NLP for Healthcare in the Absence of a Healthcare Dataset

Aditya Joshi

Based on ALTA Tutorial by Sarvnaz Karimi & Aditya Joshi

18th January, 2020

Talk Scope

NLP for Healthcare in the Absence of a Healthcare Dataset

‘In the absence of a healthcare dataset’

What is a health data?

Health data is defined as “specific information about physical or mental health of an individual/data subject.”
(https://www.lexology.com/library/detail.aspx?g=73242896-04bd-456b-b031-344f3c22bb95)

Health data is defined as “being data that is associated with the provision of healthcare services to an individual.”
(https://www.csoonline.com/article/3308877/can-digital-identity-cure-the-chronically-ill.html)

Health data is defined as “personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status” (EU General Data Protection Regulation (GDPR), 2016)

So what is a healthcare dataset?

Based on the definition for ‘health data’ by (EU GDPR, 2016)

“personal data related to the provision of health care services, which reveal information about his or her health status”.

Source of healthcare data: Healthcare agencies providing services to individuals

Why are we assuming the absence of a healthcare dataset?

Some datasets are public.

Yes. Many are structured datasets.

Some datasets are anonymised as well.

Yes. MIMIC, for example. They are useful and widely used.

BUT…

- Accessibility is limited
- Confidentiality and Privacy
- May not be complete (Not every patient goes to a hospital)
- Latency between different sources
- Maturity of healthcare systems are different

Are ‘non’-healthcare datasets any good?

- White et al. (2013) look at correlations between Internet searches for drugs and adverse events. There was clear evidence of interaction of paroxetine and pravastatin causing hyperglycemia.

- Ranney et al (2016) show a statistically significant relationship between the counts of alcohol-related tweets and counts of emergency care visits, both originating from a small north-Eastern state of the US.

- Joshi et al (2019) show that social media-based epidemic surveillance could help early detection of an asthma outbreak in Melbourne.

White, R. & Shah, Nigam & Altman, Russ & Horvitz, Eric. (2013). Web-scale pharmacovigilance: Listening to signals from the crowd. JAMIA.
Ranney, Megan L., Brian Chang, Joshua R. Freeman, Brian Norris, Mark Silverberg, and Esther K. Choo. "Tweet now, see you in the ED later? Examining the association between alcohol-related tweets and emergency care visits." *Academic Emergency Medicine* 23, no. 7 (2016): 831-834.
Joshi, Aditya, Ross Sparks, James McHugh, Sarvnaz Karimi, Cecile Paris, C Raina MacIntyre. “Harnessing Tweets for Early Detection of an Acute Disease Event.” Epidemiology, 2019.

Tutorial Objective

To show how NLP techniques can help healthcare *even* without using (in the absence of) healthcare datasets.

Advantages:

1. Potentially more real-time
2. Potentially deployable
3. Potentially applicable where digital healthcare systems are not sufficiently mature

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

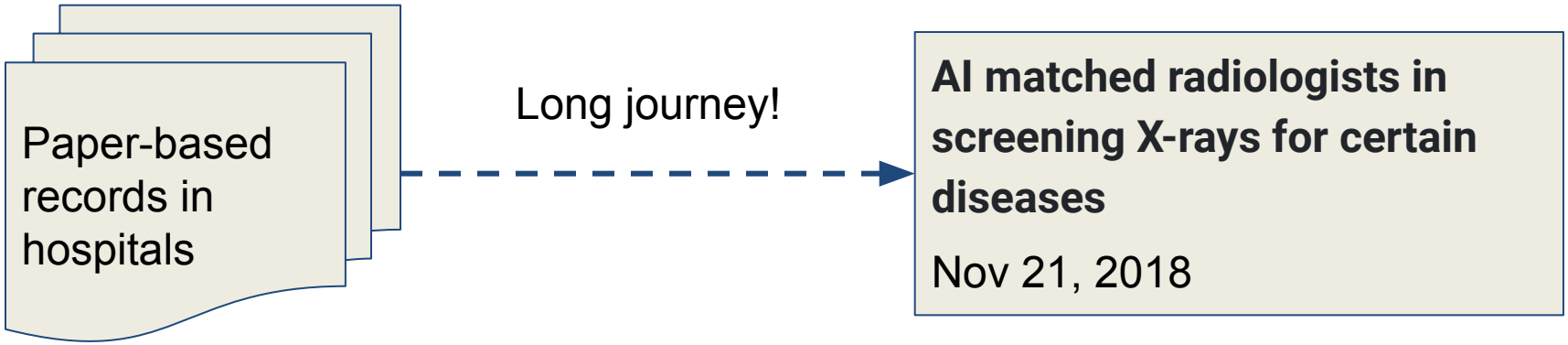
Image of coffee from wikimedia commons.

Module I: Introduction

Outline

- History
- Challenges
- Scope

Information Technology in Healthcare



Clinical Decision Support Systems (CDSS)

"Clinical decision support systems link health observations with health knowledge to influence health choices by clinicians for improved health care"

CDSS is to assist clinicians at the point of care.

GELLO Expression language to query and build CDSS on the top of health records

Early CDSS' are criticised for their 'Greek Oracle' approach. The human expert only becomes a data entry operator with no expertise of their own.

Clancey, William J., and Edward H. Shortliffe. *Readings in medical artificial intelligence: the first decade*. Addison-Wesley Longman Publishing Co., Inc., 1984.

MYCIN

Early 1970s
Stanford University
A medical diagnosis system
Questions regarding symptoms
600 rules that return a ranked list of likely bacterial infections

In practice it never actually got used!

INTERNIST-1

Computer-assisted diagnostic tool
University of Pittsburgh
Asks a bunch of questions
The medical practitioner only enters answers to the questions posed by the system. The system presents its results
Does not work well if the patient has more than one diseases

Randolph A. Miller, et al., “INTERNIST-1: An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine,” New England Journal of Medicine 307 (August 19, 1982): 468-76.

History of healthcare and NLP: Pre-1999

Computerized medication monitoring system (1976)
Sager: NLP system also applied to medical documents (1981)
SPRUS: Radiology text-processor (1994)
SYMTEXT: Automatically generate codes for admission diagnoses (1995)
MedLEE: First NLP system used for actual patient care (1994)
Geneva Hospital: French, English and German documents: Discharge summaries of patients admitted for gastrointestinal surgery (1992-93)
MENELAS: Accessing patient discharge summaries using search (1994)

Friedman, Carol, and George Hripcsak. "Natural language processing and its future in medicine." *Acad Med* 74, no. 8 (1999): 890-5.

The paper (written in 1999) also states...

NLP is likely to become more important shortly because of healthcare economics

1) Web-based technology is becoming pervasive
2) Continuous voice recognition
3) Likely produce standard terminology and output forms, due to initiatives such as UMLS

Will Doctors Be Replaced by Artificial Intelligence? (Sept 2018)

https://blog.netapp.com/is-ai-going-to-replace-doctors/

Friedman, Carol, and George Hripcsak. "Natural language processing and its future in medicine." *Acad Med* 74, no. 8 (1999): 890-5.

Digitisation of Patient Records

'Computer-based patient record (CPR)' is important for two reasons:

1. Evolving role of the primary care: Since the first point of contact determines which tests need to be done, etc., CPRs can empower them
2. The Integrated delivery system: Information systems that combine health care providers, service providers and other facilities

What are the hallmarks of 'computer'-isation of patient records?

1. Integrated view of patient data
2. Access to knowledge resources
3. Physician order entry and clinician data entry
4. Integrated communications support
5. Clinical decision support

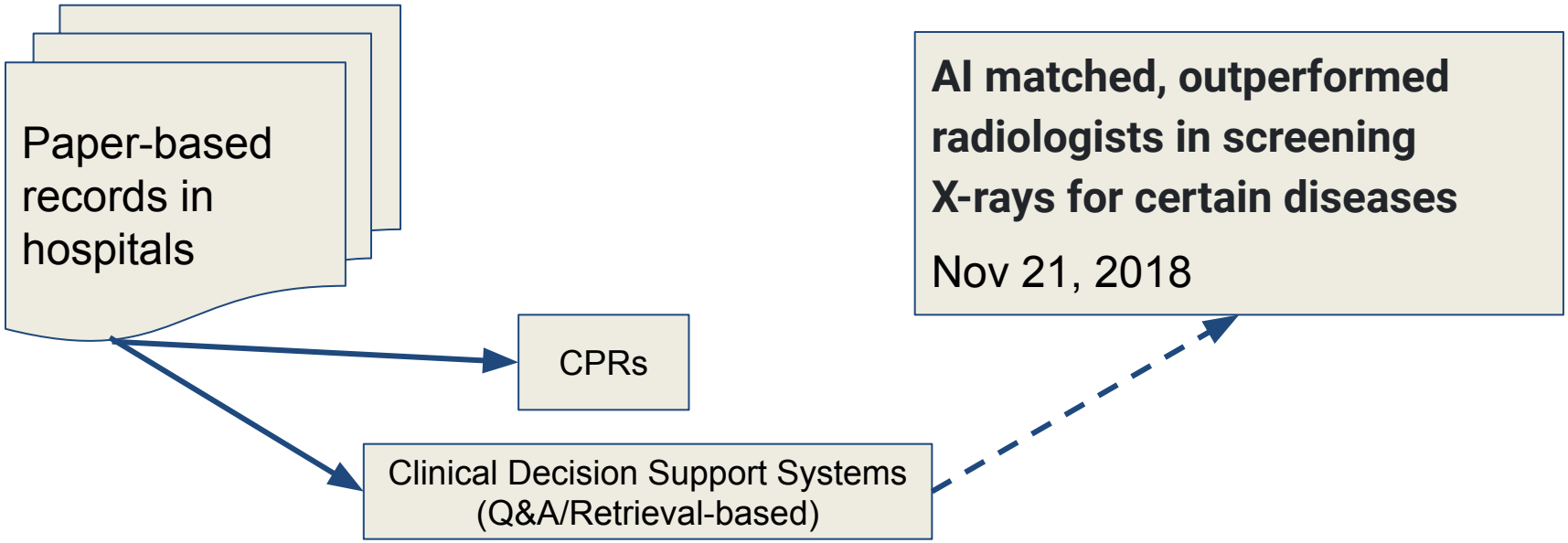
Dick, Richard S., Elaine B. Steen, and Don E. Detmer, eds. *The computer-based patient record: an essential technology for health care*. National Academies Press, 1997.

Reflecting on the hallmarks from the AI perspective

1. Integrated view of patient data: Information retrieval approaches for different forms of data
2. Access to knowledge resources: Standardised medical ontologies
3. Physician order entry and clinician data entry
4. Integrated communications support
5. Clinical decision support: Several AI approaches on different kinds of medical data

Dick, Richard S., Elaine B. Steen, and Don E. Detmer, eds. *The computer-based patient record: an essential technology for health care*. National Academies Press, 1997.

History: Revisited



Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module II: Datasets

Outline

- Sources
- Annotation
- Ethics

All cliparts in this module are from wikimedia commons.

NLP for Healthcare in the Absence of a Healthcare Dataset

Absence of a Healthcare Dataset

Non-Healthcare Datasets

Where are they obtained from?
How are they labeled?
What ethical concerns may be relevant?

Non-Healthcare Datasets

Where are they obtained from?
How are they labeled?
What ethical concerns may be relevant?

Types of sources

(Velardi et al., 2014)

1) **Demand-based data sources**: This refers to sources that reflect demand for information.
   a) Pros: Aggregate information
   b) Cons: Access may be restricted; limited context
2) **Supply-based data sources**: The data originates on large-scale platforms designed to share information.
   a) Pros: Large-scale information
   b) Cons: The text tends to be longer than search queries, bringing in typical challenges of ambiguity in NLP

Paola Velardi, Giovanni Stilo, Alberto E Tozzi, and Francesco Gesualdo. 2014. Twitter mining for fine-grained syndromic surveillance. Artificial Intelligence in Medicine 61, 3 (2014), 153–163.

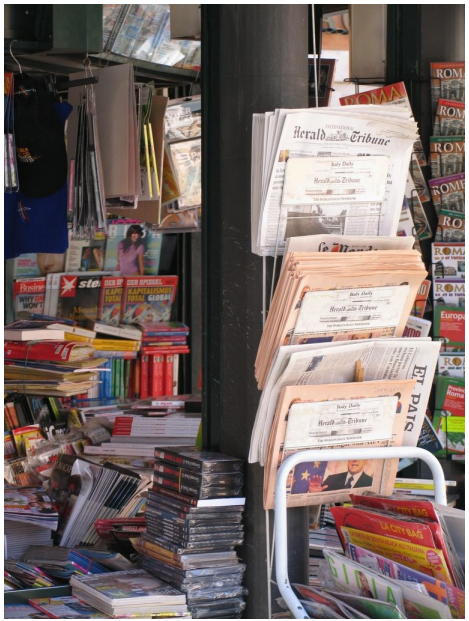
Search Queries

Anonymous search engine query data provides insights into users' beliefs, behaviors, and interests. (Hulth et al., 2009). Google Flu Trends (Ginsberg et al., 2009): Volumes of queries on Google to detect disease outbreaks. Top queries from states of the US.

Advantages: Large-scale, potentially anonymous data
Disadvantages: Access

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web queries as a source for syndromic surveillance. PloS one 4, 2 (2009), e4378
Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457, 7232 (2009), 1012.

News Articles



News articles may contain health-related news. Early work in epidemic intelligence (#PreSocialMediaEra)
Health Map uses news reports to monitor diseases (Freifeld et al, 2008).
Lejeune et al. (2010) use news articles from multilingual sources.

Advantages: Formal, likely to be well-written.
Disadvantages: Duplication, trustworthiness of sources

Gaël Lejeune, Antoine Doucet, Roman Yangarber, and Nadine Lucas. 2010. Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In 4th International workshop on cross-lingual information access (Beijing, China). 8–pages.
Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein. 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. Journal of the American Medical Informatics Association 15, 2 (2008), 150–157.

Social Media

People may report health conditions and health-related issues on social media. Social media-based epidemic intelligence is a popular area of research.

Advantages: Targeted health information without the legal and technical obstacles that exist for other sources such as official records (Yepes et al., 2015)
Disadvantages: May not be as reliable as official records. (Adam et al., 2017)

Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using twitter. In Proceedings of Workshop on Biomedical Natural Language Processing (Beijing, China). 164–170.
Dillon C Adam, Jitendra Jonnagaddala, Daniel Han-Chen, Sean Batongbacal, Luan Almeida, Jing Z Zhu, Jenny J Yang, Jumail M Mundekkat, Steven Badman, Abrar Chughtai, et al. 2017. ZikaHack 2016: A digital disease detection competition. In Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM2017). 39–46.

Discussion Forum Posts

Medical discussion forums allow users to create discussions around topics. Discussion forums may be general or specific (mumsnet, AskaPatient, IVF) (Sokolova et al, 2013)

Advantages: Topic-specific text, opportunities to understand awareness/impact/emotional implications, etc., possible anonymity
Disadvantages: May not be real-time. Topic Drift. Information may not be reliable since it is patients/families supporting each other.

Sokolova, Marina, and Victoria Bobicev. "What Sentiments Can Be Found in Medical Forums?." In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pp. 633-639. 2013.

Non-Healthcare Datasets

Where are they obtained from?
How are they labeled?
What ethical concerns may be relevant?

