

Lecture 16: Introduction to Natural Language Processing (NLP)

Course: Biomedical Data Science
Parisa Rashidi
Fall 2018

Agenda

- Introduction
- NLP problems
 - A lot of material covered at a high level
- NLP in biomedical and clinical settings

Credit

- The material in these slides are partially based on:
 1. “An Introduction to Clinical Natural Language Processing”, Leonard D’Avolio Dina Demner-Fushman Wendy W. Chapman
 2. “Machine Learning Methods in Natural Language Processing”, Michael Collins
 3. “Biomedical and Clinical Natural Language Processing”, Ozlem Uzuner Meliha Yetsgen Amber Stubbs
 4. “Speech and Language Processing” - Jurafsky and Martin
 5. Natural Language Processing for Precision Medicine, Hoifung Poon, Chris Quirk, Kristina Toutanova, Scott Wen-tau Yih

Applications

Why NLP?

- Increasing amounts of biomedical literature
 - Extracting facts, relations, events into knowledge repositories (text mining)
 - Model organism database curation
 - Question answering (TREC Genomics track)
 - Literature based discovery

Biomedical NLP

- Increasing demands for use of EHR data
 - Phenotyping for genomic-related analysis
 - Linking evidence for Evidence-based medicine
 - Biosurveillance
 - Quality measures
- Majority of EHR data is free text!

Clinical NLP

Sample ER Reports

- 1. HISTORY OF PRESENT ILLNESS:** The patient is a (XX)-month-old child. Family notes that since last night, the patient was fussy and irritable, possibly pulling at her ears. No obvious pain or discomfort with urination. No chest pain, cough. No abdominal pain. No nausea, vomiting or diarrhea and is having usual stool and wet diapers. Child, however, has been much more fussy at times and now comes to the ER for further evaluation. This is the mother's first child. The child is consolable and not constantly crying in the ER and is able to be addressed and approached by me without being upset. Family notes the child has been fully immunized and is compliant with scheduled appointments.
- 2. HISTORY OF PRESENT ILLNESS:** This (XX)-year-old very pleasant gentleman presents to the emergency room with a one day complaint of pain in his right ear. The patient states last evening he thought he had wax. He used a wax softener. He has had lots of drainage from his ear today. He is still having pain. He attributes it to his new hearing aid. He also is complaining of pain now in the right side of his head. There is no nausea, no vomiting, no tinnitus, visual, olfactory or auditory changes. He states the headache is behind the ear, and it is related directly to the pain in his ear. There is no chest pain, no shortness of breath, nausea, vomiting. No other complaints. He is very affable and in no apparent distress.

Primary Provider Clinic Note

Primary Provider Clinic Note
Patient MRN: 0000000
Created: XXXX-XX-XX XX:XX:XX.XXXX

Pt: Bob Builder
contact info: 715-788-9999

General Medicine Clinic Note - follow up visit

HPI:

77 yo old m with h/o HTN, CAD s/p CABG 1988. Endorses intermittent dyspnea. Right eye blindness. CRI (bl 1.5-1.7). Pt has persistent gas/epigastric discomfort.

SocialHx:

lives with wife and son in the Bronx. Requires help with all ADLs. History of tobacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Father died of unknown at 80, Mother died 92.

ALL: PCN (rash)

MEDS:

- 1) ASA 81mg po daily
- 3) Lisinopril 5mg po daily
- 4) Metformin 1000mg po bid
- 5) Cozaar 50mg po qd
- 6) HCTZ 25mg po qd
- 7) simethicone prn
- 8) maalox prn

PE:

97/64, 99, 16
Alert, comfortable appearing NAD
PERRLA, anicteric sclerae, OP moist, no exudates
normal rate, irreg rhythm, no murmurs or gallops
+BS, soft, nt/nd EXT: WWP, no edema.

Labs:

- Na 142, k 4.8, Cl 107, CO2 23, BUN 20, Cr 1.6, Gluc 106, Ca 9.2
- hgba1c 6.9
- urinary microalbumin 2.2

A/P:

- pt 77 yo old man with HTN CAD s/p CABG 1988, Here for f/u.
- leave patient off lasix and Ace-I
- Continue Cozaar and HCTZ
- continue metformin 1000mg po bid
- will follow Cr
- will refer to eye clinic
- f/u 1 month

Primary Provider Clinic
Patient MRN: 00000000
Created: XXXX-XX-XX

Pt: Bob Builder
contact info: 715-788

General Medicine Clinic

HPI:
77 yo old m with h/o
t eye blindness. CRI
SocialHx:
lives with wife and s
tobacco use. Smoked ab
her died of unknown a

ALL: PCN (rash)

MEDS:
1) ASA 81mg po daily
3) Lisinopril 5mg po
4) Metformin 1000mg po
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable ap
PERRLA, anicteric scl
normal rate, irreg rh
+BS, soft, nt/nd EXT

Labs:
- Na 142, k 4.8, Cl 1
- hgbaic 6.9
- urinary microalbumi

A/P:
- pt 77 yo old man w
- leave patient off la
- Continue Cozaar and
- continue metformin
- will follow Cr
- will refer to eye d
- f/u 1 month

ORIGINAL ARTICLE

A Controlled Trial of Renal Denervation for Resistant Hypertension

Deepak L. Bhatt, M.D., M.P.H., David E. Kandzari, M.D., William W. O'Neill, M.D.,
Ralph D'Agostino, Ph.D., John M. Flack, M.D., M.P.H., Barry T. Katzen, M.D.,
Martin B. Leon, M.D., Minglei Liu, Ph.D., Laura Mauri, M.D., Manuela Negoita, M.D.,
Sidney A. Cohen, M.D., Ph.D., Suzanne Oparil, M.D., Krishna Rocha-Singh, M.D.,
Raymond R. Townsend, M.D., and George L. Bakris, M.D.,
for the SYMPLICITY HTN-3 Investigators*

ABSTRACT

BACKGROUND

Prior unblinded studies have suggested that catheter-based renal-artery denervation reduces blood pressure in patients with resistant hypertension.

METHODS

We designed a prospective, single-blind, randomized, sham-controlled trial. Patients with severe resistant hypertension were randomly assigned in a 2:1 ratio to undergo renal denervation or a sham procedure. Before randomization, patients were receiving a stable antihypertensive regimen involving maximally tolerated doses of at least three drugs, including a diuretic. The primary efficacy end point was the change in office systolic blood pressure at 6 months; a secondary efficacy end point was the change in mean 24-hour ambulatory systolic blood pressure. The primary safety end point was a composite of death, end-stage renal disease, embolic events resulting in end-organ damage, renovascular complications, or hypertensive crisis at 1 month or new renal-artery stenosis of more than 70% at 6 months.

RESULTS

A total of 535 patients underwent randomization. The mean (\pm SD) change in systolic blood pressure at 6 months was -14.13 ± 23.93 mm Hg in the denervation group as compared with -11.74 ± 25.94 mm Hg in the sham-procedure group ($P < 0.001$ for both comparisons of the change from baseline), for a difference of -2.39 mm Hg (95% confidence interval [CI], -6.89 to 2.12 ; $P = 0.26$ for superiority with a margin of 5 mm Hg). The change in 24-hour ambulatory systolic blood pressure was -6.75 ± 15.11 mm Hg in the denervation group and -4.79 ± 17.25 mm Hg in the sham-procedure group, for a difference of -1.96 mm Hg (95% CI, -4.97 to 1.06 ; $P = 0.98$ for superiority with a margin of 2 mm Hg). There were no significant differences in safety between the two groups.

CONCLUSIONS

This blinded trial did not show a significant reduction of systolic blood pressure in patients with resistant hypertension 6 months after renal-artery denervation as compared with a sham control. (Funded by Medtronic; SYMPLICITY HTN-3 ClinicalTrials.gov number, NCT01418261.)

