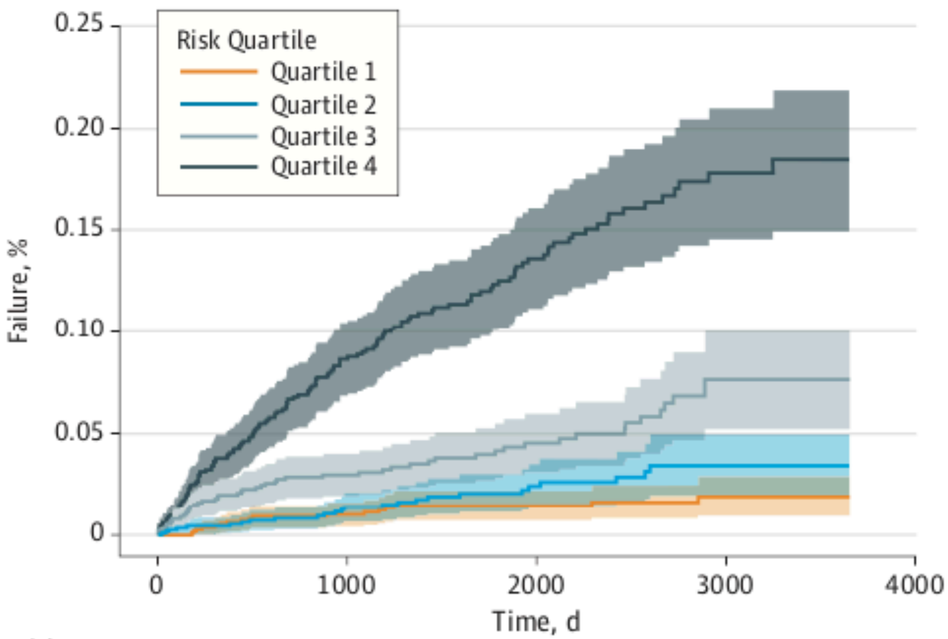
Figure 2. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using the baseline variables and 75 topics. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

Prediction of Suicide and Accidental Death After Discharge

- Very large cohort: 845 417 discharges from two medical centers, 2005–2013
  - 458 053 unique individuals
- Imbalanced: 235 suicides, but all-cause mortality was 18% during 9 years
- Censoring: median follow-up was 5.2 years
- “Positive Valence” assessed using *curated list of 3000 terms* found in discharge summaries
  - “Valence, as used in psychology, especially in discussing emotions, means the intrinsic attractiveness/"good"-ness (positive valence) or averseness/"bad"-ness (negative valence) of an event, object, or situation.[1] The term also characterizes and categorizes specific emotions. For example, emotions popularly referred to as "negative", such as anger and fear, have negative valence. Joy has positive valence.” —Wikipedia

McCoy, T. H., Jr, Castro, V. M., Roberson, A. M., Snapper, L. A., & Perlis, R. H. (2016). Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. JAMA Psychiatry, 73(10), 1064–8. http://doi.org/10.1001/jamapsychiatry.2016.2172

Figure 1. Kaplan-Meier Curves for Time to Death by Suicide Among 458 053 Individuals With at Least 1 Hospital Discharge by Predicted Risk Quartile

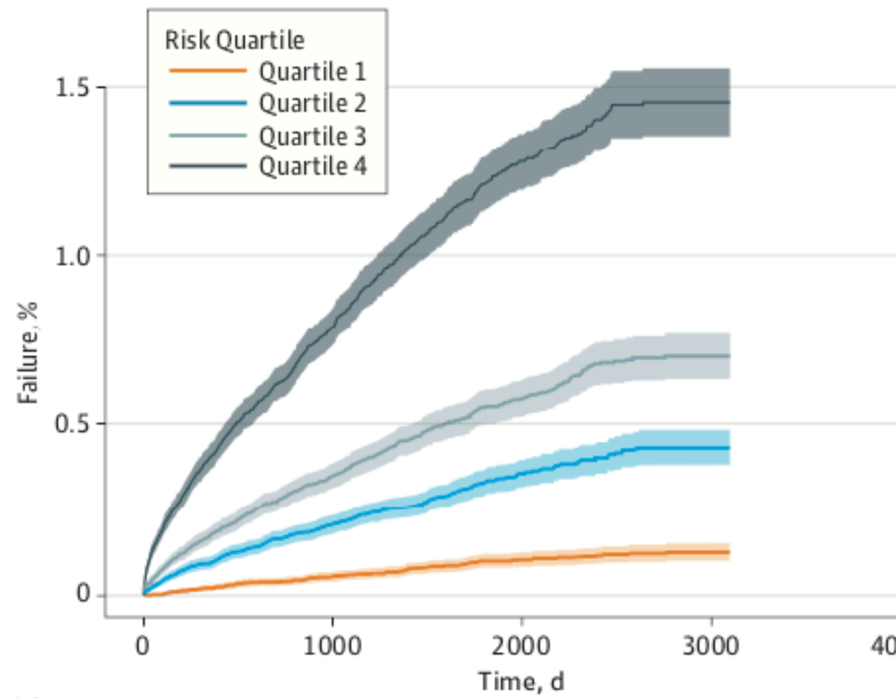


No. at risk

	0	1000	2000	3000
Quartile 1	114514	93698	63289	31025
Quartile 2	114513	85693	52292	22697
Quartile 3	114513	82810	49258	21580
Quartile 4	114513	85746	51707	21541

The axes are rescaled inside the figure to improve interpretability.

Figure 2. Kaplan-Meier Curves for Time to Death by Suicide or Accidental Death Among 458 053 Individuals With at Least 1 Hospital Discharge by Predicted Risk Quartile



No. at risk

	0	1000	2000	3000
Quartile 1	114514	99448	68260	33935
Quartile 2	114513	89270	55871	25263
Quartile 3	114513	84465	50944	21768
Quartile 4	114513	74764	41471	15877

The axes are rescaled inside the figure to improve interpretability.

Tensor Factorization for Unsupervised Exploitation of Text

- Goals:
  - Identify patients with subtypes of lymphoma by analysis of their pathology notes
- Unsupervised approach
  - Do the core “clusters” of patient descriptions correspond to known lymphoma types?
  - Can we use these to help refine out understanding of the types?

Luo, Y., Sohani, A. R., Hochberg, E. P., & Szolovits, P. (2014). Automatic lymphoma classification with sentence subgraph mining from pathology reports. Journal of the American Medical Informatics Association, 21(5), amiajnl–2013–002443–832. http://doi.org/10.1136/amiajnl-2013-002443

Generalizing Matrix to Tensor

- *N*-dimensional data structure (N ≥ 3)
- Example: patient and timed physiological measurements

t

0 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
David	78	49	143	4	111	162	5.8	3.5
Mary	123	68	140	3	108	119	9.1	2.4
Robert	127	66	140	4.3	108	158	9.2	2.4
Andrea	136	70	138	4.7	110	115	9	1.8