Annotation

The process of assigning supplementary information
Why?

I have been coughing all morning, just took my cough syrup

Personal health mention classification: **True** False

Drug mention extraction: I have been coughing all morning, just took my [cough syrup]

Symptom Normalisation: I have been coughing[R05] all morning, just took my cough syrup

Typical Annotation Strategies

Manual: Human annotators are given a set of guidelines and questions

Hybrid: A combination of manual and automatic steps

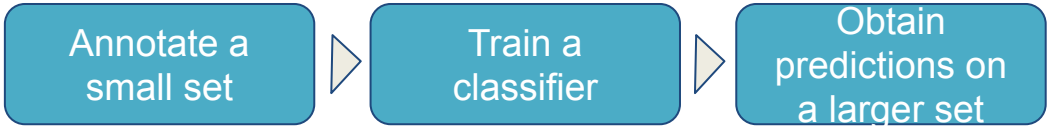
Manual Annotation

Guidelines to annotators are key.
For example,

Aramaki et al. (2011): Dataset of tweets with health mentions labeled
The annotator guidelines state that a tweet should be labeled positive if:
(a) one or more people with flu exist around the tweet author; and
(b) the tense is present or recent past.
The authors also mandate that the tweet should be affirmative and not speculative

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In Proceedings of the conference on Empirical methods in natural language processing (Beijing, China). Association for Computational Linguistics, 1568–1576.

Hybrid: Two-step



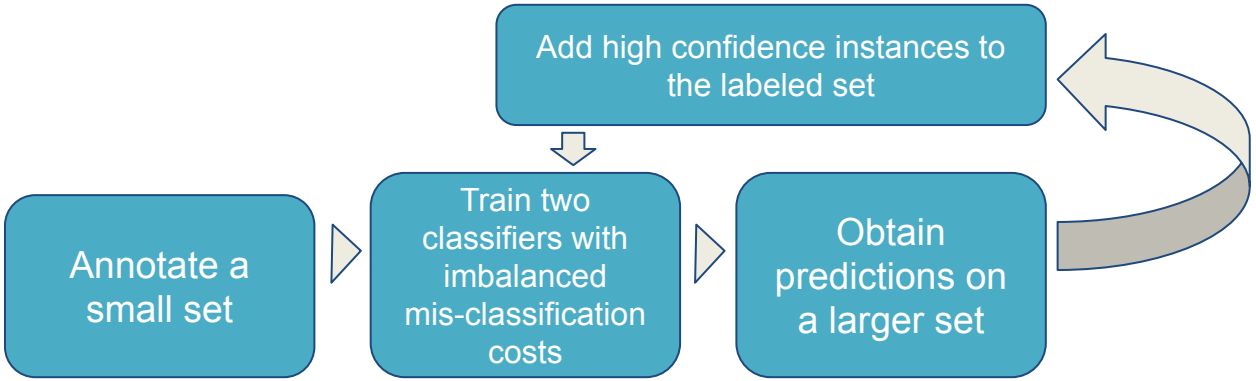
Paul and Dredze (2011&2012) download a set of 2 million tweets. A subset of 5,128 tweets are manually labeled. A classifier is trained on these manually labeled tweets and the predictions on the remaining tweets are obtained. These predictions are then used as labels for the tweets.

A topic model that uses these labels to create clusters; Therefore, a high-accuracy system is not required.

Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. International AAAI Conference on Web and Social Media 20, 265–272. [55]
Michael J Paul and Mark Dredze. 2012. A model for mining public health topics from Twitter. Health 11 (2012), 16–6

Hybrid: Iterative

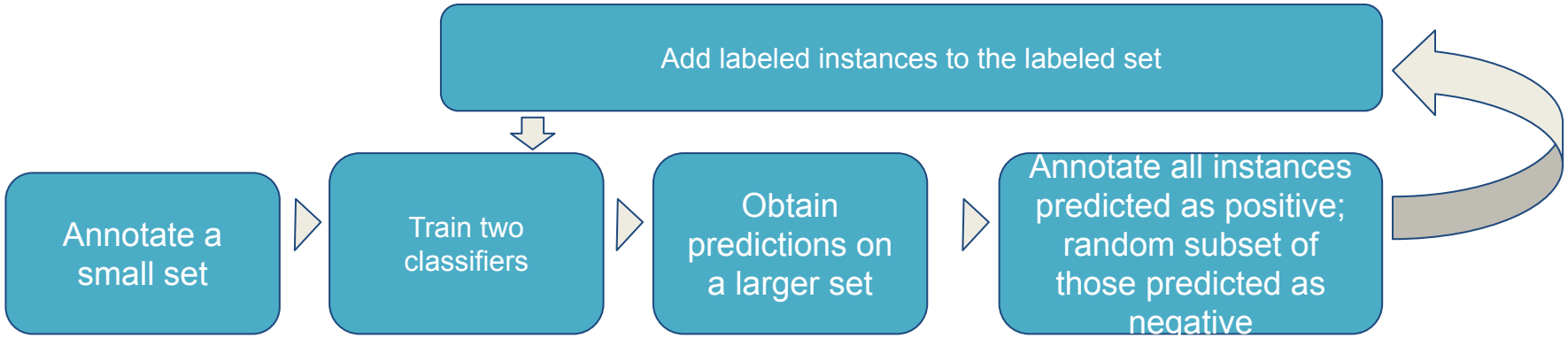
Sadilek et al. (2012):



Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Predicting disease transmission from geo-tagged micro-blog data.. In Conference on Artificial Intelligence (AAAI) (Toronto, Canada). 136–142.

Hybrid: Iterative

Jiang et al. (2016):



Keyuan Jiang, Ricardo Calix, and Matrika Gupta. 2016. Construction of a personal experience tweet corpus for health surveillance. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing (Berlin, Germany). 128–135.

Non-Healthcare Datasets

Where are they obtained from?
How are they labeled?
What ethical concerns may be relevant?

Confidentiality and Privacy in Medical Documents

Confidentiality refers to the obligation of professionals and other health workers not to disclose ***information***; codified in the Hippocratic Oath in the 4th century BC.

Privacy relates to protecting an ***individual's*** control over what personal information and decisions may or may not be shared with others.

Authorisation is the granting of permission to view confidential information.

Source: ICD-10 2016 document at https://www.who.int/classifications/icd/icdonlineversions/en/

Mapping to Non-Healthcare Data

Social stigma may adversely affect non-anonymised non-healthcare data

A twitter user posting a personal health condition may not understand the possible implications

'I pooped while sneezing' returns 12 results on twitter search (Access date: 25th November, 2019)

Anonymisation: Google Flu Trends gave specifications to anonymise

Typical Ethical Research Protocols

Benton et al (2017) discuss ethical research protocols for social media research for health. Prescribe:

- Institutional review boards
- Informed consent?
- In case of open-source data, there could be variations in what you are allowed or not allowed to do
- Statement of responsibility
- Separation of data from annotations using identifiers.
- Specific requirements of organisations

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing (Vancouver, Canada). 94–102.

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module III: Structured Knowledge Resources

Outline

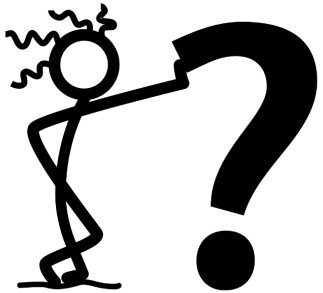
- ICD Codes
- Ontologies
- Using ICD Codes and Ontologies

Standardisation in healthcare

Digitisation of health records led to the need of standardisation:

- Multiple terms
- Understanding the hierarchy of diseases, the body parts they affect, etc.

Several repositories & ontologies have been created .. and used for healthcare applications of NLP



International Statistical Classification of Diseases and Related Health Problems (ICD)

A list of medical terms by the World Health Organisation (WHO)
Current version ICD-10
Used in 194 countries; first introduced in 1893

"To make people count, we first need to be able to count people."[1]

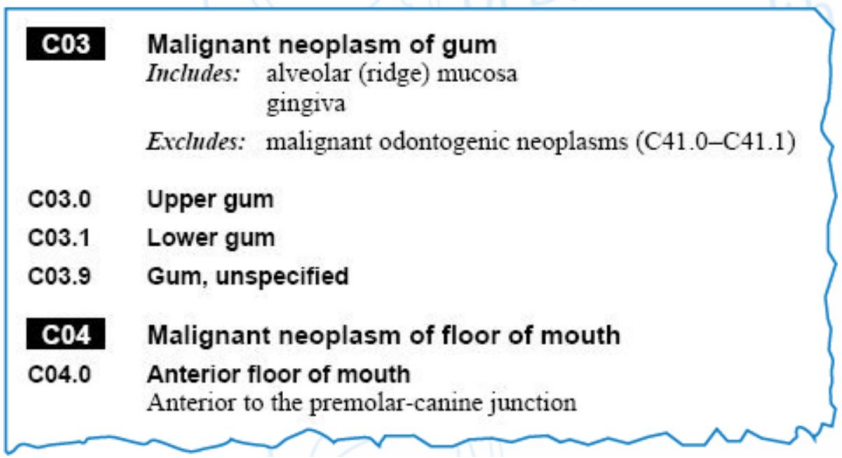
[1] http://apps.who.int/classifications/apps/icd/icd10training/ICD-10%20training/Start/index.html ; Accessed on 26th November, 2019

Groups of diseases

- Communicable diseases
- General diseases that affect the whole body
- Local diseases arranged by site
- Developmental diseases
- Injuries
- External causes

ICD Volumes

Volume 1: The Tabular List



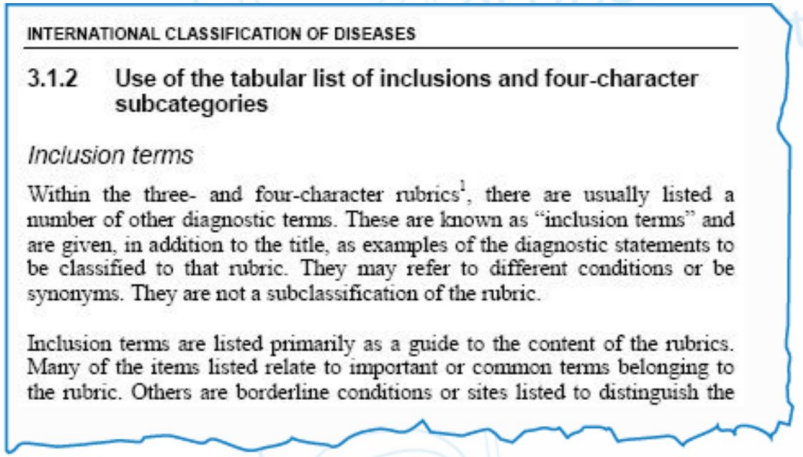
C03 **Malignant neoplasm of gum**
Includes: alveolar (ridge) mucosa
gingiva
Excludes: malignant odontogenic neoplasms (C41.0–C41.1)

C03.0 **Upper gum**
C03.1 **Lower gum**
C03.9 **Gum, unspecified**

C04 **Malignant neoplasm of floor of mouth**
C04.0 **Anterior floor of mouth**
Anterior to the premolar-canine junction

11400 character codes

Volume 2: Instruction manual



INTERNATIONAL CLASSIFICATION OF DISEASES

3.1.2 Use of the tabular list of inclusions and four-character subcategories

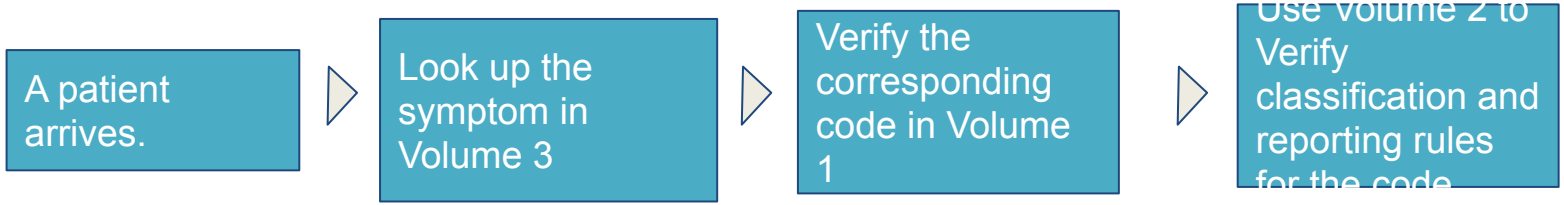
Inclusion terms

Within the three- and four-character rubrics[1], there are usually listed a number of other diagnostic terms. These are known as "inclusion terms" and are given, in addition to the title, as examples of the diagnostic statements to be classified to that rubric. They may refer to different conditions or be synonyms. They are not a subclassification of the rubric.

Inclusion terms are listed primarily as a guide to the content of the rubrics. Many of the items listed relate to important or common terms belonging to the rubric. Others are borderline conditions or sites listed to distinguish the

Instructions to encode
Morbidity: Hospital statistics
Mortality: Death statistics

Recommendation to use ICD Codes

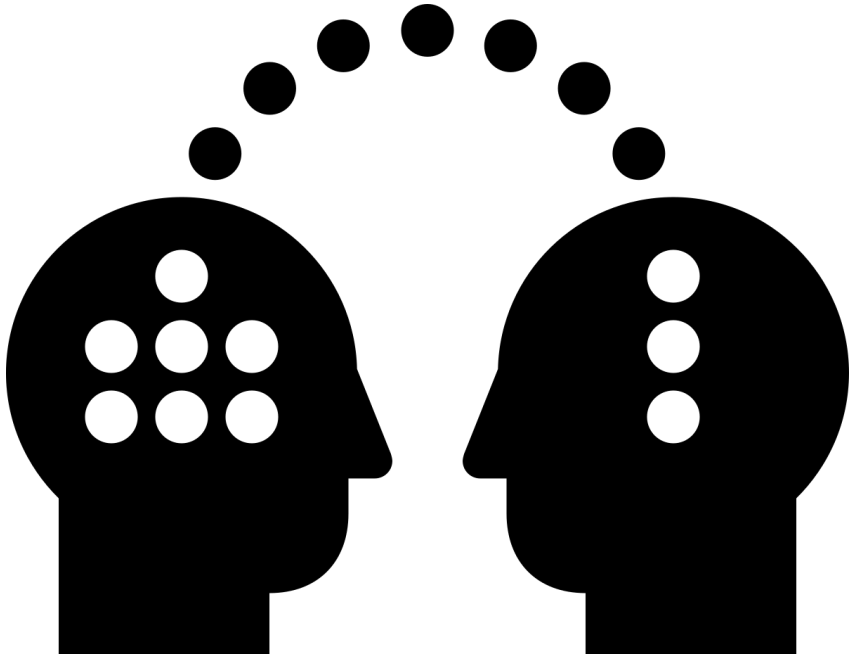


Tabular list uses British spelling
Alphabetical index uses American spelling to sort, with cross-references

Ontology

In philosophy: **Onto-: being -logia: study. Ontology** is the philosophical study of being.
In AI: Ontology is a formal specification of a shared conceptualization