From Brigham and Women's Hospital Heart and Vascular Center and Harvard Medical School (D.L.B., L.M.), Boston University School of Public Health (R.D.), and Harvard Clinical Research Institute (R.D., L.M.) — all in Boston; Piedmont Heart Institute, Atlanta (D.E.K.); the Division of Cardiology, Henry Ford Hospital (W.W.O.), and Wayne State University and the Detroit Medical Center (J.M.F.) — all in Detroit; Baptist Cardiac and Vascular Institute, Miami (B.T.K.); New York Presbyterian Hospital, Columbia University Medical Center, and Cardiovascular Research Foundation, New York (M.B.L.); Medtronic CardioVascular, Santa Rosa, CA (M.L., M.N., S.A.C.); University of Alabama at Birmingham, Birmingham (S.O.); Prairie Heart Institute, Springfield, IL (K.R.S.); Perelman School of Medicine, University of Pennsylvania, Philadelphia (S.A.C., R.R.T.); and University of Chicago Medicine, Chicago (G.L.B.). Address reprint requests to Dr. Bhatt at Brigham and Women's Hospital Heart and Vascular Center, 75 Francis St., Boston, MA 02115, or at dlbhattmd@post.harvard.edu.

*A complete list of investigators in the SYMPLICITY HTN-3 trial is provided in the Supplementary Appendix, available at NEJM.org.

This article was published on March 29, 2014, at NEJM.org.

N Engl J Med 2014;370:1393-401.
DOI: 10.1056/NEJMoa1402670
Copyright © 2014 Massachusetts Medical Society.

Free-text

Primary Provider Clin
Patient MRN: 0000000
Created: XXXX-XX-XX

Pt: Bob Builder
contact info: 715-788

General Medicine Clin

HPI:
77 yo old m with h/o
t eye blindness. CRI
SocialHx:
lives with wife and s
tobacco use. Smoked ab
her died of unknown a

ALL: PCN (rash)

MEDS:
1) ASA 81mg po daily
3) Lisinopril 5mg po
4) Metformin 1000mg po
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable ap
PERRLA, anicteric scl
normal rate, irreg rh
+BS, soft, nt/nd EXT

Labs:
- Na 142, k 4.8, Cl 1
- hgbalc 6.9
- urinary microalbumi

A/P:
- pt 77 yo old man w
-leave patient off la
- Continue Cozaar and
-continue metformin
-will follow Cr
- will refer to eye o
- f/u 1 month

ORIGINAL

A Controlled Trial for Resistant

Deepak L. Bhatt, M.D., M.P.H., David
Ralph D'Agostino, Ph.D., John M. F
Martin B. Leon, M.D., Minglei Liu, Ph.D
Sidney A. Cohen, M.D., Ph.D., Suzann
Raymond R. Townsend, M
for the SYMPLICIT

AB

BACKGROUND

Prior unblinded studies have suggested
reduces blood pressure in patients wit

METHODS

We designed a prospective, single-blind
with severe resistant hypertension were
renal denervation or a sham procedure
ing a stable antihypertensive regimen
least three drugs, including a diuretic
change in office systolic blood pressur
was the change in mean 24-hour amb
safety end point was a composite of d
resulting in end-organ damage, renov
at 1 month or new renal-artery stenosis

RESULTS

A total of 535 patients underwent ran
tolic blood pressure at 6 months wa
group as compared with -11.74 ± 2.5
($P < 0.001$ for both comparisons of the
 -2.39 mm Hg (95% confidence interv
with a margin of 5 mm Hg). The chang
sure was -6.75 ± 15.11 mm Hg in the d
the sham-procedure group, for a differ
 $P = 0.98$ for superiority with a margin
ferences in safety between the two gro

CONCLUSIONS

This blinded trial did not show a signi
patients with resistant hypertension
compared with a sham control. (F
ClinicalTrials.gov number, NCT014182

N. ENGL. J. MED. 370:75

Vitamin B12 Correlated With Higher Cancer Risk (But Don't Panic Just Yet)

Posted: 11/19/2013 10:43 am EST

Like Share 8 people like this.



6 4 0 0 0
f Share t Tweet Pin it Email Comment

GET HEALTHY LIVING NEWSLETTERS:

Enter email SUBSCRIBE

FOLLOW: Vitamin b12 Cancer, Vitamin b12 Cancer Risk, Vitamin Cancer, Healthy Living News

By Susan E. Matthews

Vitamin B12 may be known for helping people absorb iron and maintaining the nervous system, but a new study tracking hundreds of thousands of people in Denmark found that higher levels of the vitamin were associated with higher levels of cancer. But don't throw out your supplements yet -- the researchers theorize it's more likely cancer causes high levels of B12 than B12 causing cancer.

everydayHEALTH

Free-text

Primary Provider Clinic
Patient MRN: 0000000
Created: XXXX-XX-XX

Pt: Bob Builder
contact info: 715-788

General Medicine Clinic

HPI:
77 yo old m with h/o
t eye blindness. CRI
SocialHx:
lives with wife and s
tobacco use. Smoked ab
her died of unknown a

ALL: PCN (rash)

MEDS:
1) ASA 81mg po daily
3) Lisinopril 5mg po
4) Metformin 1000mg po
5) Cozaar 50mg po qd
6) HCTZ 25mg po qd
7) simethicone prn
8) maalox prn

PE:
97/64, 99, 16
Alert, comfortable ap
PERRLA, anicteric sc
normal rate, irreg rh
+BS, soft, nt/nd EXT

Labs:
- Na 142, k 4.8, Cl 1
- hgbalc 6.9
- urinary microalbumi

A/P:
- pt 77 yo old man w
- leave patient off la
- Continue Cozaar and
- continue metformin
- will follow Cr
- will refer to eye c
- f/u 1 month

THE NEW ENGLAND JOURNAL OF MEDICINE

ORIGINAL

A Controlled Trial for Resistant

Deepak L. Bhatt, M.D., M.P.H., David
Ralph D'Agostino, Ph.D., John M. F
Martin B. Leon, M.D., Minglei Liu, Ph.D
Sidney A. Cohen, M.D., Ph.D., Suzanne
Raymond R. Townsend, M.D.
for the SYMPLICITY

BACKGROUND
Prior unblinded studies have suggested
reduces blood pressure in patients with

METHODS
We designed a prospective, single-blind
with severe resistant hypertension were
renal denervation or a sham procedure
ing a stable antihypertensive regimen
least three drugs, including a diuretic
change in office systolic blood pressure
was the change in mean 24-hour amb
safety end point was a composite of d
resulting in end-organ damage, renov
at 1 month or new renal-artery stenosis

RESULTS
A total of 535 patients underwent ran
tolic blood pressure at 6 months was
group as compared with -11.74±25
(P<0.001 for both comparisons of the
-2.39 mm Hg (95% confidence interval
with a margin of 5 mm Hg). The change
sure was -6.75±15.11 mm Hg in the d
the sham-procedure group, for a differ
P=0.98 for superiority with a margin
ferences in safety between the two groups

CONCLUSIONS
This blinded trial did not show a significant
patients with resistant hypertension
compared with a sham control. (F
ClinicalTrials.gov number, NCT014182

N. ENGL. J. MED. 370:715

Downloaded from nejm.org at COLUMBIA UNIV HEALTH
Copyright © 2013

Vitamin B12 Correlated With Higher Cancer Risk (But Don't Panic Just Yet)

Posted: 11/19/2013 10:43 am EST

[Like](#) [Share](#) 8 people like this.



6 4 0 0
[Share](#) [Tweet](#) [Pin it](#) [Email](#)

FOLLOW: [Vitamin b12 Cancer](#), [Vitamin b12 Ca](#)

By Susan E. Matthews

Vitamin B12 may be known for helping
and maintaining the nervous system,
tracking hundreds of thousands of people in Denmark found that higher levels of the
vitamin were associated with higher levels of cancer. But don't throw out your
supplements yet -- the researchers theorize it's more likely cancer causes high levels of
B12 than B12 causing cancer.