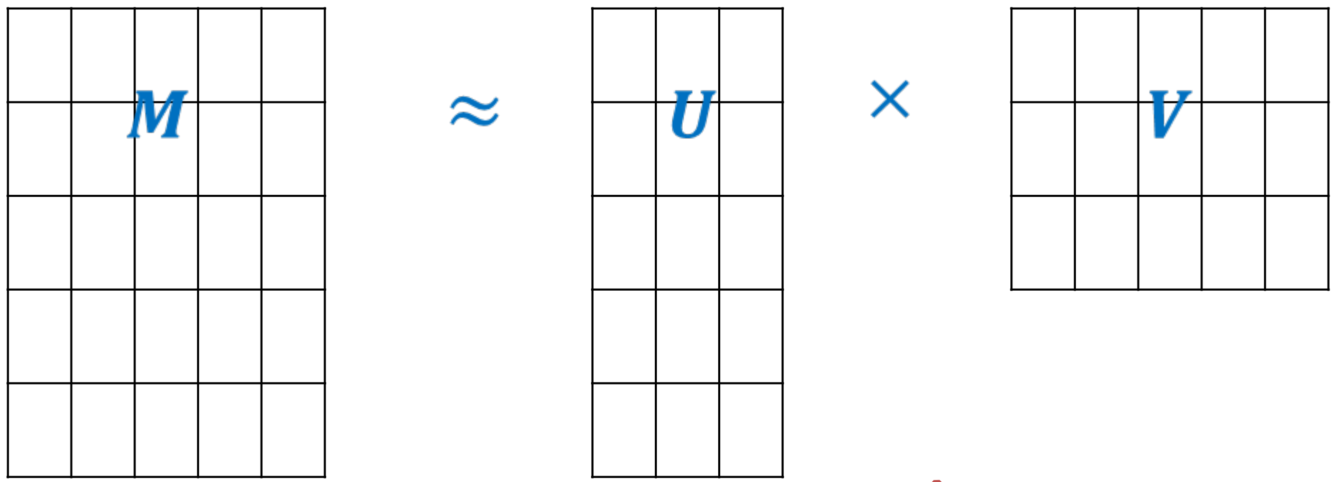
30 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.7
								2.2
								2.7
								1.9

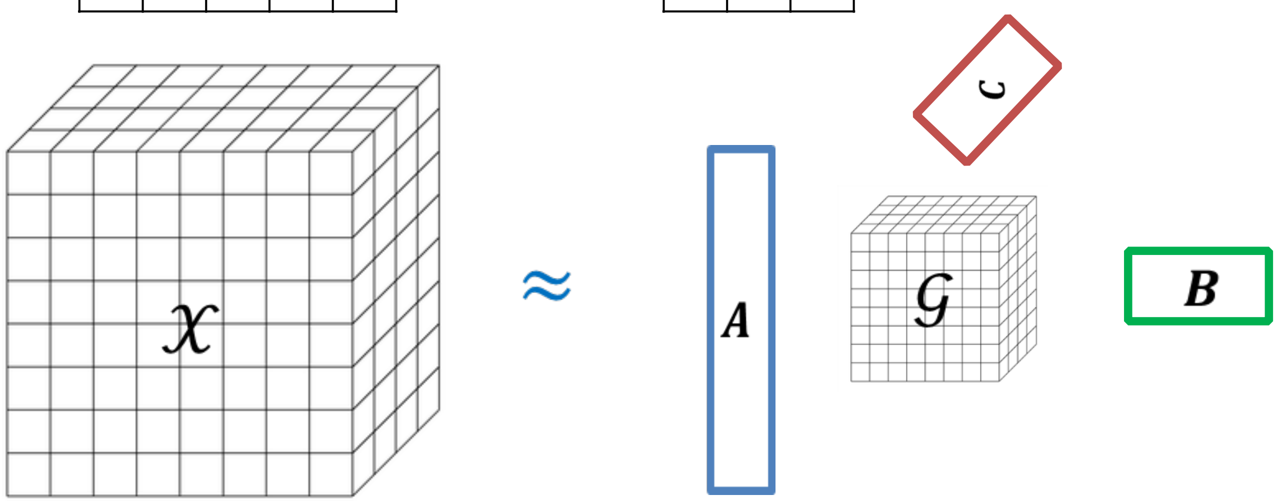
60 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.5
								2.1
								2.5
								2.0

90 min	SBP	DBP	Na	K	Cl	Glucose	Ca	Mg
								3.8
								2.2
								2.3
								1.7

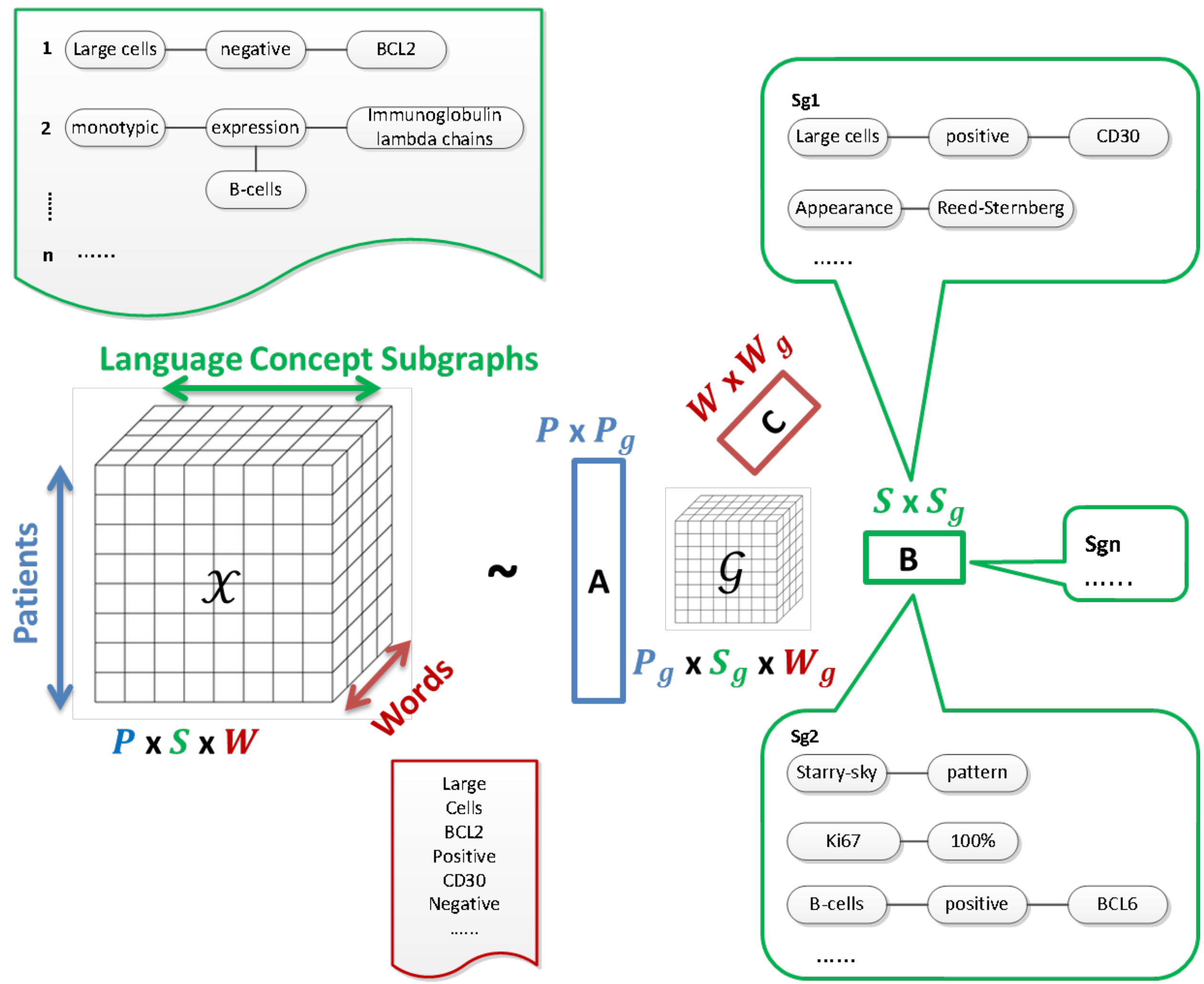
Non-Negative Matrix and Tensor Factorization

- NMF extension to tensors of arbitrary order
- Tucker model, a generalized form of spectral modeling





Multi-Mode Learning
SANTF schematic view



Unsupervised Learning – Clustering Results

- Non-negative matrix factorization as baseline
  - Traditional two-dimensional view
  - Three matrix formulation baselines
    - Patient by word
    - Patient by subgraph
    - Patient by subgraph and word
- SANTF as target (Luo et al. 2014b)
  - Patient by subgraph by word