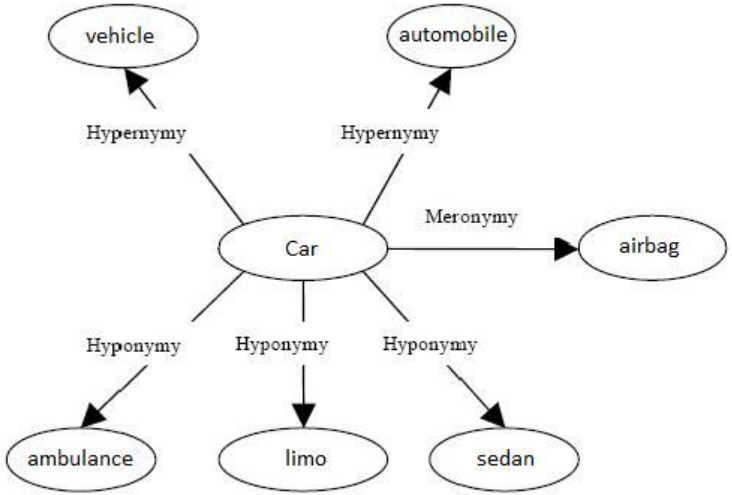
Properties of an ontology: Clarity, Coherence, Extensibility, Minimal encoding bias, Minimal ontological commitment



T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.
http://www.math.ubbcluj.ro/~didactica/pdfs/2013/didmath2013-06.pdf

Ontology fundamentals

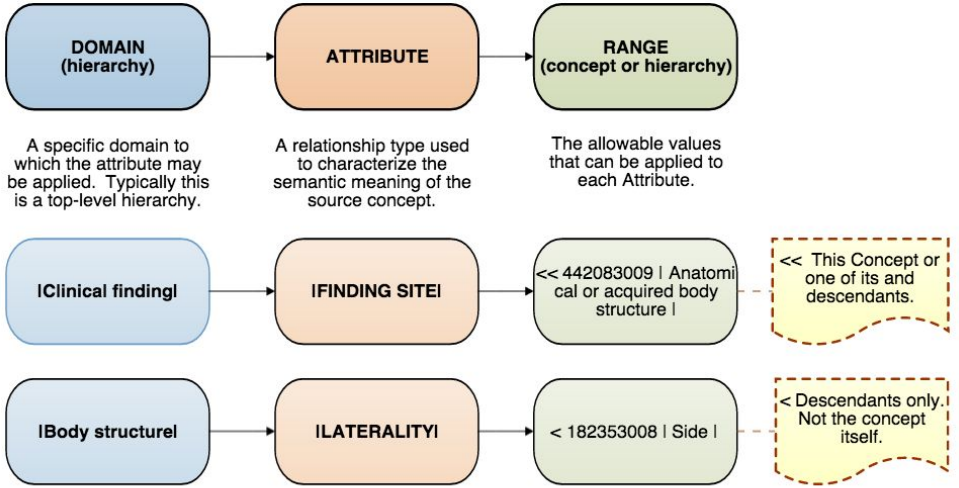
Concepts
Relationships (between concepts)



https://iaoa.org/isc2012/docs/Guarino2009_What_is_an_Ontology.pdf
George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

SNOMED

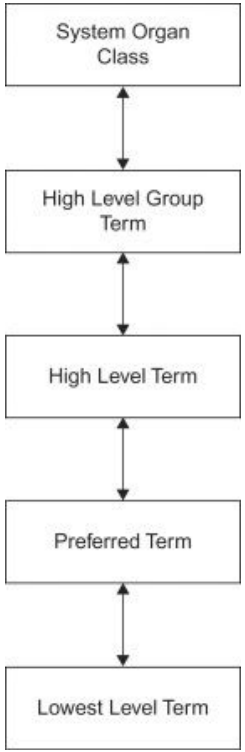
Systematized Nomenclature of Medicine -- Clinical Terms
Standardised, multi-lingual vocabulary of clinical terms



Spackman, Kent A., Keith E. Campbell, and Roger A. Côté. "SNOMED RT: a reference terminology for health care." In *Proceedings of the AMIA annual fall symposium*, p. 640. American Medical Informatics Association, 1997.Image Source: https://confluence.ihtsdotools.org/display/DOCSTART/6.+SNOMED+CT+Concept+Model

Medical Dictionary for Regulatory Activities (MedDRA)

Medical conditions and devices
Includes pharmaceuticals, vaccines and drug-device combination products.



Brown, E.G., Wood, L. and Wood, S., 1999. The medical dictionary for regulatory activities (MedDRA). *Drug safety*, *20*(2), pp.109-117.

BioCaster

Ontology that captures syndromic knowledge (Collier et al., 2007)
(a) concepts such as disease, symptom, virus, or syndrome, and
(b) relations such as has_symptom that relates a disease with a symptom, or causes that relates a disease with the virus that causes the disease.
The ontology is multilingual with support in 12 languages such as English, Japanese, French, Arabic and Thai.

N Collier, A Kawazoe, L Jin, M Shigematsu, D Dien, et al. 2007. The BioCaster Ontology: A multilingual ontology for infectious disease outbreak surveillance: Rationale, design and challenges. Journal of Language Resources and Evaluation (2007), 405–413.

Linkings between ontologies!

Many ontologies
Efforts to link them
Stearns et al (2001) describes an effort to combine two medical ontologies SNOMED RT and Clinical Terms; mostly manual by experts
Ghazvinian et al (2009) use the LOOM algorithm to automatic mappings for ontologies

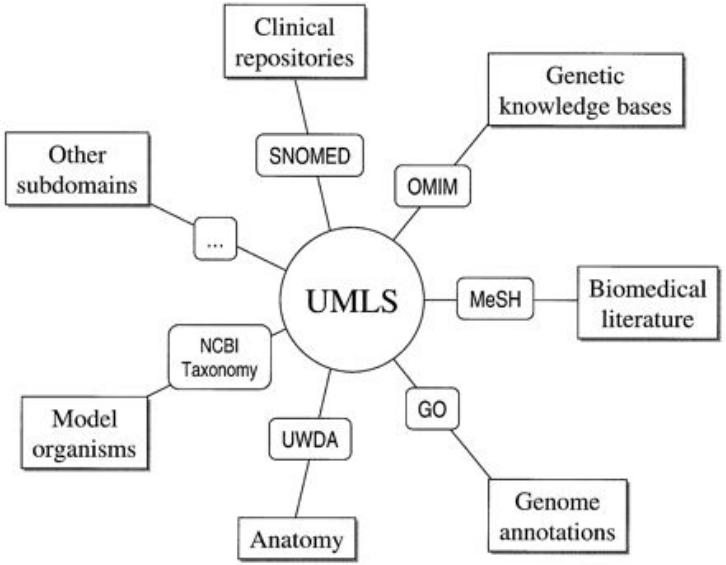
Stearns, Michael Q., Colin Price, Kent A. Spackman, and Amy Y. Wang. "SNOMED clinical terms: overview of the development process and project status." In *Proceedings of the AMIA Symposium*, p. 662. American Medical Informatics Association, 2001.
Ghazvinian, Amir, Natalya F. Noy, and Mark A. Musen. "Creating mappings for ontologies in biomedicine: simple methods work." In *AMIA Annual Symposium Proceedings*, vol. 2009, p. 198. American Medical Informatics Association, 2009.

Unified Medical Language System (UMLS)

McCray et al 1989
(https://www.nlm.nih.gov/research/umls/index.html)
15 languages
Metathesaurus: Terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT.

Semantic Network: Broad categories (semantic types) and their relationships (semantic relations).

SPECIALIST Lexicon and Lexical Tools: A large syntactic lexicon of biomedical and general English and tools for normalizing strings, generating lexical variants, and creating indexes.



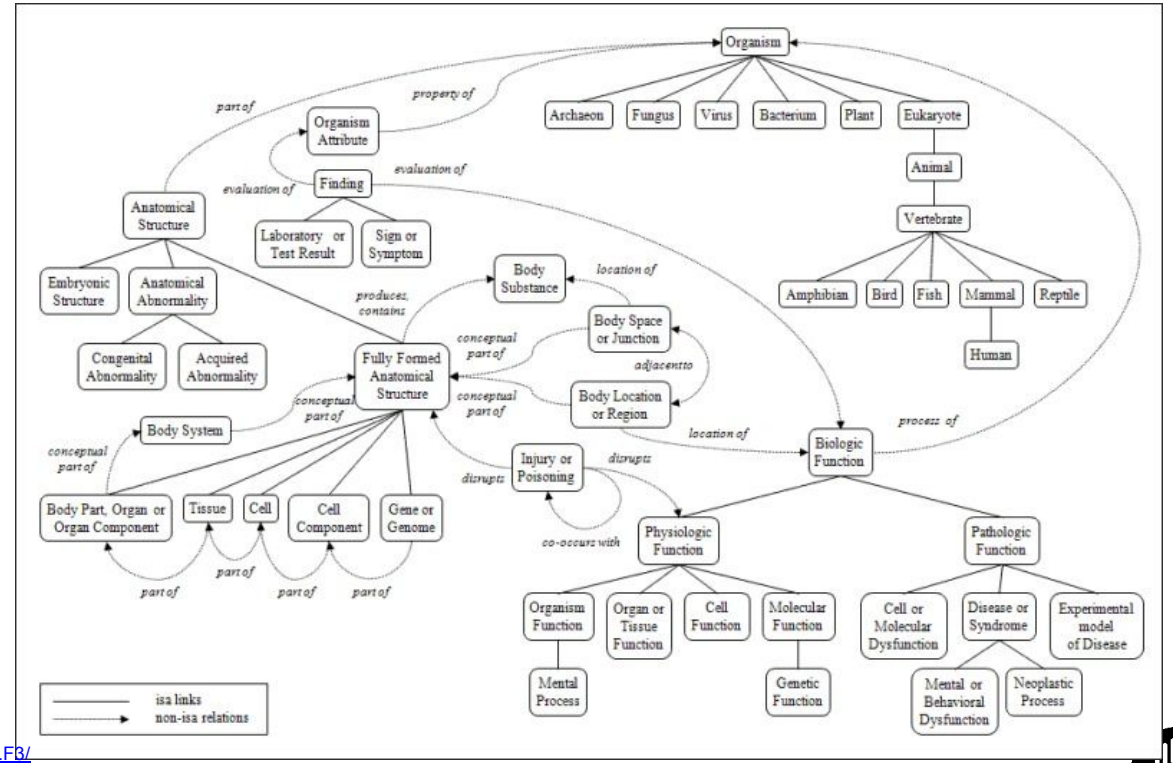
McCray, Alexa T. "The UMLS Semantic Network." In *Proceedings. Symposium on Computer Applications in Medical Care*, pp. 503-507. American Medical Informatics Association, 1989.

UMLS Meta-thesaurus

Biomedical concepts and relationships
Relationships of two types:

1. Associative (for example, symptoms to syndromes)
2. Hierarchical (for example, specialised illnesses)

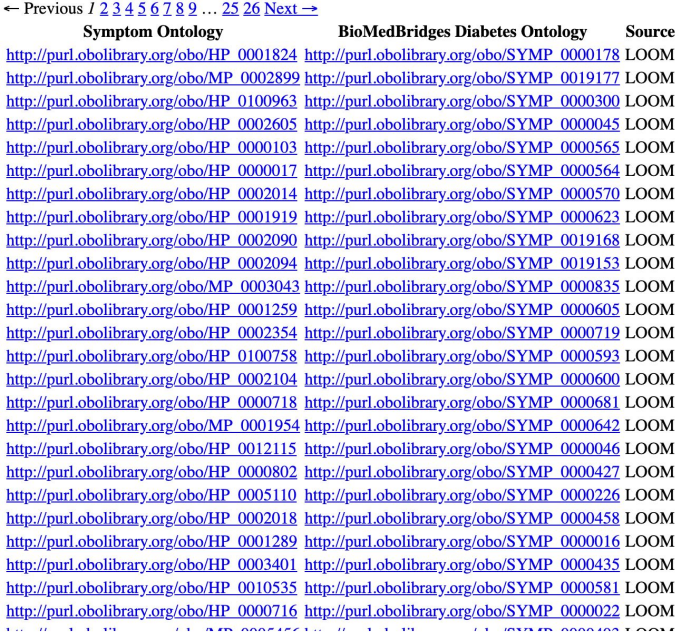
UMLS Semantic Network



https://www.ncbi.nlm.nih.gov/books/NBK9679/figure/ch05.F3/

NCBO BioPortal

Library of ontologies
Search
Browser-based
Mappings between many ontologies



← Previous *1* 2 3 4 5 6 7 8 9 … 25 26 Next →

Symptom Ontology	BioMedBridges Diabetes Ontology	Source
http://purl.obolibrary.org/obo/HP_0001824	http://purl.obolibrary.org/obo/SYMP_0000178	LOOM
http://purl.obolibrary.org/obo/MP_0002899	http://purl.obolibrary.org/obo/SYMP_0019177	LOOM
http://purl.obolibrary.org/obo/HP_0100963	http://purl.obolibrary.org/obo/SYMP_0000300	LOOM
http://purl.obolibrary.org/obo/HP_0002605	http://purl.obolibrary.org/obo/SYMP_0000045	LOOM
http://purl.obolibrary.org/obo/HP_0000103	http://purl.obolibrary.org/obo/SYMP_0000565	LOOM
http://purl.obolibrary.org/obo/HP_0000017	http://purl.obolibrary.org/obo/SYMP_0000564	LOOM
http://purl.obolibrary.org/obo/HP_0002014	http://purl.obolibrary.org/obo/SYMP_0000570	LOOM
http://purl.obolibrary.org/obo/HP_0001919	http://purl.obolibrary.org/obo/SYMP_0000623	LOOM
http://purl.obolibrary.org/obo/HP_0002090	http://purl.obolibrary.org/obo/SYMP_0019168	LOOM
http://purl.obolibrary.org/obo/HP_0002094	http://purl.obolibrary.org/obo/SYMP_0019153	LOOM
http://purl.obolibrary.org/obo/MP_0003043	http://purl.obolibrary.org/obo/SYMP_0000835	LOOM
http://purl.obolibrary.org/obo/HP_0001259	http://purl.obolibrary.org/obo/SYMP_0000605	LOOM
http://purl.obolibrary.org/obo/HP_0002354	http://purl.obolibrary.org/obo/SYMP_0000719	LOOM
http://purl.obolibrary.org/obo/HP_0100758	http://purl.obolibrary.org/obo/SYMP_0000593	LOOM
http://purl.obolibrary.org/obo/HP_0002104	http://purl.obolibrary.org/obo/SYMP_0000600	LOOM
http://purl.obolibrary.org/obo/HP_0000718	http://purl.obolibrary.org/obo/SYMP_0000681	LOOM
http://purl.obolibrary.org/obo/MP_0001954	http://purl.obolibrary.org/obo/SYMP_0000642	LOOM
http://purl.obolibrary.org/obo/HP_0012115	http://purl.obolibrary.org/obo/SYMP_0000046	LOOM
http://purl.obolibrary.org/obo/HP_0000802	http://purl.obolibrary.org/obo/SYMP_0000427	LOOM
http://purl.obolibrary.org/obo/HP_0005110	http://purl.obolibrary.org/obo/SYMP_0000226	LOOM
http://purl.obolibrary.org/obo/HP_0002018	http://purl.obolibrary.org/obo/SYMP_0000458	LOOM
http://purl.obolibrary.org/obo/HP_0001289	http://purl.obolibrary.org/obo/SYMP_0000016	LOOM
http://purl.obolibrary.org/obo/HP_0003401	http://purl.obolibrary.org/obo/SYMP_0000435	LOOM
http://purl.obolibrary.org/obo/HP_0010535	http://purl.obolibrary.org/obo/SYMP_0000581	LOOM
http://purl.obolibrary.org/obo/HP_0000716	http://purl.obolibrary.org/obo/SYMP_0000022	LOOM