Posts 1 - 18 (18 total)

chainsawz
Tucson, AZ
Joined: Sep 2008
Posts: 1,423

12 hours ago

[+ Post a Reply](#)

chainsawz wrote:

Wow wee Pee - great news!! I'm glad you found Reggie, but I hope you dump him soon for the
cuter one...NED! Herceptin sure is great stuff....so is tykerb if I must add :>

Lisa - mets to lungs & brain - clear lymph nodes.....the weakest step toward the top of the hill, toward
sunrise, toward hope, is stronger than the fiercest storm.
Diagnosis: 7/21/2008, IDC, 3cm, Stage IV, Grade 3, 0/2 nodes, mets, ER-/PR-, HER2+

jeanne46
CA
Joined: Nov 2006
Posts: 1,312

12 hours ago

[+ Post a Reply](#)

jeanne46 wrote:

Great news. Dance away with Reggie until you can meet up with Ned. Shrink, shrank, shrunk.
Way to go!

Dx 12/05, Stage IV, Grade 3 ER+/PR- Her/2- YOU NEVER KNOW HOW STRONG YOU ARE UNTIL
BEING STRONG IS THE ONLY OPTION YOU HAVE. -anonymous

justjudie
Simi valley, Ca
Joined: Apr 2010
Posts: 42

12 hours ago

[+ Post a Reply](#)

justjudie wrote:

Congratulations to you & Reggie!! Happy news is definitely the BEST!! Good for you Pee.

Judie

RobinWendy
Long Island, NY
Joined: Mar 2007
Posts: 1,841

11 hours ago

[+ Post a Reply](#)

RobinWendy wrote:

So wonderful to hear some good and great news. I'm so happy for you. I hope you and Reggie
stay together until NED cuts in on the dance!!

dx Stage II primary BC in Jan. 2001; dx with DCIS on other breast in Jan. 2003; dx with mets to lungs
in Jan. 2004 and in May 2007, mets spread to liver.
Diagnosis: ER+/PR+, HER2-

lovinmomma
Spokane, WA
Joined: Oct 2008

11 hours ago

[+ Post a Reply](#)

lovinmomma wrote:

Application: automated coding

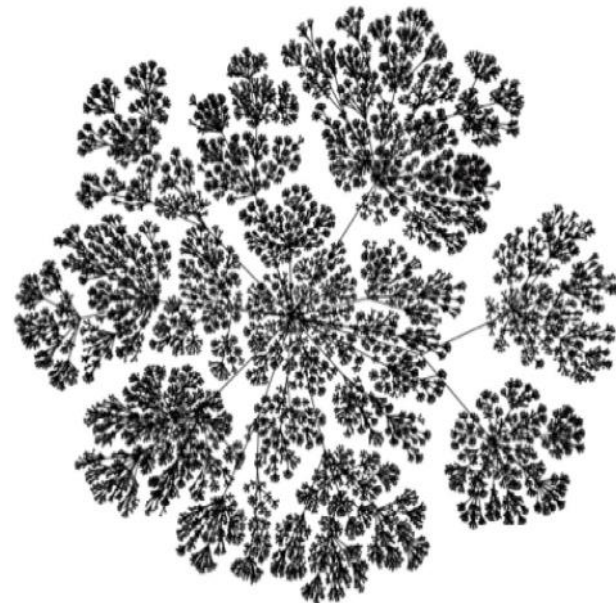
```
Primary Provider Clinic Note
Patient MRN: 0000000
Created: XXXX-XX-XX XX:XX:XX.XXXX

Pt: Bob Builder
contact info: 715-788-9999

General Medicine Clinic Note - follow up visit

HPI:
77 yo old m with h/o HTN, CAD s/p CABG 1988. Endorses intermittent dyspnea. Right eye blindness. CRI (bl 1.5-1.7). Pt has persistent gas/epigastric discomfort.
SocialHx:
lives with wife and son in the Bronx. Requires help with all ADLs. History of tobacco use. Smoked about 1 ppd from age 19 to age 65. Denies use of alcohol. Father died of unknown at 80, Mother died 92.
ALL: PCN (rash)
MEDS:
1) ASA 81mg po daily
2) Lisinopril 5mg po daily
3) Metformin 1000mg po bid
4) Cozaar 50mg po qd
5) HCTZ 25mg po qd
6) simethicone prn
7) maalox prn
PE:
37/64, 99, 16
Alert, comfortable appearing NAD
PERRLA, anicteric sclerae, OP moist, no exudates
normal rate, irreg rhythm, no murmurs or gallops
+BS, soft, nt/nd EXT: WWP, no edema.
Labs:
- Na 142, K 4.8, Cl 107, CO2 23, BUN 20, Cr 1.6, Gluc 106, Ca 9.2
- hgbalc 6.9
- urinary microalbumin 2.2
A/P:
- pt 77 yo old man with HTN CAD s/p CABG 1988, Here for f/u.
- leave patient off lasix and Ace-I
- Continue Cozaar and HCTZ
- continue metformin 1000mg po bid
- will follow Cr
- will refer to eye clinic
- f/u 1 month
```

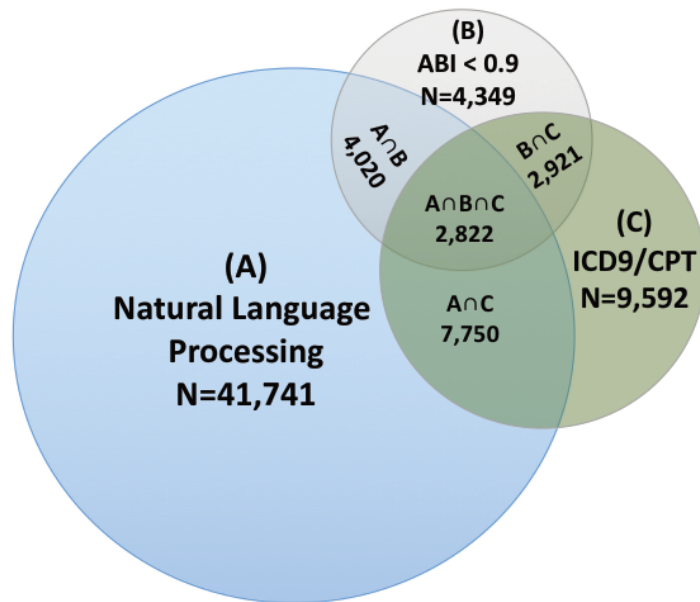
+



→ 401.1 (Hypertension);
428.0 (Congestive heart failure);
369.6 (One eye blindness)

Application: cohort detection

Patient cohort identification - define a set of inclusion/exclusion criteria and enable investigators to query both structured and unstructured clinical information simultaneously



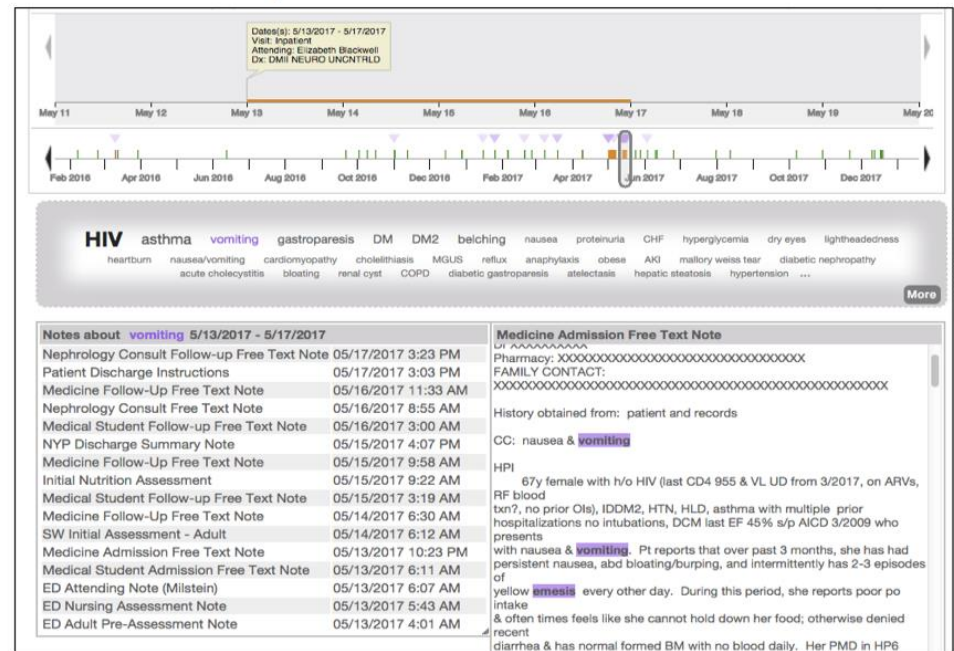
NLP detected 4x more patients than traditional algorithms. More importantly, many patients with Peripheral Arterial Disease (PAD) are missed using standard approaches.