Clinical Narrative Text			
Lymphoma	All	Train	Test
DLBCL	589	305	284
Follicular	184	101	83
Hodgkin	124	65	59

Metrics / Methods	Macro Average			Micro Average		
	Precision	Recall	F-measure	Precision	Recall	F-measure
(1) NMF pt × wd	0.492	0.495	0.428	0.626	0.626	0.626
(2) NMF pt × sg	0.621	0.765	0.601	0.605	0.605	0.605
(3) NMF pt × [sg wd]	0.637	0.787	0.615	0.614	0.614	0.614
(4) SANTF pt × sg × wd	**0.720**[1,2,3]	**0.849**[1,2,3]	**0.743**[1,2,3]	**0.751**[1,2,3]	**0.751**[1,2,3]	**0.751**[1,2,3]

Outline

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- UMLS resources
- ML to expand terms
- pre-NN ML to identify entities and relations
- **Language models**
- Neural methods

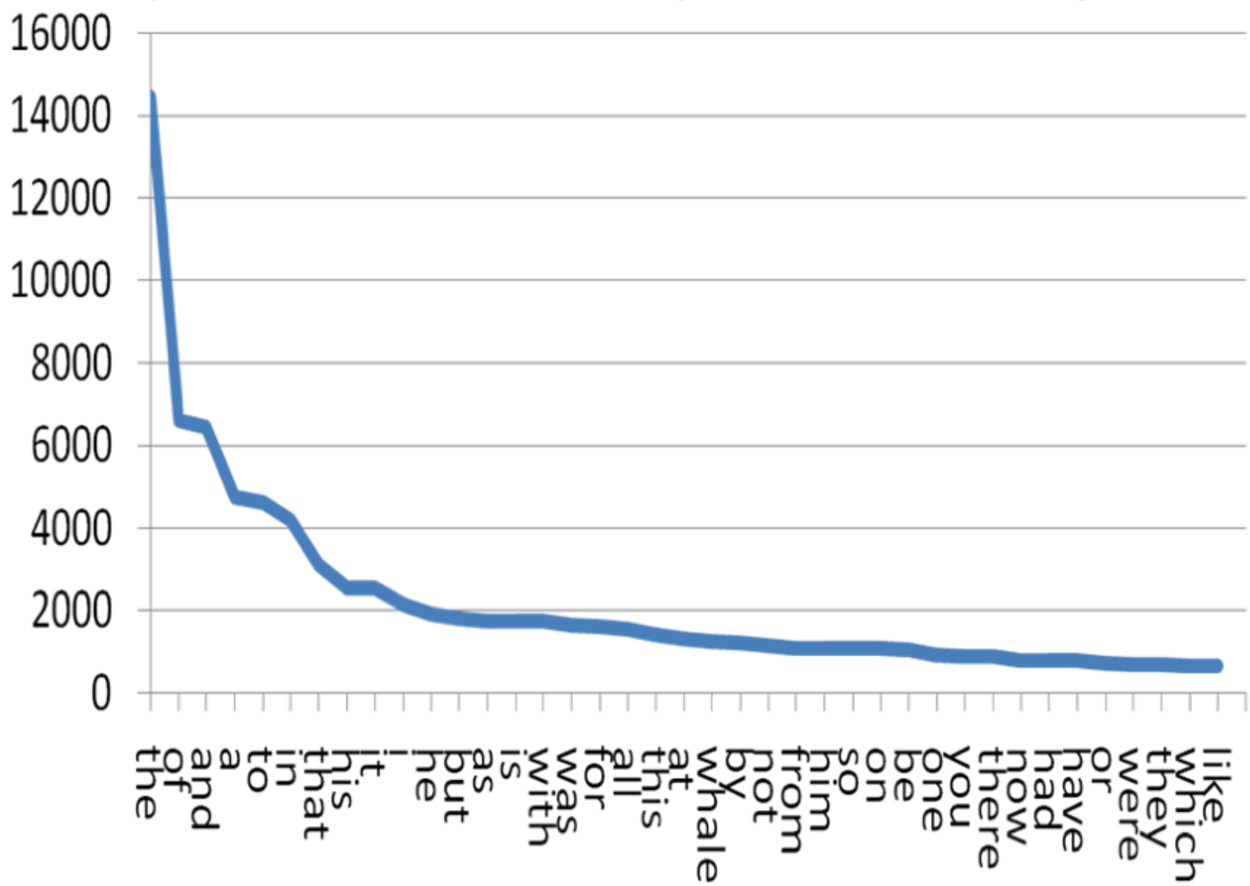
Language Modeling

- Predict the next token given the ones before it (*autoregressive* model)
  - In unigram model, P(token) is just estimated from frequency in corpus
- Markov assumption simplifies model so
  - P(token | stuff before) = P(token | previous token) [bigram model]
  - $P(t_k \mid \text{stuff before}) = P(tk \mid t_{k-1}, \ldots, t_{k-n})$ [n-gram models]

- Perplexity is an aggregate measure of the complexity of a corpus
  - $2^{H(p)}$ where H(p) is the entropy of the probability distribution
  - intuitively, the number of likely ways to continue a text
    - a perplexity of *k* means that you are as surprised on average as you would have been if you had to guess between *k* equiprobable choices at each step
  - For example, we compared perplexity of dictated doctors' notes (8.8) vs. that of doctor-patient conversations (73.1)
    - What does that tell you about the difficulty of accurately transcribing speech for these applications?

Statistical Models of Language
Zipf's law

- There are very few very frequent words
- Most words have very low frequencies
- The frequency of a word is inversely proportional to its rank – $f(k) \propto 1/k$
- In the Brown corpus, the 10 top-ranked words make up 23% of total corpus size (Baroni, 2007)



N-gram models

- Shakespeare as a Corpus
  - N=884,647 tokens, V=29,066
  - Shakespeare produced 300,000 bigram types out of V^2= 844 million possible bigrams...
    - So, 99.96% of the possible bigrams were never seen
- Google released corpus of 1,024,980,267,229 (i.e., ~1T) words in 2006
  - 13.6M unique words occurring at least 200 times
  - 1.2B five-word sequences that occur at least 40 times

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html

Example Google 3-grams

ceramics	collectables	collectibles	55
ceramics	collectables	fine	130
ceramics	collected	by	52
ceramics	collectible	pottery	50
ceramics	collectibles	cooking	45
ceramics	collection	,	144
ceramics	collection	.	247
ceramics	collection	</S>	120
ceramics	collection	and	43
ceramics	collection	at	52
ceramics	collection	is	68
ceramics	collection	of	76
ceramics	collection	|	59
ceramics	collections	,	66
ceramics	collections	.	60
ceramics	combined	with	46
ceramics	come	from	69
ceramics	comes	from	660
ceramics	community	,	109
ceramics	community	.	210
ceramics	community	for	61
ceramics	companies	.	53
ceramics	companies	cpnsultants	173

Generating Sequences

- This model can be turned around to generate random sentences that are like the sentences from which the model was derived.
- Generally attributed to Claude Shannon.
  - Sample a random bigram (<s>, w) according to its probability
  - Now sample a random bigram (w, x) according to its probability
  - Where the prefix w matches the suffix of the first.
  - And so on until we randomly choose a (y, </s>)
- Then string the words together

```
<s> I
    I want
      want to
           to get
              get Chinese
                  Chinese food
                          food </s>
```

Generating Shakespeare

Unigram	• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have • Every enter now severally so, let • Hill he late speaks; or! a more to leg less first you enter • Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like
Bigram	• What means, sir. I confess she? then all sorts, he is trim, captain. •Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. •What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman? •Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
Trigram	• Sweet prince, Falstaff shall die. Harry of Monmouth's grave. • This shall forbid it should be branded, if renown made it empty. • Indeed the duke; and had a very good friend. • Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
Quadrigram	• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; • Will you not tell me who I am? • It cannot be but so. • Indeed the short and the long. Marry, 'tis a noble Lepidus.