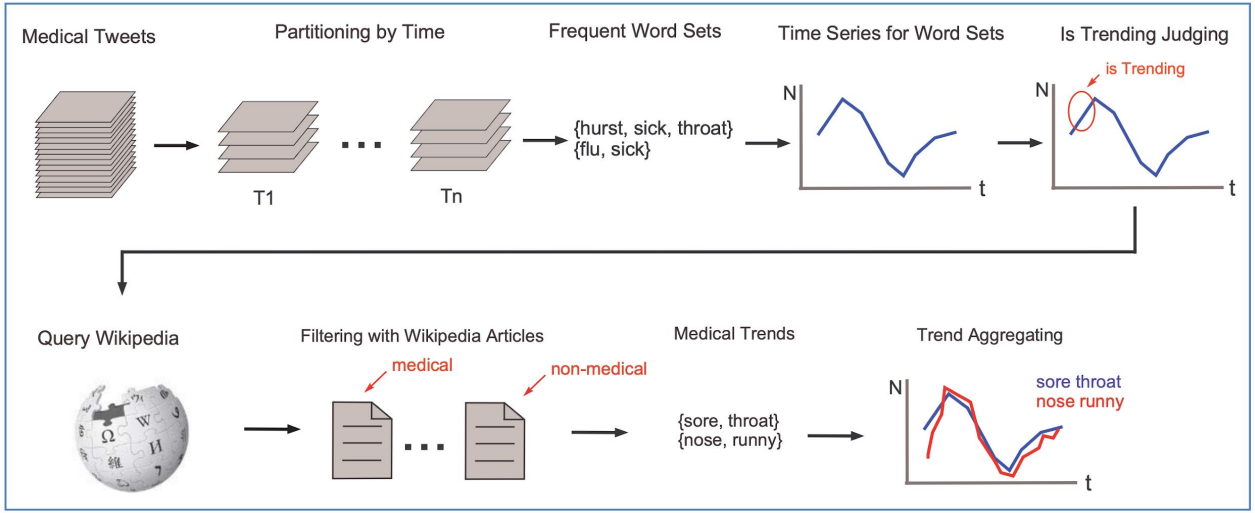
Using ICD codes

1) Automatic ICD code classification using text-based autopsy reports (Mujtaba et al, 2017)
2) Description associated with ICD codes can provide useful keywords to:
   a) To create datasets (Next slide)
   b) To assign labels to unsupervised clusters (Webb et al, 2018): LDA is run on a dataset of tweets; clusters are named based on the presence of top words (in a cluster) in the ICD code descriptions.

Mujtaba, Ghulam, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, and Mohammed Ali Al-Garadi. "Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection." PloS one 12, no. 2 (2017): e0170242.
Webb, Frank, Amir Karami, and Vanessa Kitzie. "Characterizing Diseases and disorders in Gay Users' tweets." *arXiv preprint arXiv:1803.09134* (2018).

ICD Codes for Social Media-based Monitoring

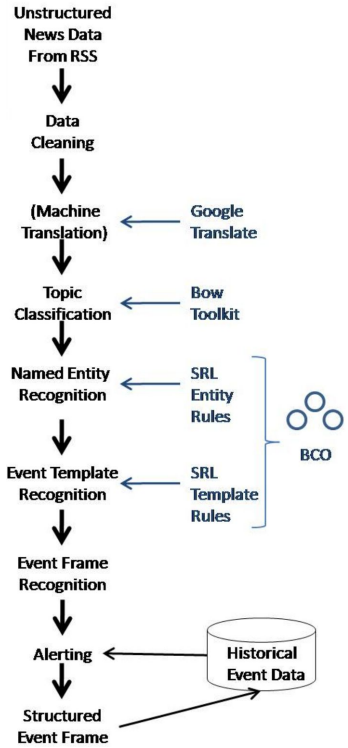
ICD Codes on the Wikipedia page are used to filter relevant keywords (Parker et al, 2013)



Parker, Jon, Yifang Wei, Andrew Yates, Ophir Frieder, and Nazli Goharian. "A framework for detecting public health trends with twitter." In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 556-563. ACM, 2013.

Using ontologies (1/2)

(Collier et al., 2010) An ontology-based approach to detect global health events
Monitors news feeds for multi-lingual health-related news
Based on the target keywords, relevant news articles are then translated to English
Topic classification to further filter news
Then, information extraction techniques, such as named entity recognition or semantic role labeling, to identify concepts/relationships in the BioCaster ontology



Nigel Collier, Reiko Matsuda Goodwin, John McCrae, Son Doan, Ai Kawazoe, Mike Conway, Asanee Kawtrakul, Koichi Takeuchi, and Dinh Dien. 2010. An ontology-driven system for detecting global health events. In Proceedings of the 23rd International Conference on Computational Linguistics (Uppsala, Sweden). Association for Computational Linguistics, 215–222.

Using ontologies (2/2)

(Huang et al, 2016)
A medical ontology provides features of a classifier for illness classification
Step 1: If a word in a tweet is predicted as an entity of interest, it is mapped to a concept present the ontology using similarity values.
Step 2: The concepts themselves become the features of a classifier that detects an illness.

Pin Huang, Andrew MacKinlay, and Antonio Jimeno Yepes. 2016. Syndromic Surveillance using Generic Medical Entities on Twitter. In Proceedings of the Australasian Language Technology Association Workshop 2016 (Melbourne, Australia). 35–44.

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module IV: Classification Problems

Outline

- Typical problem definitions
- Similarity-based approaches
- Topic Model-based approaches
- Pipeline-based approaches
- Statistical approaches
- Deep learning-based approaches

All images in this module are from wikimedia commons.

Typical problem definitions (1/2)

Health mention classification: Predict whether or not a given text is a health mention

"I have had cough all week" versus "Ice creams can cause cough"

Why? To detect health outbreaks; Individual health mentions are events

Challenges: Targets may be important; a celebrity reporting a health symptom may be widely re-tweeted or discussed but does not correspond to a separate event (Kanouchi et al, 2015).

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold?- identifying the subject of a symptom. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Beijing, China), Vol. 1. 1660–1670.

Typical problem definitions (2/2)

Adverse drug reaction (ADR) detection: ADR is an injury caused by a medication

Data Source	Type	Reliability	ADR Specific	ADR Monitoring
SRS	Structured	High	Yes	Passive
EHR	Structured	High	No	Active & passive
Administrative DB	Structured	High	No	Active
Medical literature	Unstructured	High	No	Active & passive
Medical forums	Unstructured	Low	No	Active
Social media	Unstructured	Low	No	Active
Search engine logs	Semi-structured	Low	No	Active

Sentiment detection (for example: Vaccine sentiment): Classify the sentiment expressed about vaccines in a given text

Karimi, Sarvnaz, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. "Text and data mining techniques in adverse drug reaction detection." *ACM Computing Surveys (CSUR)* 47, no. 4 (2015): 56.

Classification-based approaches

Similarity-based approaches
Topic Model-based approaches
Pipeline-based approaches
Statistical approaches
Deep learning-based approaches

Similarity-based Approaches

Notions of similarity to model classes of interest
Semantic distance between words in a text and medical words of interest

Health mention classification of news articles

(Freifeld et al, 2008): Match n-grams in a news article with known dictionary of terms based on semantic distances
Thus, classify each news article for two output labels: primary location and disease name
(Lejeune et al., 2010): Multilingual news articles. Use similarity to extract properties such as location, time and type of illness. As a result, classify a news article as relevant to health or not.

Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. 2008. HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inf. Assoc.* 15, 2 (2008), 150–157.
Gaël Lejeune, Antoine Doucet, Roman Yangarber, and Nadine Lucas. 2010. Filtering news for epidemic surveillance: Towards processing more languages with fewer resources. In Proceedings of the 4th International Workshop on Cross-lingual Information Access.

Topic model-based approaches

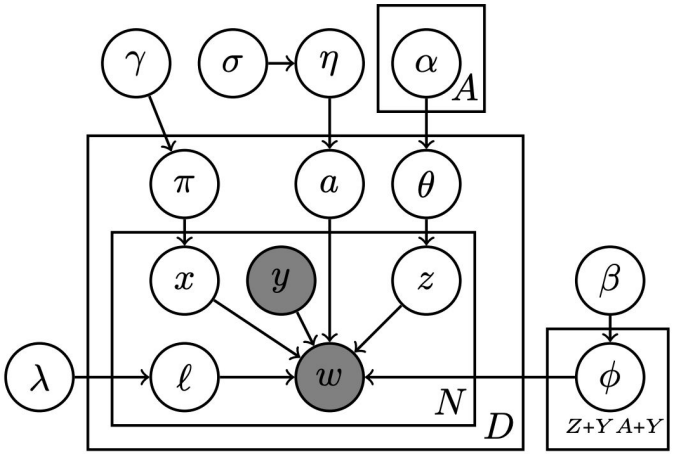
Topic models can be used to discover semantic concepts underlying large datasets, either labeled or unlabeled
Large unlabeled datasets can be used to extract structured topics of interest, and, as a result, classify tweets
Asymmetric priors are set on words to bootstrap the models

Aspect Topic Ailment Model (ATAM)

(Paul and Dredze., 2012) Three latent variables:

- Switching variable for general or health-related word (l)
- Identifying background words (x)
- An ailment (z)

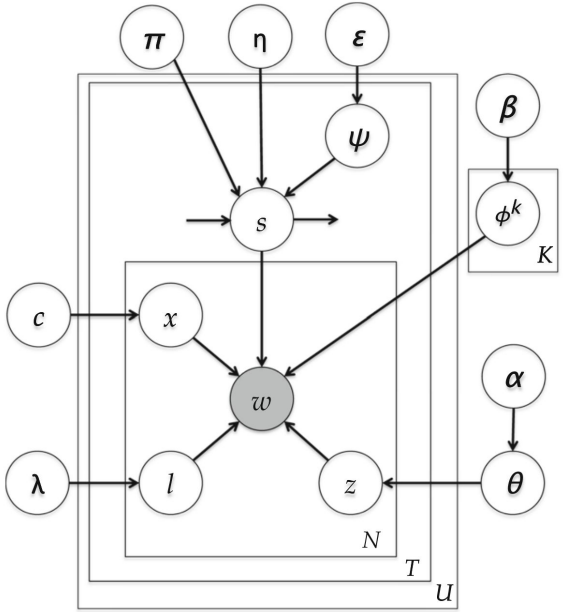
Applied to Chinese microblogs by Wang et al. (2014)



Michael J. Paul and Mark Dredze. 2012. A model for mining public health topics from Twitter. *Health* 11 (2012).
Shiliang Wang, Michael J. Paul, and Mark Dredze. 2014. Exploring health topics in Chinese social media: An analysis of Sina Weibo. In Proceedings of the AAAI Workshop on the World Wide Web and Public Health Intelligence, Vol. 31. 59.

Hidden Flu-State Tweet Model

Temporal model (Chen et al., 2016)
Word-level variable (w) each to indicate background words and general general words
Switch (x) between symptom (l) and general (z) words
Tweet-level symptom variable (s)
The symptom variable of a tweet depends on the value of the previous tweet by a user



Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Aditya Prakash. 2016. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Min. Knowl. Discov.* 30, 3 (2016), 681–710.

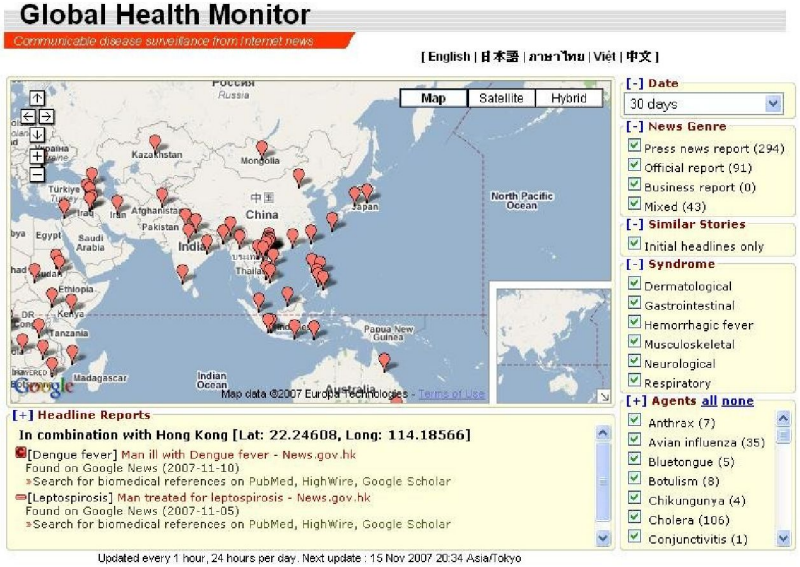
Pipeline-based Approaches

A pipeline of NLP components; Classification in addition to other information
Examples:
Global Health Monitor, MedISys

Global Health Monitor

(Doan et al., 2008)
Periodically scans news feeds

A) Topic classification using Naive Bayes
B) Named entity recognition and disease location detection to extract terms
C) Visualisation to present the news on a map



Son Doan, Ai Kawazoe, Nigel Collier, et al. 2008. Global health monitor-a web-based system for detecting and mapping infectious diseases. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*.

MedISys

(Yangarber et al., 2008)
News articles are searched from feeds
Then categorised as health-related or not
Then, information extraction system based on patterns extracts incidents
Location, Disease Name, Date/Period, Victim Information

Roman Yangarber, Peter Von Etter, and Ralf Steinberger. 2008. Content collection and analysis in the domain of epidemiology. In *Proceedings of the International Workshop on Describing Medical Web Resources (DrMED'08)*.

Statistical Approaches

	Classifier	Features
Olszewski (2003)	Naive Bayes	Unigrams and bigrams
Aramaki et al (2011)	Naive Bayes, AdaBoost, SVM	Bag of words with feature windows of multiple sizes
Névéol et al (2009)	Priority model: Probabilistic model	N-grams
Kanouchi et al (2015)		Unigrams, weblinks, word classes, length, n-grams, etc.
Jiang et al. (2016)		Emotion words, emotion scores, user mentions,number, pronominal mentions

Robert T. Olszewski. 2003. Bayesian classification of triage diagnoses for the early detection of epidemics. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*. 412–416.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1576.

Aurélie Névéol, Won Kim, W. John Wilbur, and Zhiyong Lu. 2009. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, 144–152.

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold?-identifying the subject of a symptom. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 1660–1670.

Keyuan Jiang, Ricardo Calix, and Matrika Gupta. 2016. Construction of a personal experience tweet corpus for health surveillance. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 128–135.

Twitter-based epidemic detection (1/2)

(Aramaki et al., 2011): Health mention classification for tweets
“My flu is worse than it was yesterday” : Positive
“A bad influenza is going around in our lab.”: Positive
“In the normal flu season, 80-percent-of-deaths occur in people over 65”: Negative
Annotate a dataset of 5000 tweets

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1576.