PAD Detection Algorithm	# Unique Patients	Specificity
NLP PAD Algorithm	41741	98%
Rest Pain	2498	98%
Diminished pulses	5773	92%
Ischemic Limb NLP	1339	99%
Peripheral Arterial Disease NLP	31430	99%
Claudication	15337	96%

Application: clinical decision support

- Leverage information from the clinical notes within logic of clinical decision support
 - Drug–drug interactions
 - Allergies
 - ...

Clinical decision support - Developing automated systems to assist decision making in clinical settings by utilizing clinical narratives, e.g. automating colonoscopy follow-up where NLP can be applied to extract necessary information for decision support from pathology and colonoscopy reports.



Demner-Fushman D, Chapman W, McDonald C. (2009) What can natural language processing do for clinical decision support? J Biomed Inform. 42(5):760-762
Demner-Fushman D, Elhadad N. (2016) Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. IMIA Yearbook of Medical Informatics.

Application: info-surveillance from public social media



Jacob D.
Brooklyn, NY
1 friend
66 reviews



2/1/2014

I got very sick from eating here today so be careful. I had dim sum, noodles, and fried dumplings and felt fine until about 6 hours later and then just started vomiting profusely several times. I had no other food today and no alcohol. I will never eat there again.



Jay Dabhi 92.3NowFM @JayDabhi · Mar 6

Went to @GoldenKrustBkry for Patties. They tried to charge me 50 cents for tap water! My tummy is in shambles. Poor cust service #neveragain

[Details](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)



Golden Krust Bakery @GoldenKrustBkry · Mar 6

@JayDabhi We're so sorry to hear. Which location was this at?

[Details](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)



Jay Dabhi 92.3NowFM @JayDabhi · Mar 8

@GoldenKrustBkry Location is 39th St & 8th Ave. Very rude woman at register & food caused diarrhea. Disappointing.

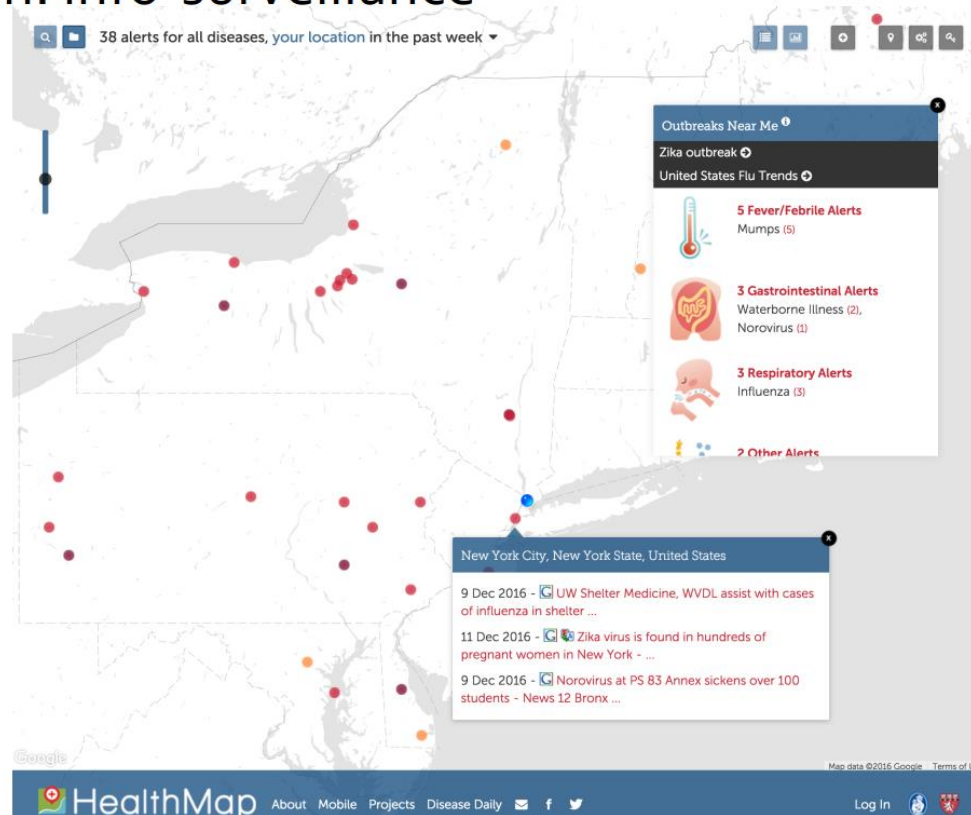
[Details](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)

Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H, Waechter H, Lowe L, Gravano L, Balter S. (2014) Using online reviews by restaurant patrons to identify unreported cases of foodborne illness— New York City, 2012–2013. Centers for Disease Control and Prevention's Morbidity and Mortality Weekly Report (MMWR), 63(20):441–445.

Paul M, Dredze M. (2011) You are what you tweet: Analyzing Twitter for public health. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

Application: info-surveillance



BioSurveillance - for detecting emerging infectious diseases and acts of bioterrorism. These data include the chief complaint fields from outpatient encounters and emergency department visits. Currently, biosurveillance systems retrieve chief complaint section of the encounter note and administrative codes such as ICD codes. Keyword searches look for the occurrence of such word forms as “sore throat” but may miss such related notions as “pain upon swallowing” or “throat feels raw”.

Application: high throughput phenotyping

Words from clinical notes

Laboratory Tests

lupus ana sle complement rheum anti mg ab rash absent esr ulcers igg
plaquenil dna alopecia wt antibody urine systematic dsdna neg rheumatology crp positive
antimalarials metamucil prednisone c4_complement
c3_complement esr rbc_urine total_hemolytic_complement dna_antibody_igg crphi
random_urine_protein antidna_antibodies urine_protein_random urine_creatinine
random_urine_creatinine 710.0 systemic lupus erythematosus

Medications

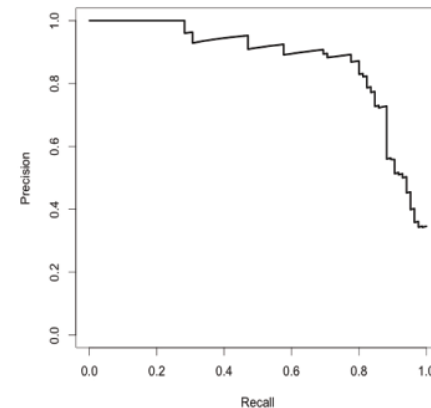
ICD9 codes

mitral valve regurgitation repair severe
replacement mvr moderate tricuspid furosemide
potassium-chloride warfarin heparin-
sodium docusate-sodium acetaminophen epinephrine
magnesium-sulfate milrinone potassium hct
hgb glucose sodium inr-pt plt-count creat mch
magnesium ptt rdw mchc pt urea-n mcv rbc total-co2
wbc chloride 424.0-mitral-valve-
disorders 398.91-rheumatic-heart-failure-congestive
397.0-diseases_of_tricuspid-valve

ct subdural head hematoma right
left hemorrhage frontal neurosurgery subarachnoid
phenytoin-sodium phenytoin-sodium-
extended phenytoin glucose potassium mchc
anion-gap inr-pt total-co2 ptt sodium chloride plt-count
pt calcium rbc wbc creat rdw hgb mcv phosphate mch
E888.9-unspecified-accidental-
fall 852.20-subdural-hemorrhage-
following-injury E880.9-accidental-fall-on-
or-from-other-stairs-or-steps 852.21-
subdural-hemorrhage-following-injury E885.9-
accidental-fall-from-other-tripping-or-stumbling 432.1-subdural-
hemorrhage 801.26-closed-fracture-of-base-skull-with-
subarachnoid-subdural-extradural-hemorrhage 852.00-
subarachnoid-hemorrhage-following-injury