Generating the *Wall Street Journal*

unigram: Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

bigram: Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

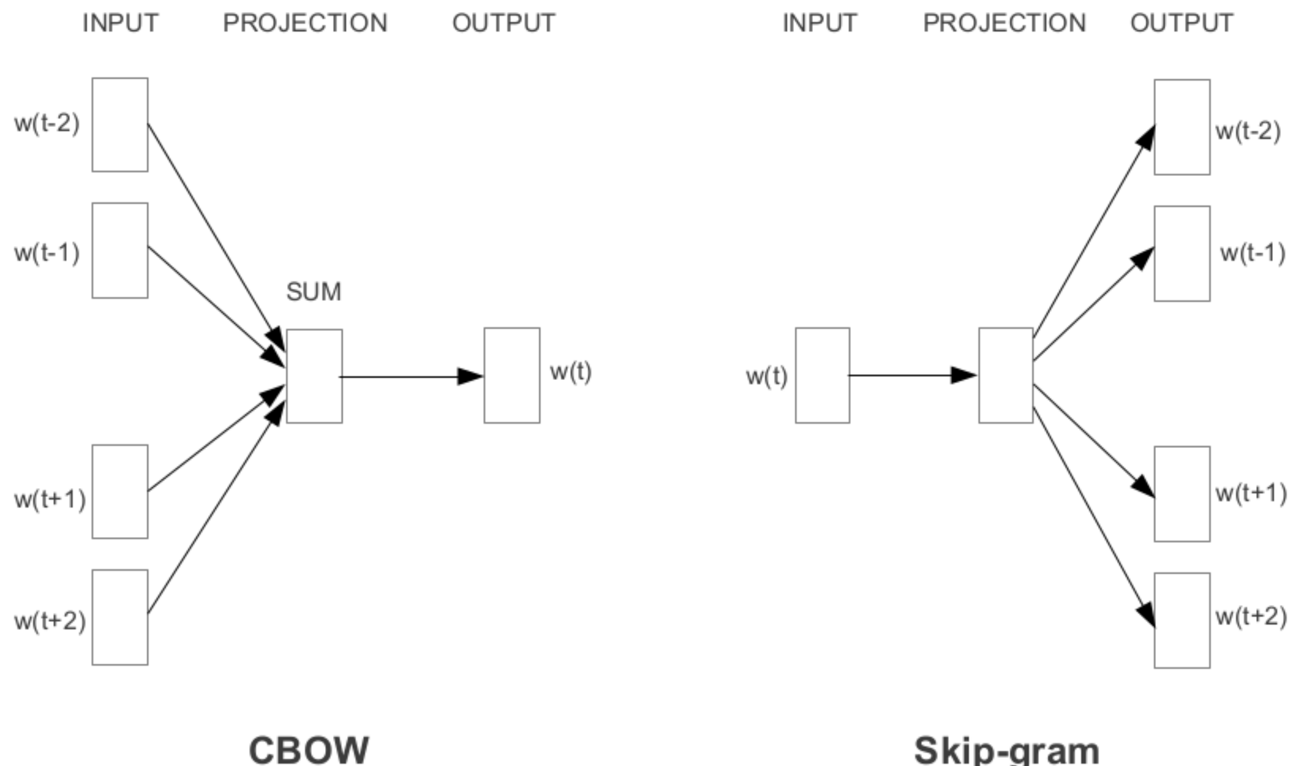
trigram: They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Outline

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- UMLS resources
- ML to expand terms
- Pre-NN ML to identify entities and relations
- Language models
- **Neural methods**

Distributional Semantics

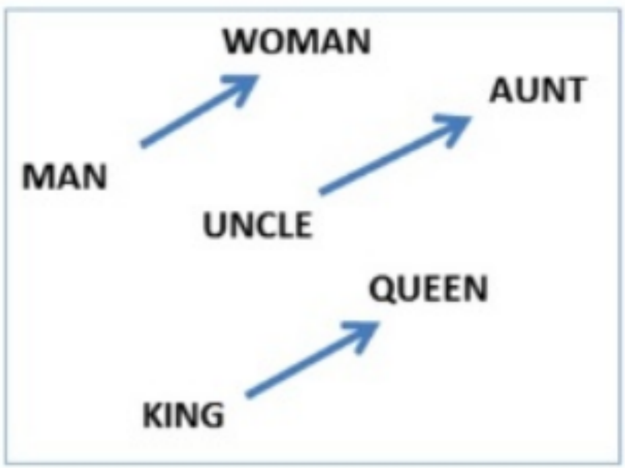
- Terms that appear in the same context of other words are (probably) semantically related
- Every term is mapped to a high-dimensional vector (the embedding space)
- Ever more sophisticated versions of embeddings, equivalent to matrix factorization
    - Word2Vec
    - GloVe
    - Elmo
    - Bert
    - GPT

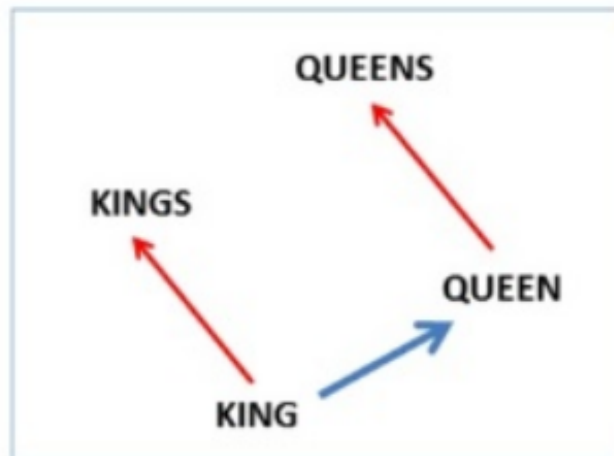


word2vec

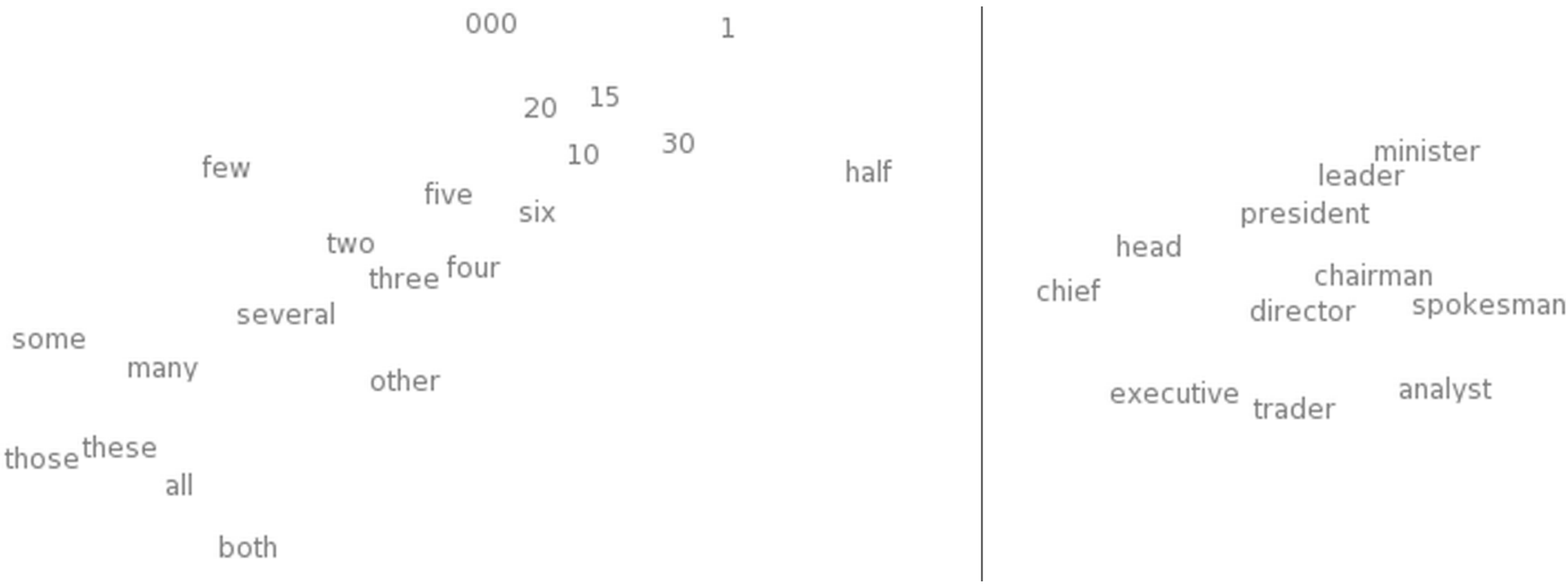
Plausibility of semantic claims

vec("man") – vec("king") + vec("woman") = vec("queen")





t-Distributed Stochastic Neighbor Embedding



van der Maaten, L., Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2579–2605.