Twitter-based epidemic detection (2/2)

Context window-based features
Support vector machine with polynomial kernel
Also experiment with other classifiers

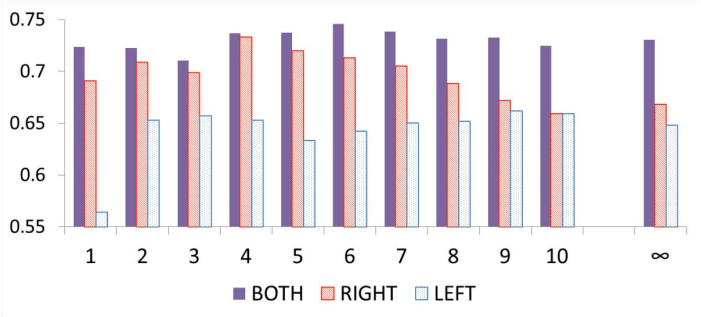


Figure 3: Window size and Accuracy (F-measure). RIGHT shows a method used only the right context. LEFT shows a method used only the left context. BOTH represents a method using both the right and left context. The number shows the window size. ∞ uses all words in each context direction.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1576.

Approaches using DL-based embeddings (1/2)

Lampos et al. (2017): Embeddings to select unigram features for flu detection: (a) Wikipedia-based embeddings, (b) Tweet-based embeddings
Dai et al. (2017): Word embeddings are used to create clusters of concepts. The clusters are used as features.
Joshi et al (2019): GloVe and Word2Vec et al. with ELMo and USE et al.

Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 695–704.
Xiangfeng Dai, Marwan Bikdash, and Bradley Meyer. 2017. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *Proceedings of the IEEE Region 3 Technical, Professional, and Student Conference (SoutheastCon'17)*. IEEE, 1–7.
Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C. Raina MacIntyre. 2019. A comparison of word-based and context-based representations for classification problems in health informatics. In *Proceedings of Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

Approaches using DL-based embeddings (2/2)

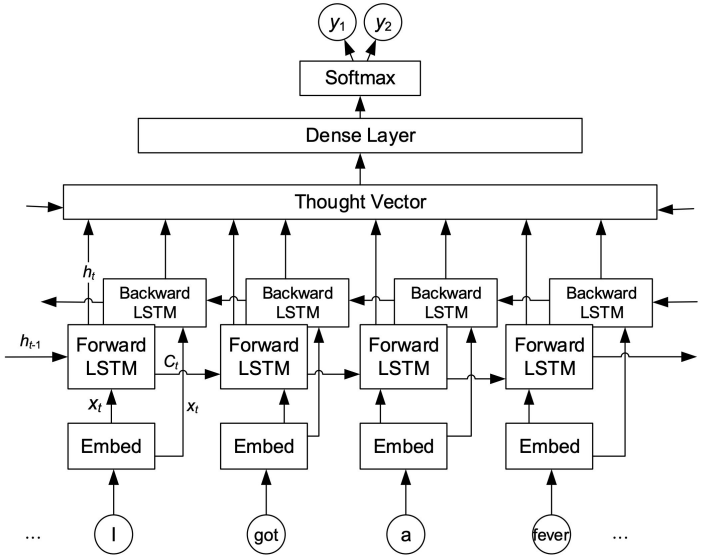
Karisani et al. (2018) :
Word Embedding Space Partitioning And Distortion:

- Word embedding space is partitioned on the basis of a labeled dataset
- A sentence vector is information gain-based weighted version
- Sentence vector is the feature of a statistical classifier

Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack?: Toward robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 137–146.

Deep learning architectures (1/2)

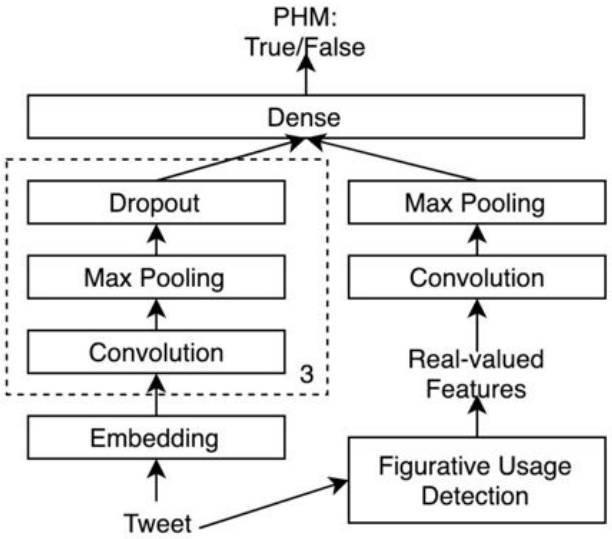
Wang et al (2017) : RNNs



Chen-Kai Wang, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. 2017. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In *Proceedings of the International Workshop on Digital Disease Detection Using Social Media 2017 (DDDSM'17)*. 33–38.

Deep learning architectures (2/2)

(Iyer et al., 2019)
Features from a figurative usage detection module appended to semantic representation
Observe an improvement of about 2.21% on the dataset by Karisani et al (2018)



Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris (2019). `Figurative Usage Detection of Symptom Words to Improve Personal Health Mention Detection', Proceedings of ACL 2019.

Summary

Approach	General Idea	Motivation	Challenges
Ontology enhanced [18, 21, 22, 34, 48]	Given a text, map the terms in the text to appropriate concepts in the ontology to determine if a syndrome can be detected.	Medical ontologies capture useful information in a structured form.	(A) Ontologies may not be complete, (B) ontologies may contain medical terms while the text may contain colloquial terms.
Similarity based [1, 26, 46, 51]	Similarity between distributions and similarity between concepts are used as indicators of an illness.	A text that is similar to illness concepts/text is likely to be about the illness.	The choice of similarity metric determines the benefit.
Topic Model based [17, 54, 55, 66]	With the help of datasets from social media topic models that are extensions of Latent Dirichlet Allocation (LDA) model have been proposed. With the use of additional latent variables, these models provide structured information about illnesses.	Topic models can process unlabeled/partially labeled data and provide valuable information.	Interpretation of generated topics and their application to Health mention classification may not be straightforward.
Pipeline based [25, 72–74]	These approaches combine existing NLP components to build effective deployments. Typical components include named entity extraction and text classification.	Health mention classification can be broken down into a sequence of NLP components fitting into one another.	NLP components may be trained on documents in domains unrelated to health-care. In such cases, their efficacy for health mention classification needs validation.
Statistical [5, 15, 37, 40, 43, 53]	Features based on words, emotion scores, medical concepts and POS tags, along with typical classifier learning algorithms have been reported.	Supervised classifiers trained on labeled datasets have been found to be useful in many applications of NLP	Selecting appropriate features and ensuring they generalise may be challenging.
Deep Learning based [23, 39, 42, 44, 65]	Features based on word embeddings and modification of general-purpose word embeddings to the specific domain space have been reported, along with typical neural network models.	Deep learning approaches have proven to be useful, since they do not rely on human-engineered features.	Lack of availability of large labeled datasets may be an impediment.

Joshi, Aditya, Sarvnaz Karimi, Ross Sparks, Cecile Paris and C Raina MacIntyre. "Text and data mining techniques in adverse drug reaction detection." *ACM Computing Surveys (CSUR)* 52, no. 6 (2019): 119.

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Biomedical/Health search, Normalisation

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Biomedical/Health search, Normalisation

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module V: Information Extraction

Outline:

- What is IE?
- Named Entity Recognition, Normalisation and Relation Extraction
- NER on User-generated Text
  - Adverse Drug Event Extraction from Patient Reports
  - Pipeline-based approaches for Epidemic intelligence

Information Extraction (IE)

- IE is the process of automatically extracting structured information from free-text.

- Applications:
  - Template-filling
  - Knowledge base population

- Techniques:
  - Named Entity Recognition (NER)
  - Relation Extraction
  - Event Extraction

Speech and Language Processing.(Chapter 18) Daniel Jurafsky & James H. Martin.2019

Biomedical NER

- Entity types of interest depends on application. In biomedical domain, some of these types are:

  Chemicals, drugs, adverse events, dosage, cell lines, diseases, genes, proteins

> "I have been using pro banthine for over 5 years now, and I really cant imagine my life without it."
>
> - review of Pro-BANTHINE from AskaPatient forum

Speech and Language Processing.(Chapter 18) Daniel Jurafsky & James H. Martin.2019

Complex Entities

- Some entities are discontinuous spans of text. Some can overlap or even nested.

> “Got severe lower back and leg pain. Could not walk.”
>
> - review of Lipitor from AskaPatient forum

Adverse events: severe lower back pain, severe leg pain

Dai, X. Recognizing Complex Entity Mentions: A Review and Future Directions, ACL SRW 2018

NER Methods

- Rule-based
- Dictionary-based

- Classification-based
- Sequence-tagging-based
  - Conditional Random Field (CRF) was SOTA for a long time

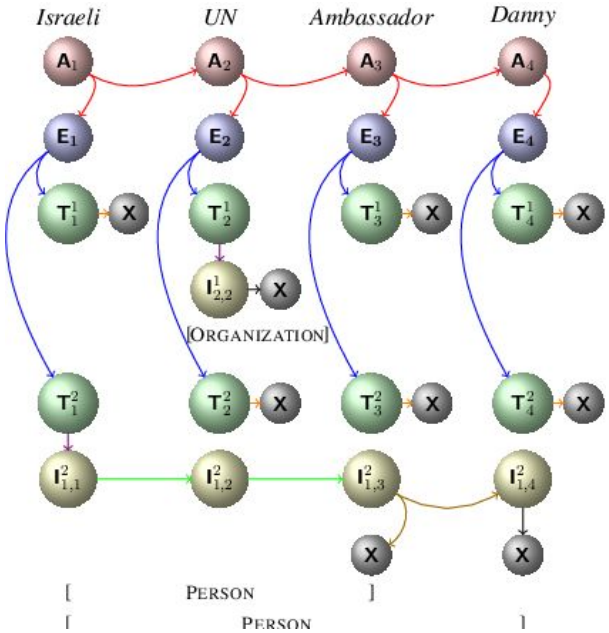
- Current SOTA for different types of text is a variation of BiLSTM-CRF
- Some add extra lexical features to BiLSTM-CRFs

Leser, U, Hakenberg, J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform. 2005 Dec;6(4):357-69
Strubell, E, P. Verga, D. Belanger, A. McCallum. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. EMNLP 2017
Ghaddar, A, P. Langlais. Robust Lexical Features for Improved Neural Network Named-Entity Recognition. COLING 2018

Complex NER Methods: Nested Entities

- Wang and Lu (2018): A neural segmental hypergraph model that captures overlapping mentions
- In mathematics, a hypergraph is a generalization of a graph in which an edge can join any number of vertices.

- Datasets: GENIA, ACE



Dai, X. Recognizing Complex Entity Mentions: A Review and Future Directions, ACL SRW 2018
Wang, B. Lu, W. Neural Segmental Hypergraphs for Overlapping Mention Recognition, EMNLP 2018
Fisher, J., Vlachos, A. Merge and label a novel neural network architecture for nested NER, ACL 2019

Complex NER Methods: Discontinuous Entities

- Bang, Lu (2019): Expand on their previous hypergraph model to identify mentions which are discontinuous.
- Justification is that in the medical area such entities can be frequent
- For the discontinuous mentions, they add an extra classifier to decide whether or not two identified text segments belong to one mention

Bang, L., Lu, W. Combining Spans into Entities: A Neural Two-Stage Approach for Recognizing Discontiguous Entities, EMNLP 2019

NER for Pharmacovigilance

- Source:
  - Formal text (literature)
  - Informal text (forums such as AskaPatient, PatientsLikeMe or DailyStrength, and twitter)

- Entity types:
  - Drug name
  - Disease
  - Symptoms
  - Adverse events - Most difficult

- Causality: Identifying potential adverse events does not imply causality

Signal detection from Social Media

- Van Stekelenborg et al (2019): “Many patients and clinicians have taken to social media to discuss their positive and negative experiences of medications, creating a source of publicly available information that has the potential to provide insights into medicinal product safety concerns.”

- They recommend using social media for pharmacovigilance areas such as exposure during pregnancy and abuse/misuse of medicines.

J van Stekelenborg. Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR. Drug Saf. 2019 Dec;42(12):1393-1407.

NER Resources: Annotated Corpus

- Some of the existing NER corpora:
  - GENIA (biomedical, MEDLINE abstracts)
  - CRAFT (biomedical, full-text biomedical journal articles)
  - CADEC (biomedical, medication forum)
  - SMM4H (biomedical, tweets)
  - NNE (nested entities, news)
  - Some of BioCreative corpora, such as CHEMNER track data

NER Resources: Tools

- Some of the freely available ones:
  - HUNER (LSTM-CRF based method)
  - GNormPlus
  - tmChem
  - BANNER
  - ABNER
  - AbGene
  - GAPScore
  - LingPipe

Leon Weber, Jannes Münchmeyer, Tim Rocktäschel, Maryam Habibi, Ulf Leser, HUNER: improving biomedical NER with pretraining, *Bioinformatics, 2019*

Named Entity Normalisation

- Once a named entity is identified, it can also be normalised (assigned) to a concept in ontologies. This task is also known as concept normalisation.
- For user-generated text, in particular, it could be necessary as a method to translate informal terms to formal:

charley horse ⟶ muscle cramp

- Depending on the application different ontologies are targeted. For example, adverse drug events may be mapped to MedDRA.

Concept Normalisation from Drug Reviews

- One popular tool is MetaMap: It works on formal text, is ngram-matching method and slow.

- Limsopatham and Collier (2016): A CNN based method with embeddings from Google News outperformed other methods, such as RNNs and CNN with embeddings trained on BioMed Central and retrieval-based methods.

- Tutubalina et al. (2018): RNNs with attention and similarity features from UMLS.

N. Limsopatham, N. Collier, Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation, in: ACL, 2016.
E. Tutubalina, Z. Miftahutdinova, S. Nikolenko, V. Malykh. Medical concept normalization in social media posts with recurrent neural networks, JBI 2018

Relation Extraction

- Biomedical literature contains knowledge that may not be found in knowledge-graphs. Such as protein-protein or drug-drug interactions.
- NLM statistics in June 2019 states that there is over 29 million MEDLINE abstracts published. It is impossible to read them all and find all those relational information manually!
- Biomedical relation extraction mines those hidden relations using NLP techniques (sometimes with added data mining methods)

Zhang, Y. et al. A hybrid model based on neural networks for biomedical relation extraction, JBI 2018

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

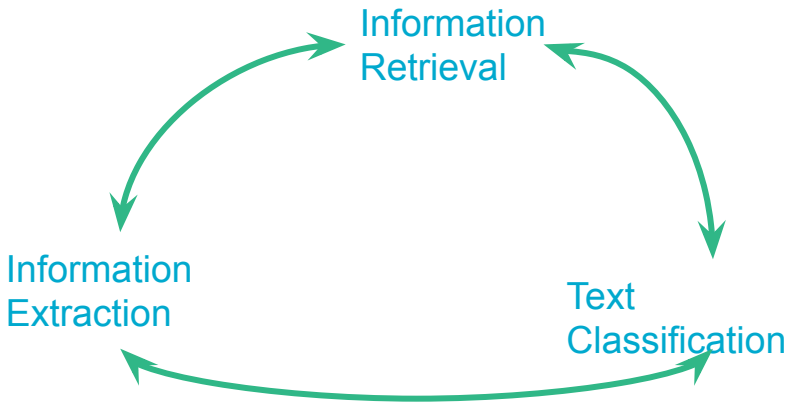
Module VI: Information Retrieval

Outline

- What is IR?
- Biomedical literature search
- Clinical trial search
- Benchmarking platform

Search and Information Retrieval

- Search is often an initial step of biomedical text mining applications to retrieve an initial set of documents to process.