DM2 cohort
identification (n=2,500)

AUC = 0.873725



Application: predictive analytics

MHs (Mental Health subreddits)
I have been considering going for some formal therapy. Any suggestions?
Everyday I feel sad and lonely
Since past sometime I think I am having panic attacks. I really need help from you guys.
It has been so many years, I feel I still can't move on. I am noticing behavior what could be considered "triggers" now.
SW (SuicideWatch)
I know I was never meant to lead this life.
Don't want to hurt the people I care but I can't take this anymore.
Today I felt I have nothing left, why am I even living... I don't see a point.
I'd kill myself, but the other part of me tells me not to waste all the money my parents invested on me..

Table 1: Example titles of posts in the MHs and SW datasets; content has been carefully paraphrased to protect the privacy of the individuals.

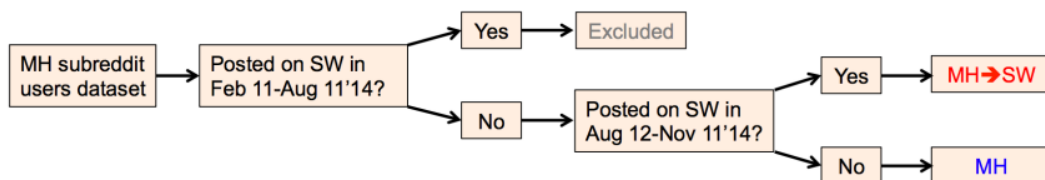
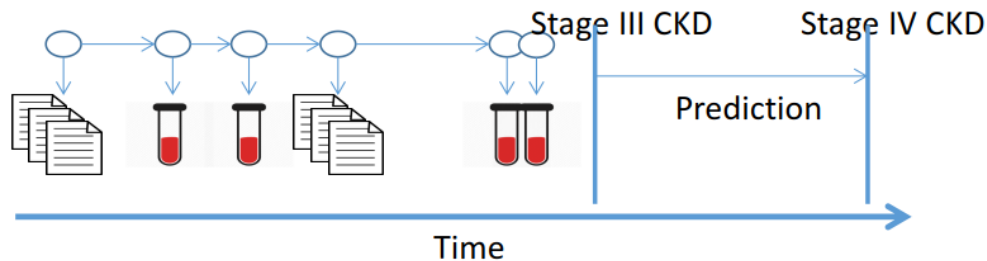


Figure 1: Schematic diagram of obtaining MH → SW and MH classes of users.

	MH	MH → SW	z	p
Linguistic Structure				
nouns	0.294	0.125	6.51	***
verbs	0.045	0.107	2.19	**
abverbs	0.048	0.099	4.87	***
readability index	0.609	0.232	5.51	***
accommodation	0.857	0.487	5.46	**
Interpersonal Awareness				
1st person singular	0.018	0.086	-10.6	***
1st person plural	0.093	0.078	4.53	*
2nd person	0.058	0.031	8.01	*
3rd person	0.087	0.042	6.32	***
Interaction				
posts authored	18.97	10.31	2.53	*
post length	215.62	443.73	-15.4	***
comments authored	122.42	106.22	0.95	-
comments received	19.862	13.414	1.05	*
comment length authored	63.417	87.116	-1.88	*
comment length received	42.323	26.362	5.44	**
response velocity (mins)	7.746	6.966	0.84	-
vote difference	28.788	7.681	7.18	***

Table 2: Differences between MH → SW and MH user classes based on linguistic structural, interpersonal awareness and interaction measures. Statistical significance is reported based on Wilcoxon signed rank tests at levels $p = .05/N; .01/N; .001/N$, ($N = 17$), following Bonferroni correction.

Application: predictive analytics



Survival Model (n=2,617)	Concordance (n=291)
(Text + Lab) Kalman Filter	0.849
Lab Kalman Filter	0.836
Recent Labs	0.819
Text Kalman Filter	0.733
eGFR risk score	0.779

(Heart Failure)

Lasix
Volume
Edema
Heart
Failure
Worsening
Diuresis
Severe
Diastolic
Overload

(Diabetes)

Units
Insulin
Subcutaneous
Lantus
Glucose
Diabetes
Times
70/30
Diabetic
Days

(Dialysis)

q15
Dialysis
Fistula
Volume
Bid
Lasix
Placement
Improved
Heparin
Examined

(Health Maintenance)

Flu
Visit
Fasting
Colonoscopy
Year
Shot
Vaccine
wnl
Check
Primary

(Gynecological)

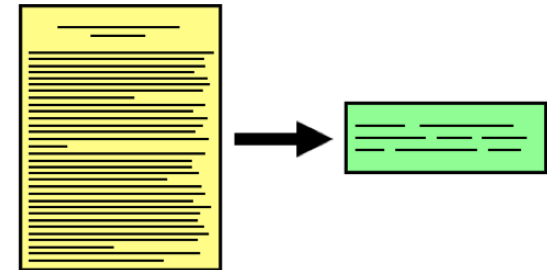
Breast
Vaginal
Mammo
Cancer
Hx
Pap
nl
Age
will
Endometrial

(Asthma)

Albuterol
Asthma
Inhaled
Lung
Obstructive
Wheezing
Advair
Pulm
Restrictive
Puffs

Clinical Notes Use Cases

- **Text Summarization** - There are two different use cases of text summarization.
 - One is to accelerate the chart review process so that clinical information in multiple reports of the same patient can be extracted and visualized so that clinicians can quickly grasp major past medical history of a patient without performing lengthy chart reviews.
 - The second use case is to summarize clinical information across a group of patients enabled using NLP techniques. The new annotation process will be able to group and summarize annotation text hits at both patient and document levels.



Biomedical Documents



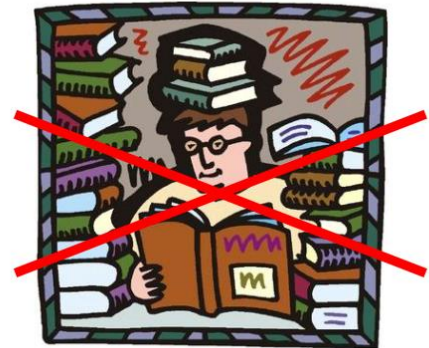
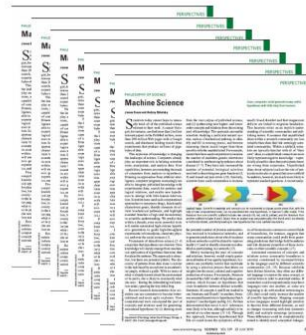
Grows in size at a dramatic pace