Feature extraction for phenotyping from semantic and knowledge resources (SEDFE)

- Goal: "fully automated and robust unsupervised feature selection method that leverages only publicly available medical knowledge sources, instead of EHR data"
  - Surrogate features derived from knowledge sources
- Method:
  - Build a word2vec skipgram model from .5M Springer articles (2006-08) to yield 500-D vectors for each word
  - Sum vectors for each word in the defining strings for UMLS Concepts, weighted by IDF
  - For each disease in Wikipedia, Medscape eMedicine, Merck Manuals Professional Edition, Mayo Clinic Diseases and Conditions, and MedlinePlus Medical Encyclopedia use NER to find all concepts related to the phenotype
- Retain only concepts that occur in at least 3 of 5 knowledge sources
- Choose top *k* concepts whose embedding vectors are closest (by cos distance) to the embedding of the phenotype
- Define the phenotype as a linear combination of its related concepts, learn weights by least squares, and choose *k* to minimize BIC

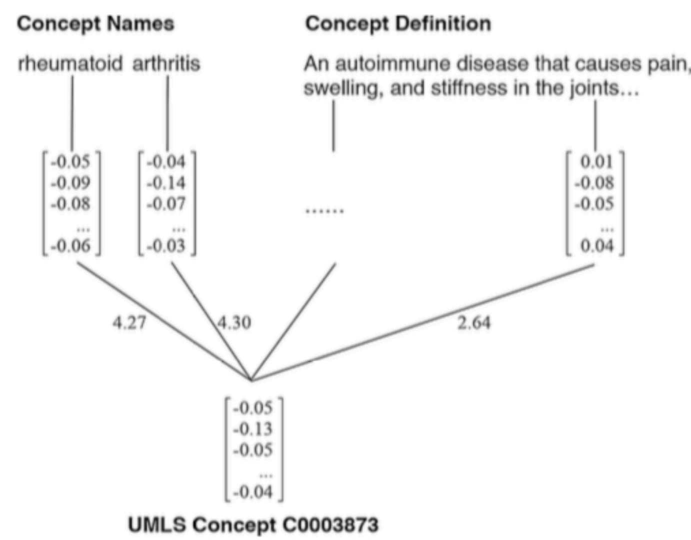


Fig. 1. Generating concept vector representations from word vectors in the paraphrase.

Ning, W., Chan, S., Beam, A., Yu, M., Geva, A., Liao, K., et al. (2019). Feature extraction for phenotyping from semantic and knowledge resources. *Journal of Biomedical Informatics*, *91*, 103122. http://doi.org/10.1016/j.jbi.2019.103122

Evaluating SEDFE

- Used to create phenotypes for coronary artery disease (CAD), rheumatoid arthritis (RA), Crohn's disease (CD), ulcerative colitis (UC), and pediatric pulmonary arterial hypertension (PAH)

Number of features from various methods.

	Phenotype				
	CAD	RA	CD	UC	PAH
Number of concepts extracted from source articles	805	1067	1057	700	58
Number of expert-curated features[a]	34	21	47	48	24
Number of features from SAFE	19	15	16	17	28
Number of features from SEDFE	36	26	18	27	35

[a] The source of PAH features in the original study includes both expert curation and algorithm selection.

	AFEP	SAFE	SEDFE
Commonality	Applies NER to online articles about the target phenotype to find an initial list of clinical concepts as candidate features		
Feature selection method	Frequency control, then threshold by rank correlation with the NLP feature representing the target phenotype	Frequency control, majority voting, then use sparse regression to predict the silver-standard labels derived from surrogate features	Majority voting; Use concept embedding to determine feature relatedness; Use semantic combination and the BIC to determine the number of needed features
Data requirement	EHR data (hospital dependent and not sharable)	EHR data (hospital dependent and not sharable)	A biomedical corpus for training word embedding (usually sharable)
Tuning parameters	Threshold for the rank correlation	(1) Upper and lower thresholds of the surrogate features for creating the silver standard labels, which are affected by the distribution of the features, and therefore phenotype dependent; (2) The number of patients to sample, which affects the number of selected features	The word embedding parameters, which are not overly sensitive. The embedding is done only once for all phenotypes

ANN model for de-identification

- Label-sequence optimization layer

$$s(y_{1:n}) = \sum_{i=1}^{n} \mathbf{a}_i[y_i] + \sum_{i=2}^{n} T[y_{i-1}, y_i]$$