Information Retrieval 101

- Indexing a document collection (e.g., MEDLINE abstracts)
- Querying the index

```
treatments for lung cancer
```

- Ranking function decides in what order documents are shown to the users (e.g., Okapi BM25)
- Search result representation
  - Snippet
  - Highlight key-terms

Biomedical Literature Search

- PubMED is the major search engine on biomedical literature, but it is largely Boolean search
- TREC (Text Retrieval Evaluation Conference) ran Clinical Decision Support Track 2014-2016
- Goal: Retrieving biomedical articles relevant for answering clinical questions about medical records:
  - What test
  - What diagnosis
  - What treatment

* ADCS'19 has a keynote by Guido Zuccon on Search Engines, their Evaluation and the Impact on Health Decisions

Clinical Decision Support Track: Example Topic

```
<topic number="1" type="diagnosis">
<note>
78 M w/ pmh of CABG in early [**Month (only) 3**] at [**Hospital6 4406**]
(transferred to nursing home for rehab on [**12-8**] after several falls out of bed.)
He was then readmitted to [**Hospital6 1749**] on [**3120-12-11**] after developing
acute pulmonary edema/CHF/unresponsiveness?. There was a question whether he had a
small MI; he reportedly had a small NQWMI. He improved with diuresis and was not
intubated. . Yesterday, he was noted to have a melanotic stool earlier this evening
and then approximately 9 loose BM w/ some melena and some frank blood just prior to
transfer, unclear quantity.
</note>
<description>
78 M transferred to nursing home for rehab after CABG. Reportedly readmitted with a
small NQWMI. Yesterday, he was noted to have a melanotic stool and then today he had
approximately 9 loose BM w/ some melena and some frank blood just prior to transfer,
unclear quantity.
</description>
<summary>
A 78 year old male presents with frequent stools and melena.
</summary>
</topic>
```

Clinical text w/abbrv. & jargon

Text summarisation?

Search over Clinical Trials

- Finding the right trial for a patient
- Clinical trials can be accessed through websites such as clinicaltrials.gov

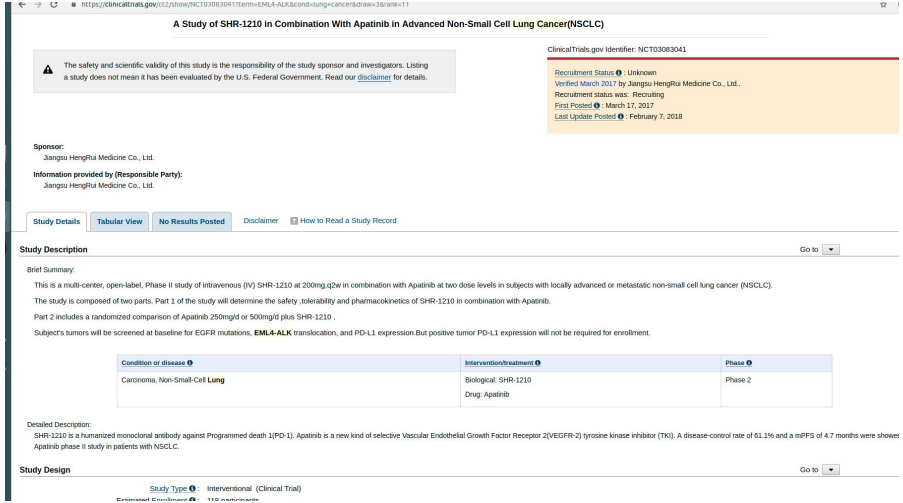
- Search over clinical trials in TREC precision medicine track (2017 to present)
  - Oncologists created synthetic cases (120 cases so far)
  - Each patient is represents by: Disease, Variant, Demographic
  - 2017 topics included other medical conditions as well

Precision Medicine Example Topic

```
<topic number="8">
<disease>Lung cancer</disease>
<gene>EML4-ALK Fusion transcript</gene>
<demographic>52-year-old male</demographic>
<other>Hypertension, Osteoarthritis</other>
</topic>
```

Inclusion and exclusion criteria of a matching clinical trial should match to all these.

```
<topic number="8">
<disease>Lung cancer</disease>
<gene>EML4-ALK Fusion transcript</gene>
<demographic>52-year-old male</demographic>
<other>Hypertension, Osteoarthritis</other>
</topic>
```



https://clinicaltrials.gov/ct2/show/NCT03083041?term=EML4-ALK&cond=lung+cancer&draw=3&rank=11

A Study of SHR-1210 in Combination With Apatinib in Advanced Non-Small Cell Lung Cancer(NSCLC)

The safety and scientific validity of this study is the responsibility of the study sponsor and investigators. Listing a study does not mean it has been evaluated by the U.S. Federal Government. Read our disclaimer for details.

ClinicalTrials.gov Identifier: NCT03083041

Recruitment Status : Unknown
Verified March 2017 by Jiangsu HengRui Medicine Co., Ltd..
Recruitment status was: Recruiting
First Posted : March 17, 2017
Last Update Posted : February 7, 2018

Sponsor:
Jiangsu HengRui Medicine Co., Ltd.

Information provided by (Responsible Party):
Jiangsu HengRui Medicine Co., Ltd.

Study Details | Tabular View | No Results Posted | Disclaimer | How to Read a Study Record

Study Description Go to

Brief Summary:

This is a multi-center, open-label, Phase II study of intravenous (IV) SHR-1210 at 200mg,q2w in combination with Apatinib at two dose levels in subjects with locally advanced or metastatic non-small cell lung cancer (NSCLC).

The study is composed of two parts. Part 1 of the study will determine the safety ,tolerability and pharmacokinetics of SHR-1210 in combination with Apatinib.

Part 2 includes a randomized comparison of Apatinib 250mg/d or 500mg/d plus SHR-1210 .

Subject's tumors will be screened at baseline for EGFR mutations, **EML4-ALK** translocation, and PD-L1 expression.But positive tumor PD-L1 expression will not be required for enrollment.

Condition or disease	Intervention/treatment	Phase
Carcinoma, Non-Small-Cell **Lung**	Biological: SHR-1210 Drug: Apatinib	Phase 2

Detailed Description:

SHR-1210 is a humanized monoclonal antibody against Programmed death 1(PD-1). Apatinib is a new kind of selective Vascular Endothelial Growth Factor Receptor 2(VEGFR-2) tyrosine kinase inhibitor (TKI). A disease-control rate of 61.1% and a mPFS of 4.7 months were shown Apatinib phase II study in patients with NSCLC.

Study Design Go to

Study Type : Interventional (Clinical Trial)

Liang *et al. BMC Cancer* (2016) 16:62
DOI 10.1186/s12885-016-2088-5

BMC Cancer

CASE REPORT Open Access

Metastatic EML4-ALK fusion detected by circulating DNA genotyping in an EGFR-mutated NSCLC patient and successful management by adding ALK inhibitors: a case report

Wenhua Liang[1,2†], Qihua He[1,2†], Ying Chen[1,2†], Shaokun Chuai[3], Weiqiang Yin[1,2], Wei Wang[1,2], Guilin Peng[1,2], Caicun Zhou[1,2,4*] and Jianxing He[1,2*]

Abstract

Background: Rebiopsy is highly recommended to identify the mechanism of acquired resistance to EGFR-TKIs in advanced lung cancer. Recent advances in multiplex genotyping based on circulating tumor DNA (ctDNA) provide a strong and non-invasive alternative for detection of the resistance mechanism.

Case presentation: Here we report a multiple metastatic NSCLC patient who was detected to have pure EGFR 19 exon deletion (negative for EML4-ALK and ROS1 in both IHC-based and sequencing assay) in the primary lesion and responded to first-line and second-line EGFR-TKI treatments (erlotinib then HY-15772). At 8 months, most lesions remained well controlled except for the liver metastases which presented dramatic progression. Considering the high risk of bleeding in rebiopsy of hepatic lesions, we conducted a multiplex genomic profiling with ctDNA. Results reported coexistence of EGFR mutation and EML4-ALK gene translocation in plasma which heavily indicated that ALK was the primary reason for progression of the liver lesions. This deduction was supported by the repeated response to ALK inhibitors (crizotinib then AP26113) of the hepatic metastases.

Conclusions: This is the first report of the existence of ALK rearrangement in metastatic lesions in an EGFR mutated patient. It highlighted the feasibility and advantages of using ctDNA multiplex genotyping in identifying the heterogeneity across lesions and the resistance mechanism of targeted treatments.

Keywords: NSCLC, EGFR mutation, EML4-ALK rearrangement, Co-existence

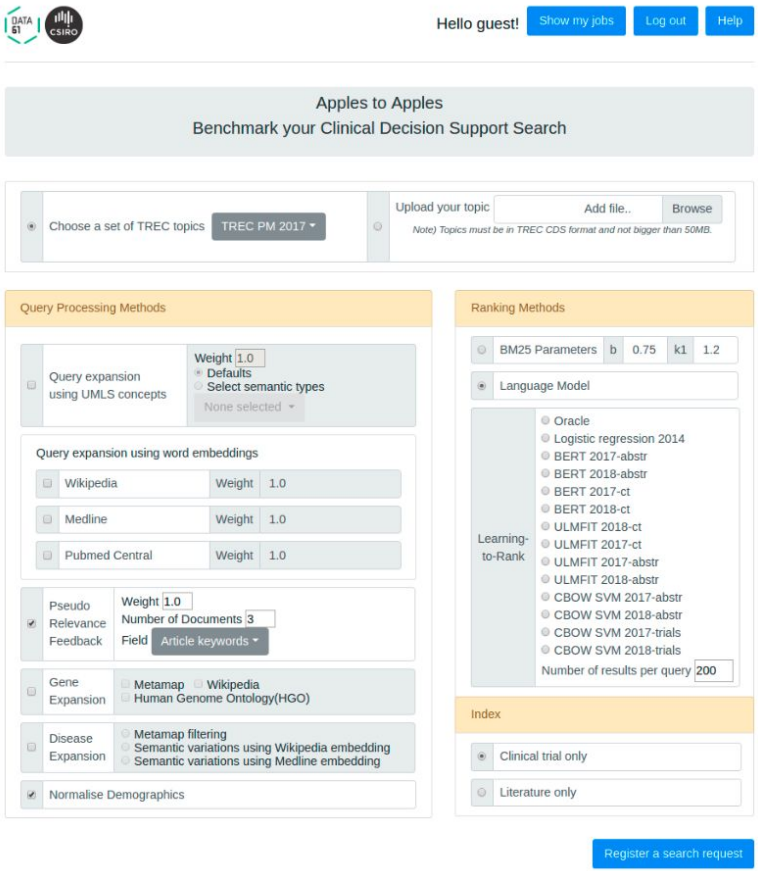
Background

Advances in geno-typing have changed the clinical practice of treatment of non-small cell lung cancer (NSCLC), especially non-squamous types where driver mutations, e.g. epidermal growth factor receptor (EGFR) mutations and echinoderm microtubule-associated protein-like 4-anaplastic lymphoma kinase (EML4-ALK) translocation are commonly present. Agents that target EGFR activating mutations (gefitinib, erlotinib, and afatinib, etc.) or ALK rearrangement (crizotinib, etc.) derive significantly greater benefits than cytotoxic chemotherapy in patients who harbor these gene alterations, which is consistently proved by extensive large-scale randomized controlled trials [1, 2]. In order to deliver an appropriate first-line treatment regimen, detection of EGFR mutation and ALK rearrangement are recommended as routine genetic profiling for non-squamous NSCLC or non-smoking populations [3]. In recent years, some selective inhibitors

* Correspondence: caicunzhoudr@163.com; dirjianxing.he@gmail.com
Co-first authors are Wenhua Liang, Qihua He and Ying Chen
†Equal contributors
[1]Department of Thoracic Surgery and Oncology, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China
[2]Guangzhou Institute of Respiratory Disease & China State Key Laboratory of Respiratory Disease & National Clinical Research Center for Respiratory Disease, No. 151, Yanjiang Rd, Guangzhou 510120, Guangdong Province, PR China
Full list of author information is available at the end of the article

A2A: Benchmarking Platform

- Allow a comparison of standard methods on the same index
- Can single-out different methods, e.g., negation detection



a2a.csiro.au

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Retrieval Problems

Biomedical/Health search, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Information Extraction Problems

NER from user-generated content

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module VII: Question Answering

Outline

- Patient Question Answering
- Shared tasks (BioASQ, MediQA)

Image source: https://images.app.goo.gl/n2tsgLDYFVbswz7P9

Patient Question Answering

- Question answering is more specific than IR: Users look for specific answers to their questions
- When it comes to medical area it becomes crucial to have higher confidence in answers or not to provide any
  - Pew (2013) survey: 35% of the US Adults search for their systems and ⅓ of these self-diagnosers do not seek professional help
- Patient often do not know the correct terminology to search over credible sources such as medical journals
- Even if presented relevant resources, may not be able to interpret them correctly

BioASQ: A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering

- QA challenges have been running for 7 years (2013 to date)
  - Pre-2019: 2747 biomedical questions and their gold standard answers (relevant concepts, articles, snippets, exact answers, summaries)
  - 2019 shared task was 5 batches of 100 biomedical questions each

- QA task uses benchmark datasets containing English training and test biomedical questions along with gold standard (reference) answers.

BioASQ: http://www.bioasq.org/participate/challenges

MediQA: Shared Task on Textual Inference, Question Entailment and Question Answering

- MedQuAD: Medical Question Answering Dataset of 47,457 QA pairs created from 12 NIH websites
- The collection covers 37 question types (e.g. Treatment, Diagnosis, Side Effects) associated with diseases, drugs and other medical entities such as tests.
- Two sets of medical questions and the associated lists of answers retrieved by the medical QA system CHiQA and reranked manually

MediQA: https://sites.google.com/view/mediqa2019

Abacha, Asma Ben, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the MediQA 2019 shared task on textual inference, question entailment and question answering. *BioNLP Workshop and Shared Task*, pp. 370-379. 2019.

MediQA: Example

<QuestionText>
abetalipoproteimemia hi, I would like to know if there is any support for those suffering with abetalipoproteinemia? I am not diagnosed but have had many test that indicate I am suffering with this, keen to learn how to get it diagnosed and how to manage, many thanks
</QuestionText>

MediQA: https://sites.google.com/view/mediqa2019

Abacha, Asma Ben, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the MediQA 2019 shared task on textual inference, question entailment and question answering. *BioNLP Workshop and Shared Task*, pp. 370-379. 2019.