PubMed

27 million abstracts

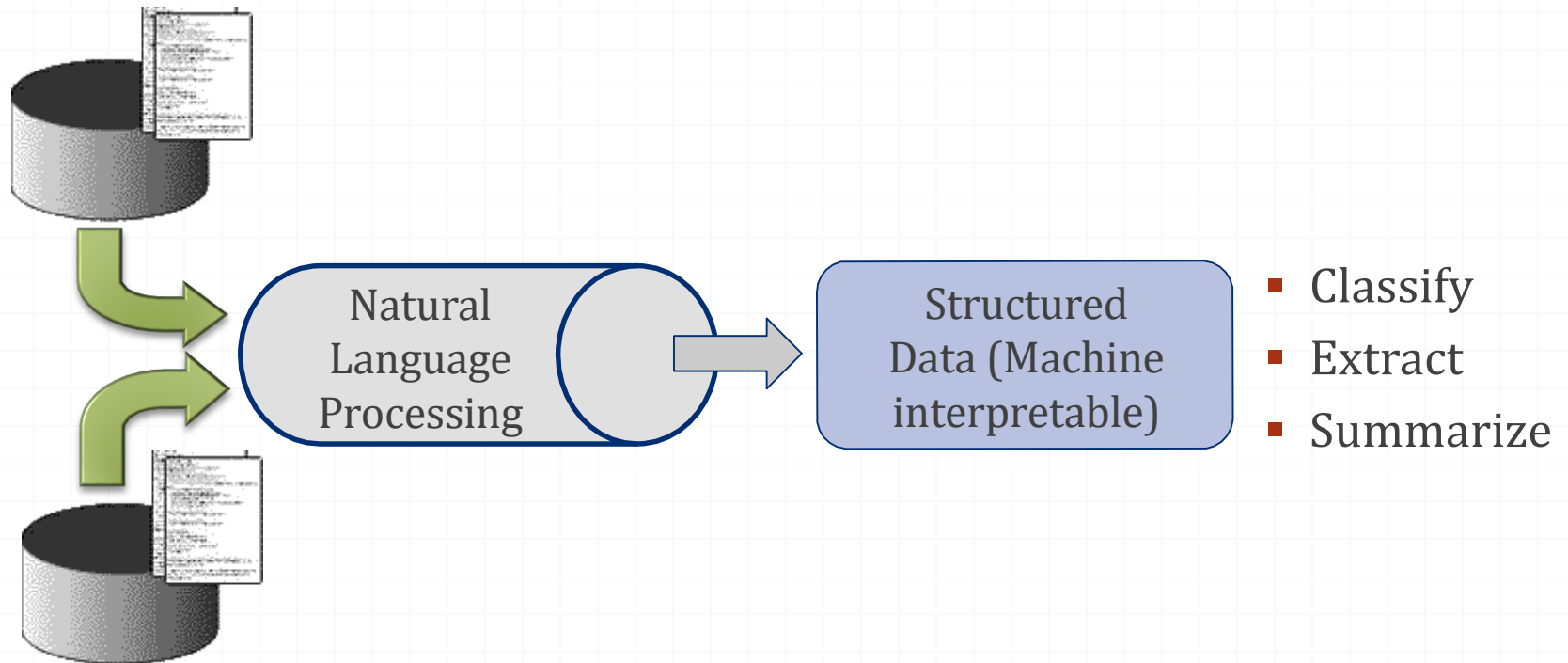
Two new abstracts every minute

Adds over one million every year



Clinical & Biomedical NLP

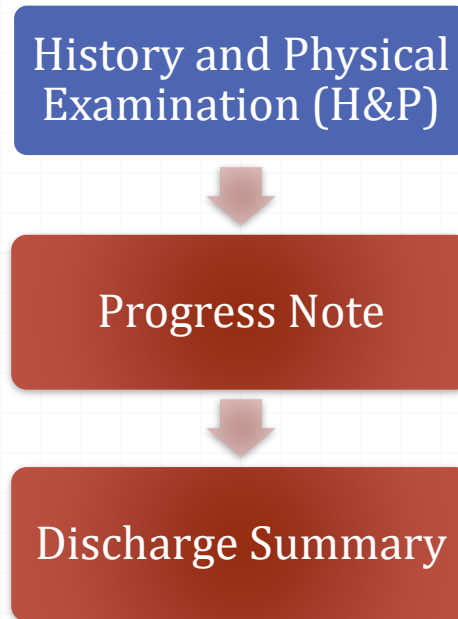
Clinical NLP: Electronic Medical Records



BioNLP: PubMed Articles

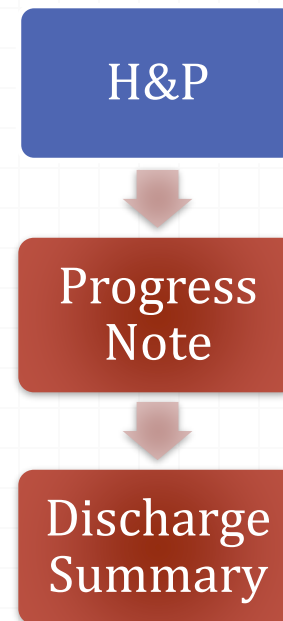
Closer Look at Clinical Notes

Clinical Documentation

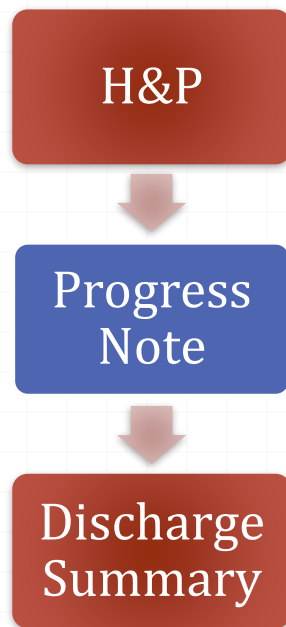


Clinical Documentation: H&P

- History & Physical:
 - Chief Complaint
 - History of Present Illness
 - Past Medical, Surgical, Psychiatric Histories
 - Family & Social History
 - Medications and Allergies
 - Vitals & Labs & Imaging
 - Review of Systems
 - Problem List
 - Assessment & Plan

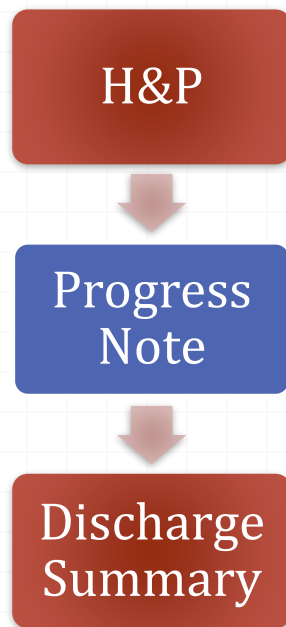


Clinical Documentation: Progress Note



- SOAP
 - Subjective: Qualitative narrative
 - Objective: Vitals, Exam, Labs, Imaging, etc.
 - Assessment: What does it all suggest?
 - Plan: Next steps.

Clinical Documentation: Discharge Summary



Final Diagnosis:

Procedures:

History of Present Illness

Laboratory/Data Hospital Course (by
PROBLEM LIST.... NOT BY DATE ---)

Discharge Medications

Discharge Instructions (diet, activity,
discharged to home/nursing facility, etc)

NLP Challenges

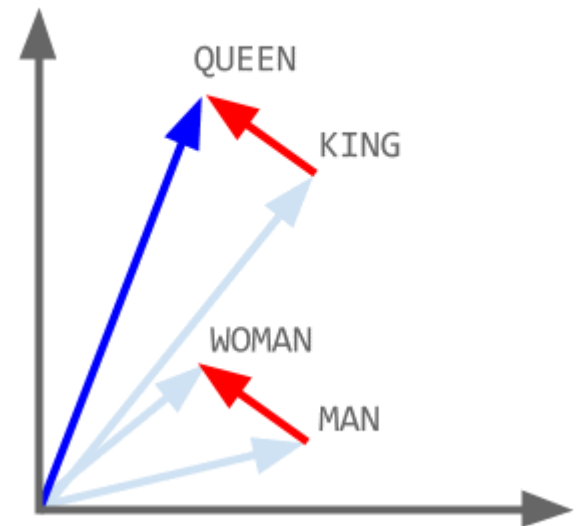
Major Elements

1. Words
2. Syntax
3. Meaning
4. Discourse

From ML to NLP

- Features from words and sentences?
 - Bag of words
 - TF-IDF
 - N-grams
 - Word vectors: word2vec, doc2vec
- Models?
 - Linear?
 - Deep learning?