- Label prediction layer
- Character-enhanced token-embedding layer

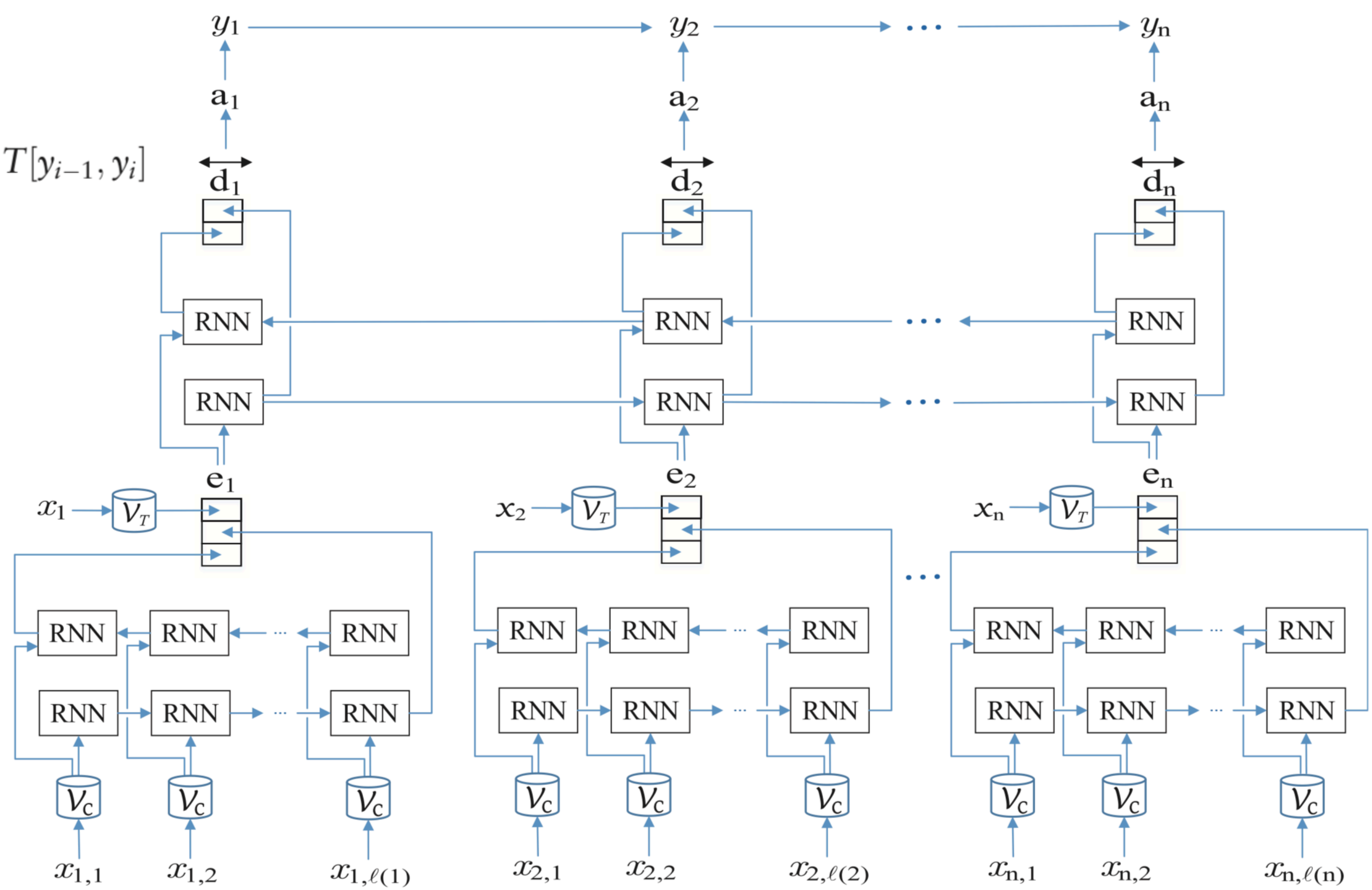


Figure 1. Architecture of the artificial neural network (ANN) model. (RNN, recurrent neural network.) The type of RNN used in this model is long short-term memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. $\mathcal{V}_T$ is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. $\mathcal{V}_C$ is the mapping from characters to character embeddings. $\mathbf{e}_i$ is the character-enhanced token embeddings of the i^{th} token. $\overleftrightarrow{\mathbf{d}}_i$ is the output of the LSTM of the label prediction layer, $\mathbf{a}_i$ is the probability vector over labels, y_i is the predicted label of the i^{th} token.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, ocw156. http://doi.org/10.1093/jamia/ocw156

De-Identifier performance

	Binary HIPAA (optimized by F1-score)			Binary HIPAA (optimized by recall)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
No feature	99.103	99.197	99.150	98.557	99.376	98.965
EHR features	99.100	99.304	99.202	98.771	**99.441**	99.105
All features	**99.213**	**99.306**	**99.259**	**98.880**	99.420	**99.149**

Table 2: Binary HIPAA token-based results (%) for the ANN model, averaged over 5 runs. The metric refers to the detection of PHI tokens versus non-PHI tokens, amongst PHI types that are defined by HIPAA. "No feature" is the model utilizing only character and word embeddings, without any feature. "EHR features" uses only 4 features derived from EHR database: patient first name, patient last name, doctor first name, and doctor last name. "All features" makes use of all features, including the EHR features as well as other engineered features listed in Table 1. "Optimized by F1-score" and "optimized by recall" means that the epochs for which the results are reported are optimized based on the highest F1-score or the highest recall on the validation set, respectively.

“Revolutionary Advances” in Embeddings

- The year 2018 has been an inflection point for machine learning models handling text (or more accurately, Natural Language Processing or NLP for short). Our conceptual understanding of how best to represent words and sentences in a way that best captures underlying meanings and relationships is rapidly evolving.
  —Jay Alammar (http://jalammar.github.io/illustrated-bert/ — *good tutorial*)

- Bidirectional LSTM applied to learn context-specific embeddings (ELMo)
- Transformer architecture — focus on attention mechanism
- Bidirectional Encoder Representations from Transformers (BERT)
- Generative Pre-Training (GPT-2) — transformer with multi-task training, huge corpus, huge model

Sequence-to-Sequence models

- Natural application: machine translation
  - But also usable for question-answer problems
  - Equivalence and natural implication problems
  - Conversion from text to some formal representation
- One of a variety of RNN models

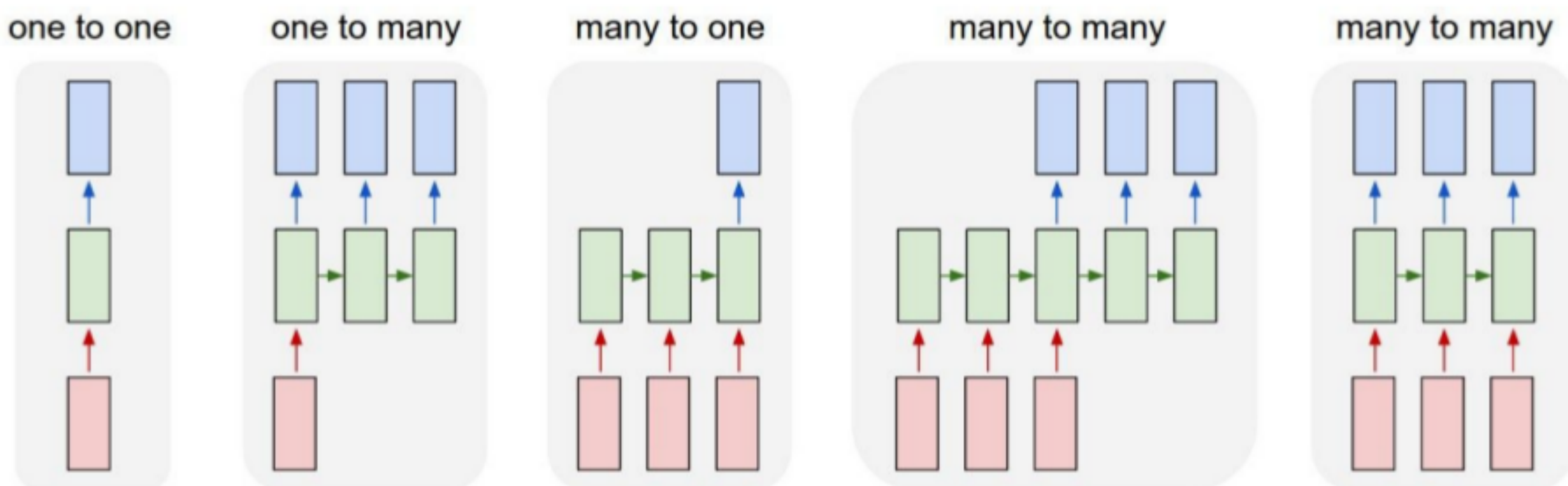


Image Captioning

Translation

Vanilla NN

Sentence Classification

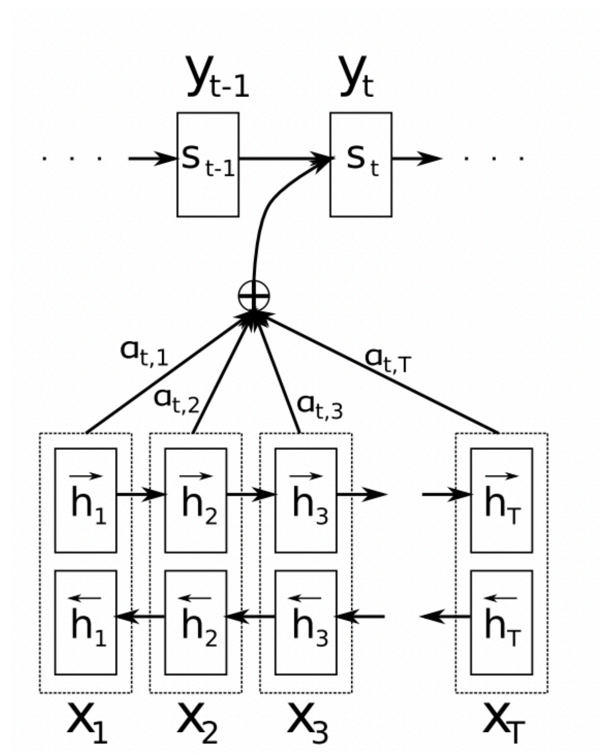
Sequence Classification

- For translation, odd to encode entire meaning of source into one state!