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Biomedical/Health search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Tutorial Outline

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Biomedical/Health search

Conclusion

Summary, Future Directions

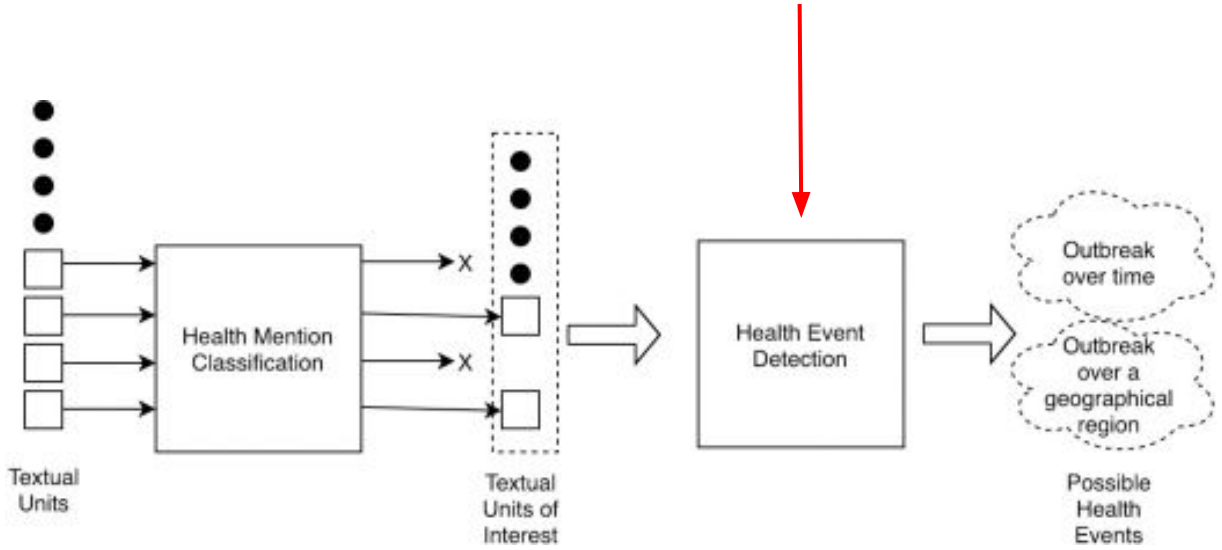
Image of coffee from wikimedia commons.

Module VIII: Time Series Monitoring Problems

Outline

- Typical problem definitions
- Predicting counts
- Predicting events

Why time series?



Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, C Raina MacIntyre, 'Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective', ACM Computing Surveys (CSUR) Surveys, Volume 52 Issue 6, Article No. 119, November 2019.

Event detection

Monitoring time series data to detect events of interest

Example: A disease outbreak, a surge in a certain sentiment about vaccines

Requirement: Textual units must be ordered (by timestamps, for example)

Typical time series monitoring problems

Two broad areas of related work:

- Predicting counts of infections
- Detecting outbreaks

Prediction of counts (1/3)

Ginsberg et al (2009): Google Flu Trends
Use search counts of the 45 million top queries across a subset of states from the US
Weekly counts of top queries are normalised
A model trained to predict influenza-like illness (ILI) counts

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012

Prediction of counts (2/3)

Woo et al (2016): Support vector regression to predict influenza counts based on unigram counts in social media data
Sparks et al (2017): Predict tweet counts using Poisson regression model
Features such as hour, day of week and day number

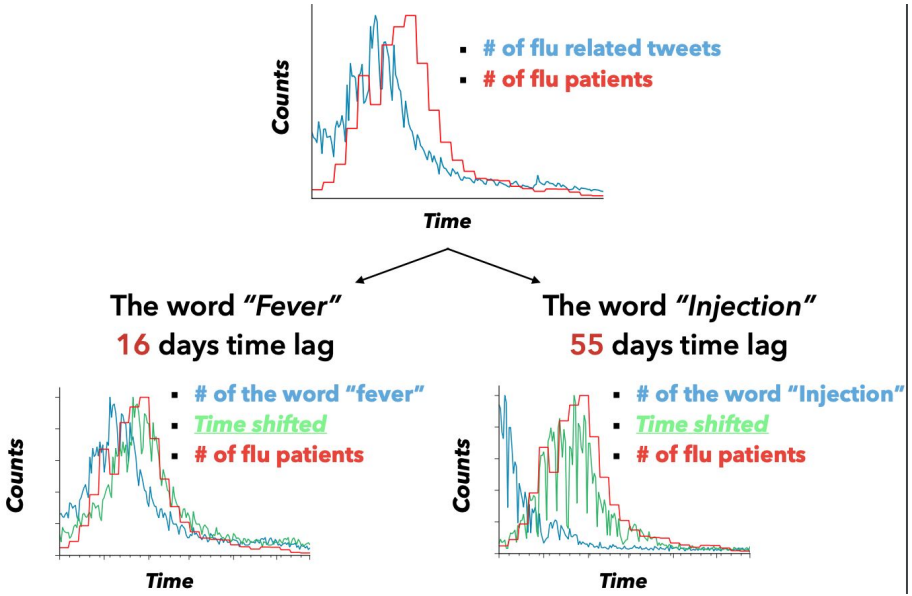
Ross S. Sparks, Bella Robinson, Robert Power, Mark Cameron, and Sam Woolford. 2017. An investigation into social media syndromic monitoring. *Commun. Stat. Simul. Comput.* 46, 8 (2017), 5901–5923.
Hyekyung Woo, Youngtae Cho, Eunyoung Shim, Jong-Koo Lee, Chang-Gun Lee, and Seong Hwan Kim. 2016. Estimating influenza outbreaks using both search engine query data and social media data in South Korea. *J. Med. Internet Res.* 18, 7 (2016).

Prediction of counts (3/3)

Hayate et al (2016): They include factors that capture time lag for different words.

I. S. O. Hayate, Shoko Wakamiya, and Eiji Aramaki. 2016. Forecasting word model: Twitter-based influenza surveillance and prediction. In *Proceedings of the 26th International Conference on Computational Linguistics*. 76–86.

Time lag between infection counts and social media counts



Estimate the expected lag between tweet-keyword counts and the flu counts
Incorporate that lag in flu count prediction based on tweet-keyword counts

So where do we get these counts from?

National Notifiable Diseases Surveillance System (NNDSS) in Australia

- Counts of notifiable diseases by state/year/age group/sex
- A detailed report for laboratory-confirmed influenza cases

Other sources:

Center for Disease Control and Prevention

Prediction of events (1/3)

A known outbreak is validated using public data
Yangarber et al (2008): Use news articles to detect health outbreaks based on a set of health-related keywords
Huang et al (2016): Prevalence of flu and Lyme can be detected 1 week ahead of reported CDC data, based on tweets

Pin Huang, Andrew MacKinlay, and Antonio Jimeno Yepes. 2016. Syndromic surveillance using generic medical entities on Twitter. In *Proceedings of the Australasian Language Technology Association Workshop 2016*. 35–44.
Roman Yangarber, Peter Von Etter, and Ralf Steinberger. 2008. Content collection and analysis in the domain of epidemiology. In Proceedings of the International Workshop on Describing Medical Web Resources (DrMED'08)

Prediction of outbreaks (3/3)

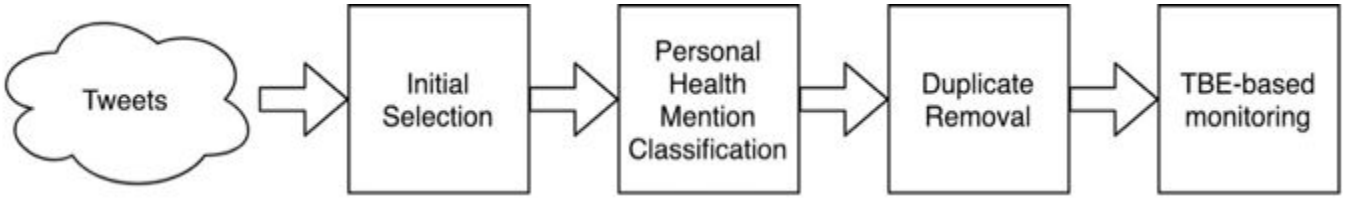
An event/geographical region as the focus
Ebola outbreak in London (Ofoghi et al., 2016)
Across different states of the United States (Zou et al., 2018)

Bahadorreza Ofoghi, Meghan Mann, and Karin Verspoor. 2016. Towards early discovery of salient health threats: A social media emotion classification technique. In *Proceedings of the Pacific Symposium on Biocomputing (Biocomputing'16)*. World Scientific, 504–515.

Bin Zou, Vasileios Lampos, and Ingemar Cox. 2018. Multi-task learning improves disease models from web search. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 87–96

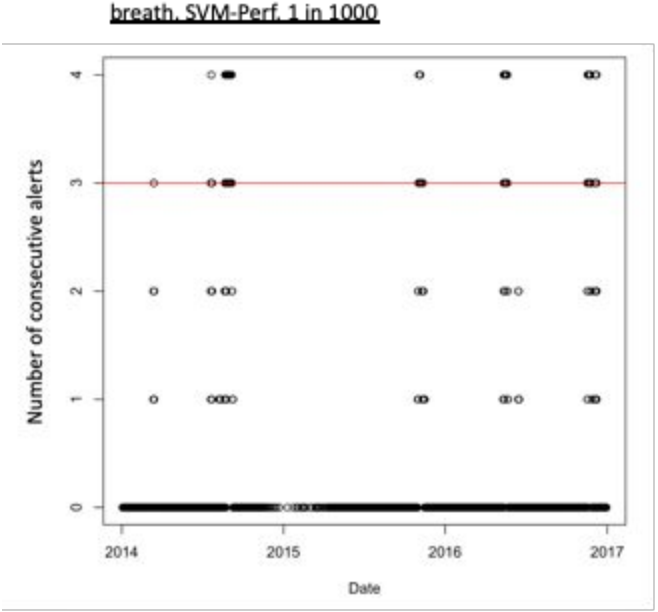
A case study (1/2)

Thunderstorm asthma outbreak in Melbourne: On 21st November 2016, a thunderstorm asthma outbreak occurred in Melbourne, Australia (Thien et al., 2018)



Francis Thien, Paul J Beggs, Danny Csutoros, Jai Darvall, Mark Hew, Janet M Davies, Philip G Bardin, Tony Bannister, Sara Barnes, Rinaldo Bellomo, et al. 2018. The Melbourne epidemic thunderstorm asthma event 2016: an investigation of environmental triggers, effect on health services, and patient risk factors. The Lancet Planetary Health, 2(6):e255–e263.
Aditya Joshi, Ross Sparks, James McHugh, Sarvnaz Karimi, Cecile Paris, C Raina MacIntyre, `Harnessing tweets for early detection of an acute disease event', Epidemiology, Wolters Kluwer Health, 2019.

A case study (2/2)



Official Reports		**First News Report**: 22 Nov 2016, 7:05 [4],	**Official Report**: 21st Nov 2016, 18:00
Dataset	Classifier	False Discovery Rate: 1 in 1000	False Discovery Rate: 1 in 2000
BreathMelbourne	Heur	None	None
CoughMelbourne	Heur	None	None
OtherMelbourne	Heur	**21st Nov 2016, 21:42:09**	**21st Nov 2016, 18:47:17**
BreathMelbourne	Stat-SVM	None	**20 Nov 2016, 19:40:20**
CoughMelbourne	Stat-SVM	None	None
OtherMelbourne	Stat-SVM	None	None
BreathMelbourne	Stat-SVMPerf	**21st Nov 2016, 8:55:19**	**21st Nov 2016, 8:55:19**
CoughMelbourne	Stat-SVMPerf	23rd Nov 2016, 13:27:08	None
OtherMelbourne	Stat-SVMPerf	22nd Nov 2016, 8:52:55	22nd Nov 2016, 10:06:56

Earliest Alert on 21st November, 2016: 8:55:19

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, Typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Roadmap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Module IX: Conclusion

Outline

- Summary
- Future directions

Recap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Datasets

Sources, Annotation, Ethics

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems

Predicting counts, predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Recap

Introduction

History, Challenges, Scope

Classification Pr

Examples, typical approaches

Problems

h, Clinical

Conclusion

Summary, Future Directions

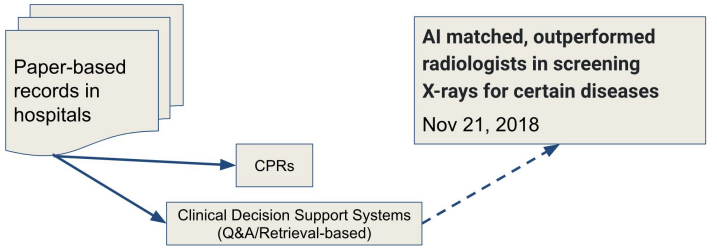
History of healthcare and NLP: Pre-1999

Computerized medication monitoring system (1976)
Sager: NLP system also applied to medical documents (1981)
SPRUS: Radiology text-processor (1994)
SYMTEXT: Automatically generate codes for admission diagnoses (1995)
MedLEE: First NLP system used for actual patient care (1994)
Geneva Hospital: French, English and German documents: Discharge summaries of patients admitted for gastrointestinal surgery (1992-93)
MENELAS: Accessing patient discharge summaries using search (1994)

Friedman, Carol, and George Hripcsak. "Natural language processing and its future in medicine." *Acad Med* 74, no. 8 (1999): 890-5.

15 | NLP for Healthcare in the Absence of a Healthcare Dataset | Sarvnaz Karimi & Aditya Joshi

History: Revisited



19 | NLP for Healthcare in the Absence of a Healthcare Dataset | Sarvnaz Karimi & Aditya Joshi

Image of coffee from wikimedia commons.

Recap

Introduction

History, Challenges, Scope

Datasets

Sources, Annotation, Ethics

Structured Resources

ICD Codes, Ontologies, Using the two

Types of sources

(Velardi et al., 2014)

1) **Demand-based data sources**: This refers to sources that reflect demand for information.
   a) Pros: Aggregate information
   b) Cons: Access may be restricted; limited context
2) **Supply-based data sources**: The data originates on large-scale platforms designed to share information.
   a) Pros: Large-scale information
   b) Cons: The text tends to be longer than search queries, bringing in typical challenges of ambiguity in NLP

Paola Velardi, Giovanni Stilo, Alberto E Tozzi, and Francesco Gesualdo. 2014. Twitter mining for fine-grained syndromic surveillance. Artificial Intelligence in Medicine 61, 3 (2014), 153–163.

QA Problems

QA, shared tasks Q, MediQA)

eries Monitoring Problems

ting counts,

content, Normalisation

Typical Annotation Strategies

Manual: Human annotators are given a set of guidelines and questions

Hybrid: A combination of manual and automatic steps

Retrieval Problems

Literature search, Clinical trial search

Image of coffee from wikimedia commons.

Recap

Introduction

History, Challenges, Scope

Classification Problems

Examples, typical approaches

Datasets

Sources, Annotation, Ethics

Information Extraction Proble

predicting outbreaks

Structured Resources

ICD Codes, Ontologies, Using the two

Conclusion

Summary, Future Directions

Recommendation to use ICD Codes

A patient arrives. → Look up the symptom in Volume 3 → Verify the corresponding code in Volume 1 → Verify classification and reporting rules for the

Tabular list uses British spelling
Alphabetical index uses American spelling to sort, with cross-references

8 | ALTA 2019 Tutorial | Sarvnaz Karimi & Aditya Joshi

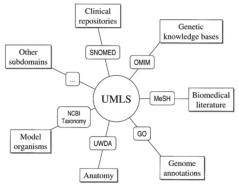
Unified Medical Language System (UMLS)

McCray et al 1989 (https://www.nlm.nih.gov/research/umls/index.html)
15 languages

Metathesaurus: Terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT.