So $\text{king} + \text{man} - \text{woman} = \text{queen!}$



Common NLP Tasks

- **Co-reference resolution**

- Given a sentence, determine which words refer to the same objects:
He entered **John's house** through the **front door**

- **Named entity recognition (NER)**

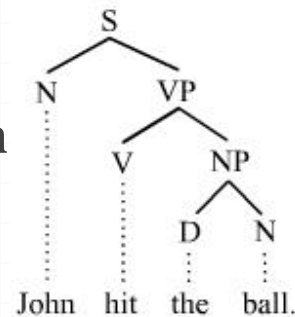
- Given a sentence, determine which items map to proper names, such as people or places. **Mr. Jones, Epilepsy**

- **Part-of-Speech tagging (POS)**

- Given a sentence, determine the part of speech for each word.
Discharge (verb/Noun)

- **Parsing**

- Determine the parse tree (grammatical analysis) of a given sentence.



Common NLP Tasks

- **Relationship extraction**
 - Given a sentence, identify the relationships among named entities (e.g. *what is inhibiting expression of what*).
- **Word sense disambiguation**
 - Many words have more than one meaning (e.g. *Discharge*)
- **Speech processing**
 - This covers speech recognition, text-to-speech and related tasks.
- **Discourse Parsing**
 - Identifying the discourse structure of connected text (e.g. elaboration, contrast), or classifying the speech acts (e.g. assertion, question).

Clinical vs. Biomedical NLP

- **Clinical NLP**

- **Task:** Focus on documentation related to patient care
- **Data source:** Electronic Health Records (EHR)
 - Both inpatient and outpatient documentation

- **Biomedical NLP**

- **Task:** Focus on scientific discoveries about biology, physiology, and medicine
- **Data source:** Journal articles, clinical trials, webpages, ...

Example: Classification

Clinical Examples

- **Classify**
- Extract
- Summarize

Classify a chief complaint into
a syndrome category

“coughing” to Respiratory

Biomedical Examples

- **Classify**
- Extract
- Summarize

Triage of articles likely to have
experimental evidence

Find evidence to assign top-level
GO (Gene Ontology) terms

Example: Extraction

Clinical Examples

- Classify
- **Extract**
- Summarize

of lymph nodes removed
during colorectal cancer
surgery

Biomedical Examples

- Classify
- **Extract**
- Summarize

#Extract bio-molecular events
What [ANTIBODIES] have been
used to detect protein TLR4?

Example: Summarization

Clinical Examples

- Classify
- Extract
- Summarize

From a H&P note, list chronic condition

Summarize family history of prostate cancer

Biomedical Examples

- Classify
- Extract
- Summarize

Summarize full text documents

Gene Reference into function
(GeneRef)

Biomedical Language

- Contains domain specific rich and evolving vocabulary
- Journal abstracts and articles usually follow similar section structure
- Sentences are very grammatical but include highly ambiguous terms
 - Neurofibromatosis 2 [disease]
 - Neurofibromin 2 [protein]
 - Neurofibromatosis 2 gene [gene]

Clinical Language

- Domain-specific, jargon, idioms
- Telegraphic, with misspellings, incomplete sentences
- Speculations, hypotheses, and negations
- Some structure

Sample Clinical Documents



HISTORY OF PRESENT ILLNESS: Mrs. [Hun] is a 77 year-woman with long standing **hypertension** who presented as a Walk-in to me at the [Bronx] Health Center on September 15. **Recently** had been started **q.o.d.** on Clonidine since August 2 to taper off of the drug. **Was** told to start Zestril 20 **mg. q.d.** again. The patient was sent to the Emergency Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [Swasissz] to follow.

SOCIAL HISTORY: **Lives** alone, **has** one daughter living in [Spring]. **Is** a non-smoker, and **does** not drink alcohol.

HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [Tylenol], was started on **IV** Heparin, Sotalol 40 **mg PO b.i.d.** increased to 80 **mg b.i.d.**, and had an echocardiogram. By the next day the patient had better rate control and blood pressure control but remained in atrial fibrillation. On Saturday the patient was felt to be medically stable...

Sample Clinical Documents

The patient is a 46 year old woman with a history of **Q wave myocardial infarction** with right ventricular **infarct** in October 1992. Peak **CK's** were 2300. Catheterization showed 100% **RCA** lesion which was treated with angioplasty **reduced** to 20-30% stenosis. Subsequent catheterization October 92 , July 92 and September 92 for atypical chest pain , showed clean coronaries. **Exercise tread mill test in September 92** , the patient went three minutes and 31 seconds with standard Bruce protocol and stopped secondary to atypical chest pain. **Maximum heart rate 162 , blood pressure 176/90 , no ST or T wave changes**. In April 92 she ruled out for myocardial infarction **by enzymes** and EKG , after presenting with prolonged chest pain. **VQ scan was low probability**. Chest CT ruled out aortic dissection. The patient now presents to the hospital with 24 hours of right sided chest pain , stating that it was squeezing in her right breast , **felt to be between the shoulder blades**. She complained of shortness of breath , dizziness , weakness and nausea , **no palpitations were noted...**



Sample Clinical Documents

Pt recently hospitalized 7/19/06 for **chf** exacerbation (diastolic dysfunction) **2nd** to dietary and medicine noncompliance (salty foods , stopped her **HCTZ**) and continued to smoke. **Pt diuresed** and sent home on new **lasix 60qam 40qpm regimen**. **Pt** noticed steady decline in functional status during the last 3 weeks because of **SOB**. **at** baseline should **sat** 85% on **ra** , 95% on **6L02NC** at rest and ambulation. (on home **o2**) but now , can't ambulate , **sating** 83-89% on 6l at rest. **also** notes **pnd** , orthopnea. **Pt** notes intermif ent chest pain on and off lasting 5 minutes not associated with exertion or any other cardiac **sx**. 8/15 dobuta mibi→ ischemia in d1 territory. **11/19 :echo→ef 60% , Pa pressure 48 + RA. no valve dz. rv enlarged and hypokinetic. A/P: pump: decompesated CHF (diastolic dysfxn , ? cor pulmonale component) 2nd to diet/med non-compliance. uptitrate captopril , continue iv lasix 60 qd with goal net neg 2 liters , daily weights , strict land O. check cxray. Switched to po lasix 10/06 , back to lisinopril for d/c Fri. ischemia: has + mibi in past , but no further workup to d1 lesion. can't get ecasa 2nd to vWD. continue BB , will hold off on statin since not hyperlipidemic. rate:tele. ...**



Why is NLP Difficult?

Named entity recognition

Linguistic variation
Polysemy Finding
validation
Implication

Contextual attribute
assignment

Negation
Uncertainty
Temporality

Discourse processing

Report structure
Coreference

Linguistic Variation

Different Words with the Same Meaning

Derivation

mediastinal = mediastinum

Inflection

opacity = opacities; cough = coughed

Synonymy

Addison's Disease: Addison melanoderma, adrenal insufficiency, adrenocortical insufficiency, asthenia pigemntosa, bronzed disease, melasma addisonii, ...

Chest wall tenderness: chest wall did demonstrate some slight tenderness when the patient had pressure applied to the right side of the thoracic cage

Polysemy

One Word With Multiple Meanings

General polysemy

Patient was prescribed codeine upon discharge

The discharge was yellow and purulent

Acronyms and Abbreviations

APC: activated protein c, adenomatosis polyposis coli, adenomatous polyposis coli, antigen presenting cell, aerobic plate count, advanced pancreatic cancer, age period cohort, alfalfa protein concentrated, allophycocyanin, anaphase promoting complex, anoxic preconditioning, anterior piriform cortex, antibody producing cells, atrial premature complex, ...

Negation

Approximately half of all clinical concepts in dictated reports are negated*

Explicit negation

“The mediastinum is not widened”

Mediastinal widening: absent

Implied absence without negation

“Lungs are clear upon auscultation”

Rales/crackles:

absent

Rhonchi: absent

Wheezing: absent

Uncertainty

Unsure

treated for a **presumptive** sinusitis

Reasoning

It was felt that the patient **probably** had a cerebrovascular accident involving the left side of the brain. Other differentials entertained were **perhaps seizure** and the patient being postictal when he was found, although this consideration is **less likely**.