Slide adapted from Anna Rumshisky

Attention tells where in the source to focus

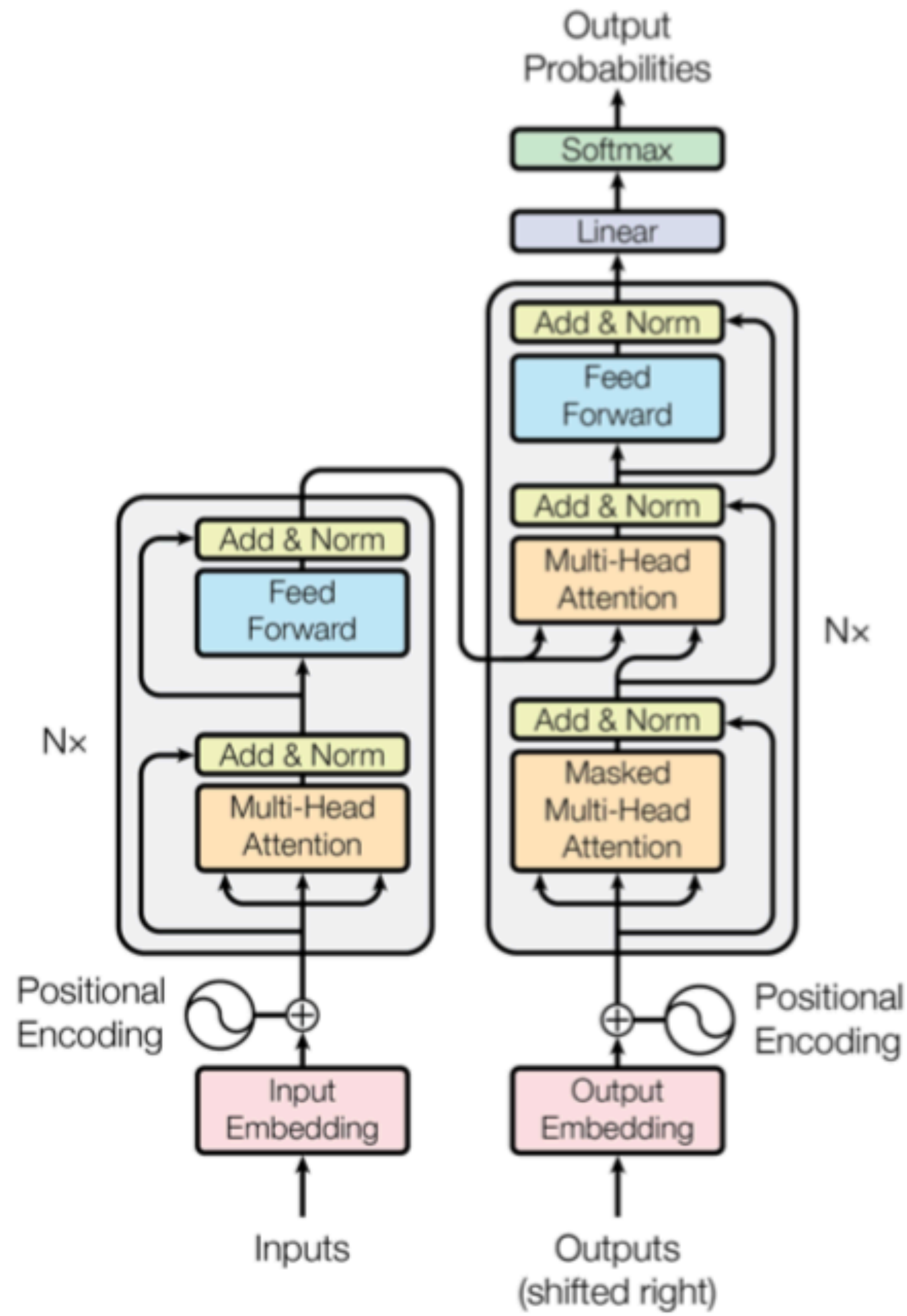
- Each decoder output word y_t now depends on a weighted combination of all the input states, not just the last state.
- The α's are weights that define how much of each input state should be considered for each output.
- Application: Automatic "alignment" of source and target languages in MT



Bahdanau, D., Cho, K., & Bengio, Y. (2014, September 1). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv*.