Reason for exam

R/O out pneumonia.

Temporality

Clinical reports tell a story

Past medical history

History of CHF presenting with shortness of left-sided chest pain.

Hypothetical or non-specific mentions

He should return for fever or increased shortness of breath.

Temporal course of disease

Patient presents with chest pain ... After administration of nitroglycerin, the chest pain resolved.

Finding Validation

Mention of a finding in the text does not guarantee
the patient has the finding

She received her **influenza** vaccine

His **temperature** was taken in the ED

Some findings require values

Fever

Temperature 38.5C

Oxygen desaturation

Oxygen saturation low

Oxygen saturation 85% on room air

Implication

Audience for patient reports is physicians

- Pneumonia not mentioned in many positive reports
- Sentence level inference
 - “There were hazy opacities in the lower lobes” -> Localized infiltrate
- Report level inference
 - Localized infiltrates -> Probable pneumonia

Report Structure

- Anatomic Location sometimes in section header
NECK: no adenopathy.

- Some sections carry more weight

IMPRESSION: atelectasis

- Some reports contain pasted text difficult to process

Cardiovascular: [] Angina [] MI [x] HTN [] CHF [] PVD [] DVT
[] Arrhythmias [] Previous PTCA [] Previous Cardiac Surgery []
] Negative -Denies CV problems

Coreference

Chest x-ray again shows a well-circumscribed nodule located in the left upper lobe. The tumor has increased in size since the last exam with a diameter of approximately 2 cm.

- How big is the nodule?
- Has the nodule increased in size?
- Where is the tumor?

Sample Clinical NLP Application

- Does the patient drink?
 - Classify patient into 3 classes: Heavy consumption, Moderate consumption, None
- Application:
 - Retrospective cohort study as ICU mortality predictor
- Structured data:
 - ICD9 codes of alcohol-based illness, e.g., delirium tremens (291*)

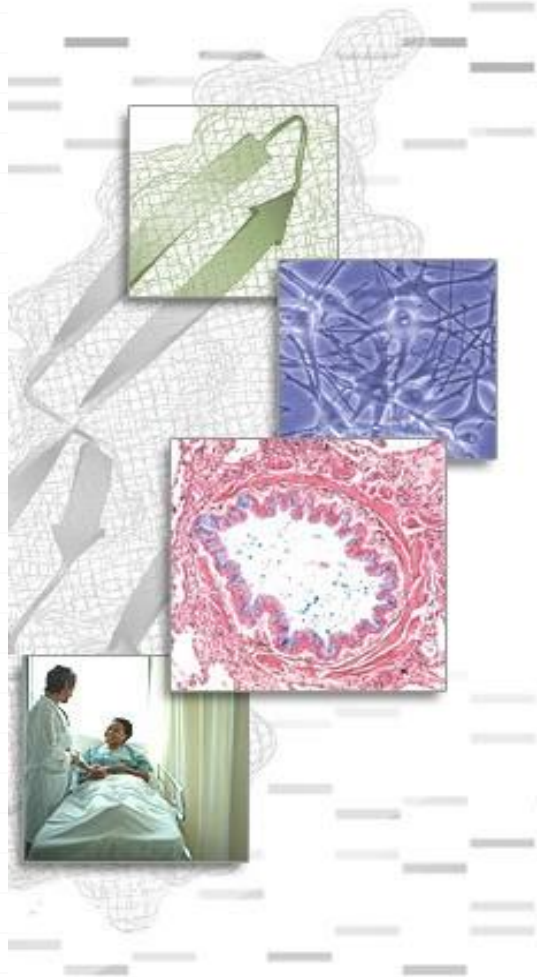


Sample Clinical Documents

- Search for alcohol OR drink OR...
- Retrieval results:
 - Yes: The patient is known to have a history of alcohol abuse
 - Maybe: tox screen only significant for alcohol level of 273
 - No: She denied any intravenous drug use, tobacco, or alcohol
 - Maybe: He was counseled against the use of alcohol
 - Maybe: He stated that he would not drink alcohol in the future
 - No: Major Surgical or Invasive Procedure: Alcohol septal ablation.
 - Yes: Quit smoking 'many' yrs ago. smoked 1 pp week. used to drink 10-12 hard drinks every other day. last drink before go to OSH. resolved to quit now. no IVDU.
 - No: her husband is a heavy drinker and often isolates himself to drink alone
 - No: the patient was seen to drink a tremendous amount of water



Clinical NLP Tasks



- Does the patient drink alcohol or not?
 - **Negation detection**
- Is alcohol use asserted, equivocal, modal or hypothetical?
 - **Uncertainty**
- Who is drinking?
 - **Experiencer, relation extraction**
- What is the patient drinking?
 - **Relation, event extraction**
- Is the patient drinking currently and regularly?
 - **Temporal relations, timeline generation**

More Clinical NLP Tasks

- Low-level tasks
 - Sentence boundary detection
 - Part of speech tagging
 - Shallow parsing (chunking)
- High-level Tasks
 - Spelling/grammatical error identification and recovery
 - Named entity recognition and information extraction
 - Word sense disambiguation



Challenges in Clinical NLP Tasks

- Sentence boundary detection: complicated by
 - abbreviations and titles, e.g., m.g., Dr.
 - lists and templates, e.g., MI [x], SOB[]
- Tokenization: complicated by
 - characters typically used as token boundaries, e.g.,
 - 10 mg/day, N--acetylcysteine.
- Morphological decomposition: complicated by
 - compound words, e.g., nasogastric
- Text segmentation: complicated by
 - problem-specific needs, e.g., sections, including Chief Complaint, Past Medical History, HEENT, etc.
- In general, systems developed for non-clinical text often work less well on clinical narratives.



Biomedical and Clinical Corpora

- There are various biomedical corpora annotated for syntax and semantics
 - [MedTag](#): A collection of biomedical annotations (MEDLINE abstracts): the AbGene corpus of annotated sentences of genes and protein named entities, the MedPost corpus of part of speech tagged sentences and the GENETAG corpus for named entity identification used for BioCreative I.
 - [TREC Genomics Track](#): A set of data collections provided by TREC Genomics Track useful for development and evaluation of retrieval and text categorization strategies in the biomedical domain.
 - [BioCreative corpus](#): Dataset produced by the BioCreative assessment, text passages relevant for GO annotations of human proteins.
 - [GENIA corpus](#): Annotated corpus of literature related to the MeSH terms: Human, Blood Cells, and Transcription Factors.
 - [Yapex corpus](#): Training and test data for the protein tagger (NER) YAPEX.
 - [PASBio](#): Predicate-argument structures of biomedical literature.
 - [LLti5 dataset](#): Genic Interaction Extraction Challenge: protein/gene interactions IE data set
 - [IEPA corpus](#): The Interaction Extraction Performance Assessment corpus
 - [BioText Data](#): Dataset for extraction of disease/treatment entities relations
 - [BioText NC Semantics Dataset](#): Dataset of Noun Compound Semantics used in experiments described in articles
 - [PennBioIE](#): UPenn Biomedical Information Extraction datasets of annotated PubMed abstracts: CYP45ti domain and oncology domain
 - [Medstract corpus](#): Biomedical annotation corpus useful for acronym definition and coreference resolution
 - [Medstract corpus](#): Biomedical annotation corpus useful for acronym definition and coreference resolution
 - [OHSUMED text collection](#): Document collection used for the TREC-9 contest.
 - [BMC corpus](#): Open access corpus of full text articles provided by BioMed Central.
 - [FetchProt corpus](#): Full text journal articles from the biological domain analyzed for experiments on proteins.
 - [PDG Bio-sentence splitting corpus](#): Small collection of text data sets derived from PubMed abstracts to develop and assess sentence splitting tools.
 - [Bio1 corpus](#): annotated corpus, same field as GENIA, but annotated to small top-level ontology.
 - ...