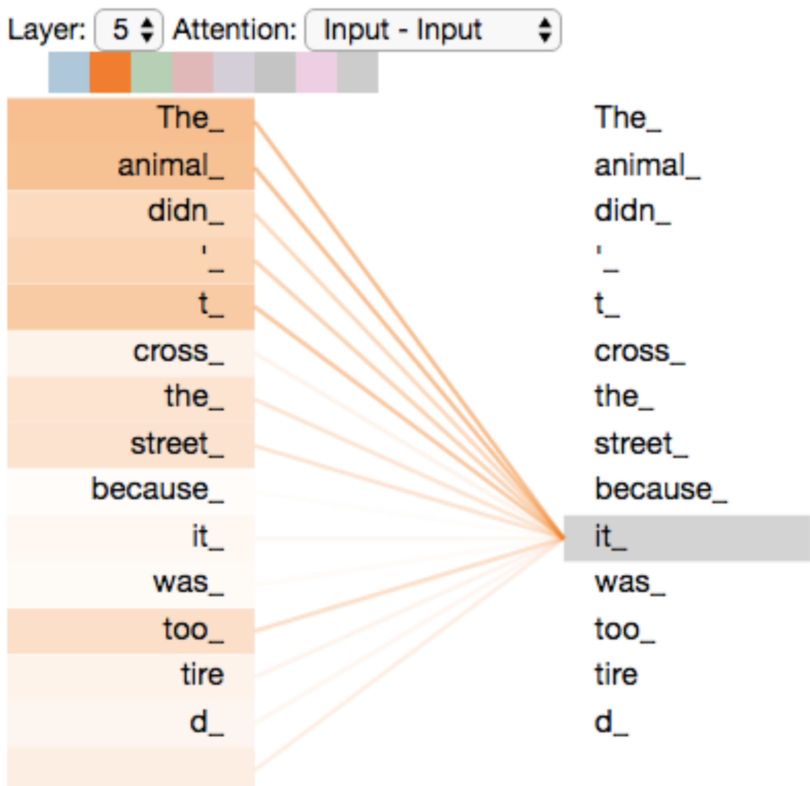
Transformer architecture

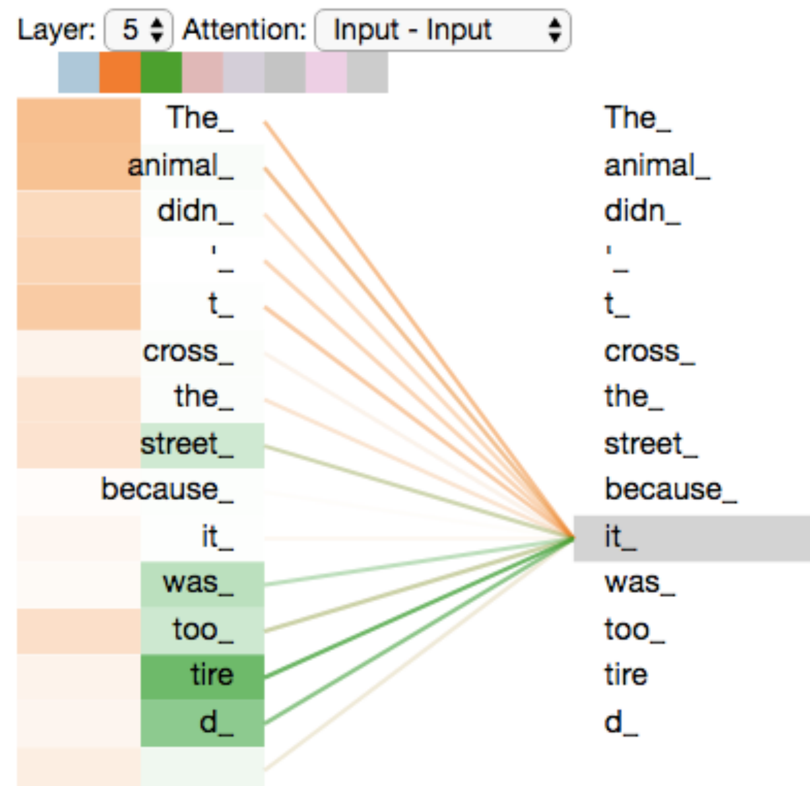
- Details well explained at https://jalammar.github.io/illustrated-transformer/
- Self-attention — vaguely reminiscent of CNNs
- Multi-headed attention — like multiple convolution kernels in CNN
- Key-value pairs passed from encoder to decoder
- Positional encoding
- Only look to left in decoder
- Scaling



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017, June 12). Attention Is All You Need. Lrec 2018.

Multi-headed attention





ELMo—Embeddings from Language Models

- Bidirectional LSTM
- Builds models for every *token*, not just for every *type*
  - i.e., different embeddings for the same word in different contexts
  - basis for word-sense disambiguation
- Significantly improves performance on nearly all NLP tasks

	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 .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

Peters, M. E., Neumann, M., Iyyer, M., 0001, M. G., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *Naacl-Hlt*.

BERT
Bidirectional Encoder Representations from Transformers

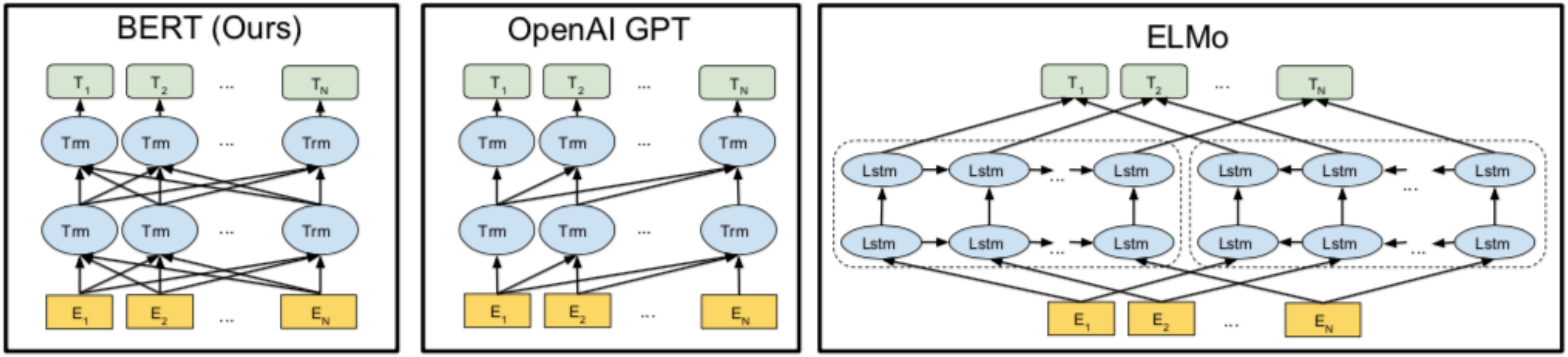


Figure 1: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTM to generate features for downstream tasks. Among three, only BERT representations are jointly conditioned on both left and right context in all layers.

- Word-piece tokens
- Predict masked tokens (~15%)
- Predict next sentence
- Trained on 800M word Books, 2,500M word Wikipedia corpus

- Large performance improvement on many tasks

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 10). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*.

BERT Performance Improvements

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

Table 1: GLUE Test results, scored by the GLUE evaluation server. The number below each task denotes the number of training examples. The "Average" column is slightly different than the official GLUE score, since we exclude the problematic WNLI set. OpenAI GPT = (L=12, H=768, A=12); $\text{BERT}_{\text{BASE}}$ = (L=12, H=768, A=12); $\text{BERT}_{\text{LARGE}}$ = (L=24, H=1024, A=16). BERT and OpenAI GPT are single-model, single task. All results obtained from https://gluebenchmark.com/leaderboard and https://blog.openai.com/language-unsupervised/.

- MNLI Multi-Genre Natural Language Inference
- QQP Quora Question Pairs
- QNLI Question Natural Language Inference
- SST-2 The Stanford Sentiment Treebank
- CoLA The Corpus of Linguistic Acceptability
- STS-B The Semantic Textual Similarity Benchmark
- MRPC Microsoft Research Paraphrase Corpus
- RTE Recognizing Textual Entailment

GPT-2 (Generative Pre-Training)

see https://blog.openai.com/better-language-models/

Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

- Unified transformer-based architecture for many tasks
- Task is itself given as a sequence of tokens, e.g.,
  - (translate to french, english text, french text)
  - (answer the question, document, question, answer)
- WebText: scrape of 8M documents linked from Reddit with “karma” ≥ 3
  - ~40GB
- Byte-level language models
- PPL = perplexity
- ACC = accuracy on a choose best next word out of 10

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019, February 14). Language Models are Unsupervised Multitask Learners.

Open-AI's GPT-2 model
Generative Pre-trained Transformer; ~1.5B parameters

prompt completion

- A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.
- The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

  In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

  "The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

  The stolen material was taken from the University of Cincinnati's Research Triangle Park nuclear research site, according to a news release from Department officials.

  The Nuclear Regulatory Commission did not immediately release any information.

  According to the release, the U.S. Department of Energy's Office of Nuclear Material Safety and Security is leading that team's investigation.

  "The safety of people, the environment and the nation's nuclear stockpile is our highest priority," Hicks said. "We will get to the bottom of this and make no excuses."

https://blog.openai.com/better-language-models/#sample2

Evolving Ideas and Systems
XLNet

- Autoregressive models learn

$$\max_{\theta} \log p_\theta(x_t \mid x_{<t}) = \sum_{t=1}^{T} \log \frac{\exp(h_\theta(x_{1:t-1})^\top e(x_t))}{\sum_{x'} \exp(h_\theta(x_{1:t-1})^\top e(x'))}$$

  - where $e(x)$ is the embedding of x and $h_\theta(x_{1:t-1})$ is the context representation produced by some neural model (RNN, Transformer, …)
  - Learns dependency on only left (or right) context
- BERT (de-noising autoencoder) learns

$$\max_{\theta} \log p_\theta(\bar{x} \mid \hat{x}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{x}) = \sum_{t=1}^{T} m_t \log \frac{\exp(H_\theta(\hat{x})_t^\top e(x_t))}{\sum_{x'} \exp(H_\theta(\hat{x})_t^\top e(x'))}$$

  - where $m_t = 1$ iff x_t is masked, and H_θ is a Transformer that maps a text sequence to a sequence of hidden vectors $H_\theta(x) = [H_\theta(x)_1, \ldots, H_\theta(x)_T]$
  - $\bar{x}$ are the masked tokens, $\hat{x}$ is the corrupted sequence of the original tokens
  - Assumes all masked tokens are independent and [MASK] doesn't appear naturally

- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. NeurIPS 2019.

XLNet

- Permutation language modeling
  - $\max_{\theta} \mathbb{E}_{z \sim Z_T} \Big[\sum_{t=1}^{T} \log p_\theta(x_{z_t}) \,|\, x_{z_{<t}}) \Big]$
  - where Z_T is the set of all permutations of of the index sequence $[1, \ldots, T]$
- Plus many other details
- Overall, better performance