Semantic Network: Broad categories (semantic types) and their relationships (semantic relations).

SPECIALIST Lexicon and Lexical Tools: A large syntactic lexicon of biomedical and general English and tools for normalizing strings, generating lexical variants, and creating indexes.



McCray, Alexa T. "The UMLS Semantic Network." In *Proceedings. Symposium on Computer Applications in Medical Care*, pp. 503-507. American Medical Informatics Association, 1989.

15 | ALTA 2019 Tutorial | Sarvnaz Karimi & Aditya Joshi

Image of coffee from wikimedia commons.

Recap

Typical problem definitions (1/2)

Health mention classification: Predict whether or not a given text is a health mention

"I have had cough all week" versus "Ice creams can cause cough"

Why? To detect health outbreaks; Individual health mentions are events

Challenges: Targets may be important; a celebrity reporting a health symptom may be widely re-tweeted or discussed but does not correspond to a separate event (Kanouchi et al, 2015).

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold?- identifying the subject of a symptom. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Beijing, China), Vol. 1. 1660–1670.

Classification Problems

Examples, typical approaches

Hidden Flu-State Tweet Model

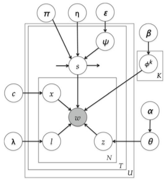
Temporal model (Chen et al., 2016)

Word-level variable (w) each to indicate background words and general general words

Switch (x) between symptom (l) and general (z) words

Tweet-level symptom variable (s)

The symptom variable of a tweet depends on the value of the previous tweet by a user



Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Aditya Prakash. 2016. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Min. Knowl. Discov.* 30, 3 (2016), 681–710.

Information Extraction Problems

NER f... user-... malisa...

Sources, Annotation, Ethics

Predicting counts ...ks

Struc...

ICD Cod...
Using the...

...eval Pr...
...search

...n
...irections

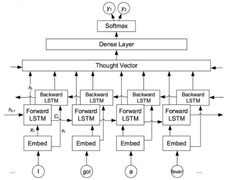
Statistical Approaches

	Classifier	Features
Olszewski (2003)	Naive Bayes	Unigrams and bigrams
Aramaki et al (2011)	Naive Bayes, AdaBoost, SVM	Bag of words with feature windows of multiple sizes
Névéol et al (2009)	Priority model: Probabilistic model	N-grams
Kanouchi et al (2015)		Unigrams, weblinks, word classes, length, n-grams, etc.
Jiang et al. (2016)		Emotion words, emotion scores, user mentions,number, pronominal mentions

Robert T. Olszewski. 2003. Bayesian classification of triage diagnoses for the early detection of epidemics. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*. 412–416.
Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1568–1576.
Aurélie Névéol, Won Kim, W. John Wilbur, and Zhiyong Lu. 2009. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics, 144–152.
Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold?-identifying the subject of a symptom. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 1660–1670.
Keyuan Jiang, Ricardo Calix, and Matrika Gupta. 2016. Construction of a personal experience tweet corpus for health surveillance. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. 128–135.

Deep learning architectures (1/2)

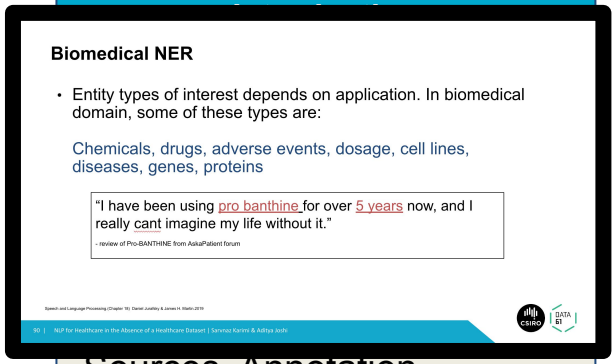
Wang et al (2017) : RNNs



Chen-Kai Wang, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. 2017. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In *Proceedings of the International Workshop on Digital Disease Detection Using Social Media 2017 (DDDSM'17)*. 33–38.

Image of coffee from wikimedia commons.

Recap



Sources, Annotation, Ethics

Classification Problems

Examples, typical approaches

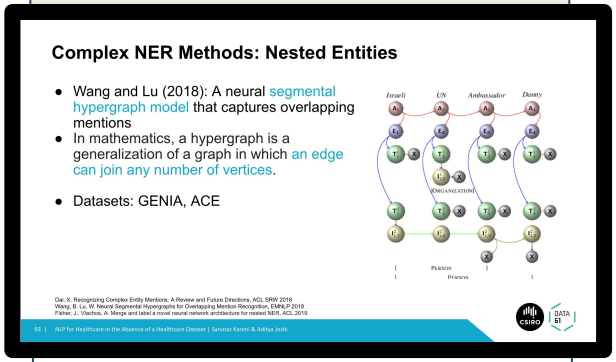
QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Information Extraction Problems

NER from user-generated content, Normalisation

Time Series Monitoring Problems



Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical trial search

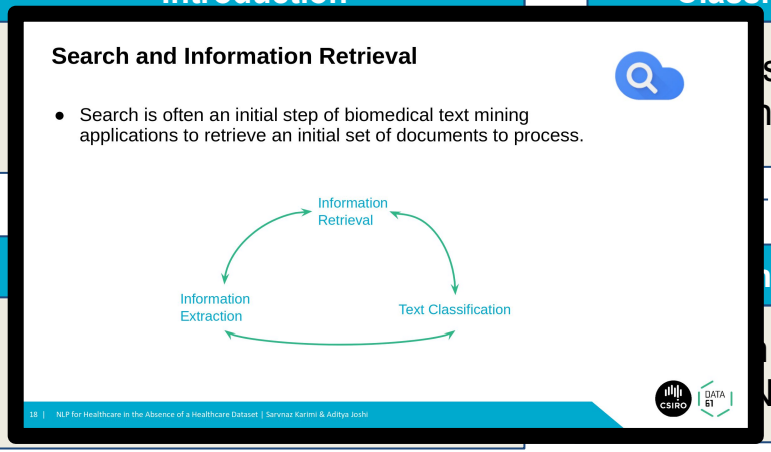
Image of coffee from wikimedia commons.

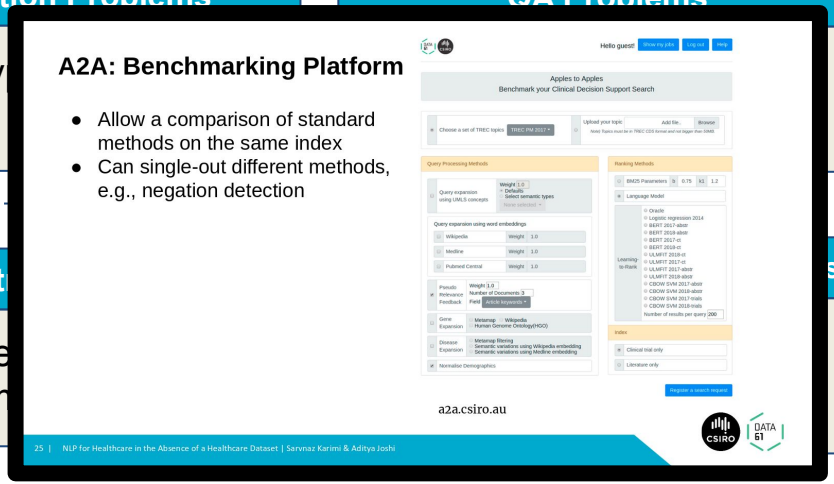
Recap

Introduction

Classification Problems

QA Problems





Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval Problems

Literature search, Clinical Trial Search

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Recap

Classification Problems

ples, typical
aches

QA Problems

Patient QA, shared tasks (BioASQ, MediQA)

Time Series Monitoring Problems

rections

Structured Resources

ICD Codes, Ontologies, Using the two

Retrieval

Literature sear
Trial Search

Patient Question Answering

- Question answering is more specific than IR: Users look for specific answers to their questions
- When it comes to medical area it becomes crucial to have higher confidence in answers or not to provide any
  - Pew (2013) survey: 35% of the US Adults search for their systems and ⅓ of these self-diagnosers do not seek professional help
- Patient often do not know the correct terminology to search over credible sources such as medical journals
- Even if presented relevant resources, may not be able to interpret them correctly

115 | NLP for Healthcare in the Absence of a Healthcare Dataset | Sarvnaz Karimi & Aditya Joshi

MediQA: Example

<QuestionText>
abetalipoproteimemia hi, I would like to know if there is any support for those suffering with abetalipoproteinemia? I am not diagnosed but have had many test that indicate I am suffering with this, keen to learn how to get it diagnosed and how to manage, many thanks
</QuestionText>

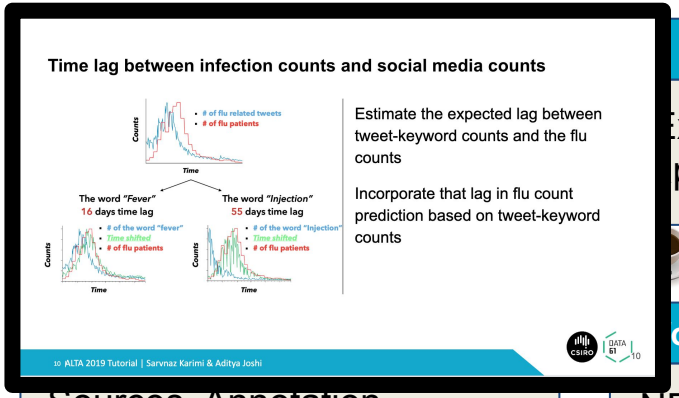
MediQA: https://sites.google.com/view/mediqa2019

Abacha, Asma Ben, Chaitanya Shivade, and Dina Demner-Fushman. Overview of the MediQA 2019 shared task on textual inference, question entailment and question answering. *BioNLP Workshop and Shared Task*, pp. 370-379. 2019.

118 | NLP for Healthcare in the Absence of a Healthcare Dataset | Sarvnaz Karimi & Aditya Joshi

Image of coffee from wikimedia commons.

Recap



Sources, Annotation, Ethics

Classification Problems

Examples, typical approaches

QA Problems

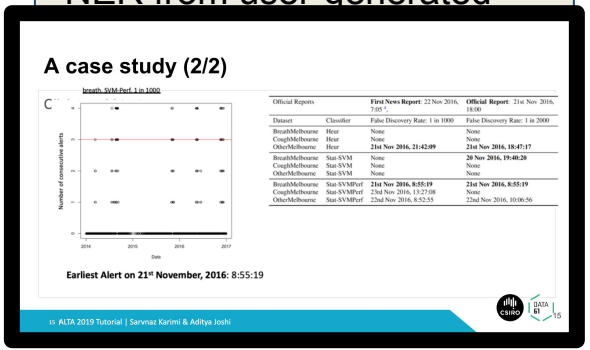
Patient QA, shared tasks (BioASQ, MediQA)

Information Extraction Problems

NER from user-generated

Time Series Monitoring Problems

Predicting counts, predicting outbreaks



Structured Resources

ICD Codes, Ontologies, Using the two

Conclusion

Summary, Future Directions

Image of coffee from wikimedia commons.

Future Directions

Adaptation to different locations and languages

Adaptation to different illnesses

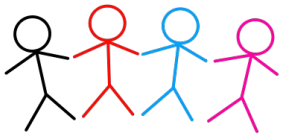
Reliability of information

False alerts due to ambiguity

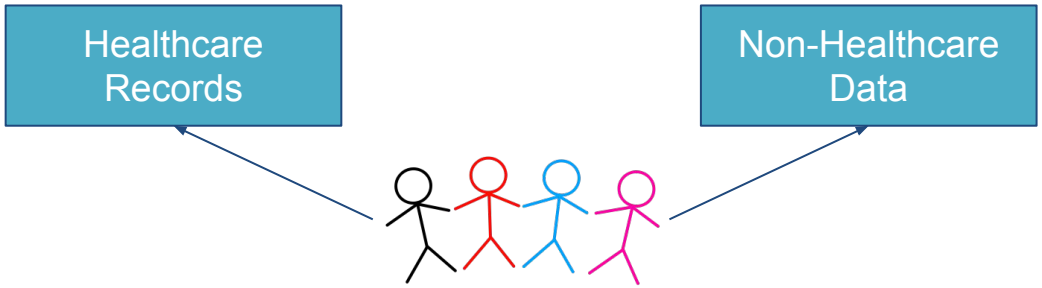
Accounting for changes in social networks

Under-represented health events in social media: Zoonotic diseases?

Final Remarks

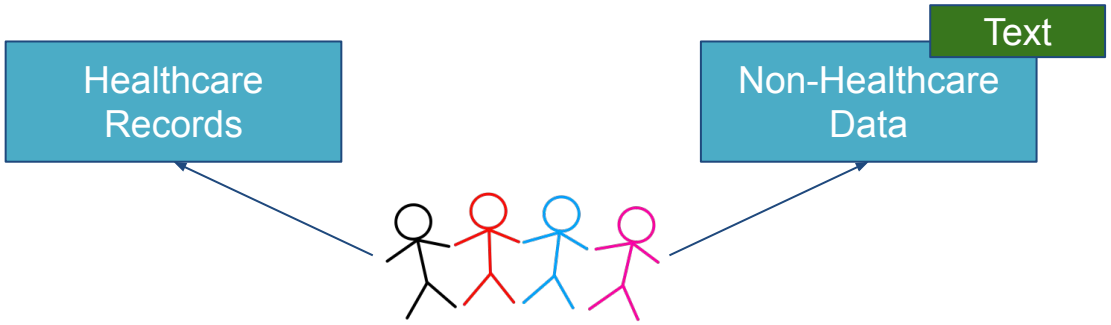


Final Remarks



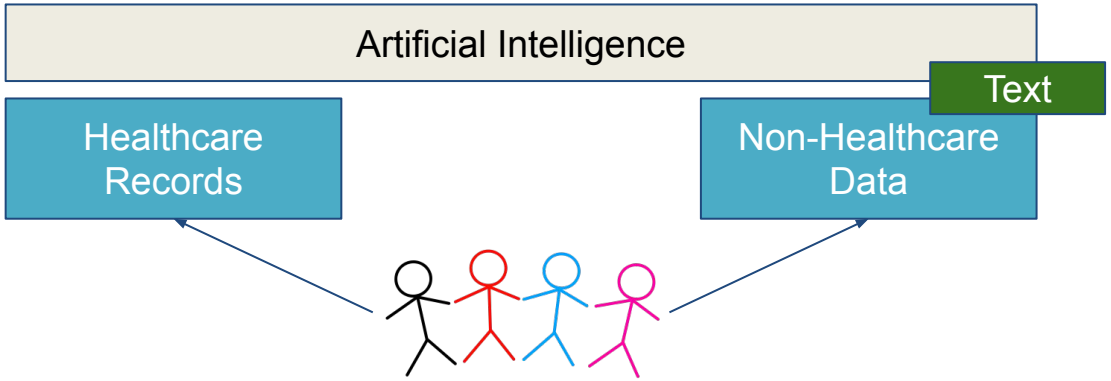
Final Remarks

- Peculiar advantages of a non-healthcare dataset



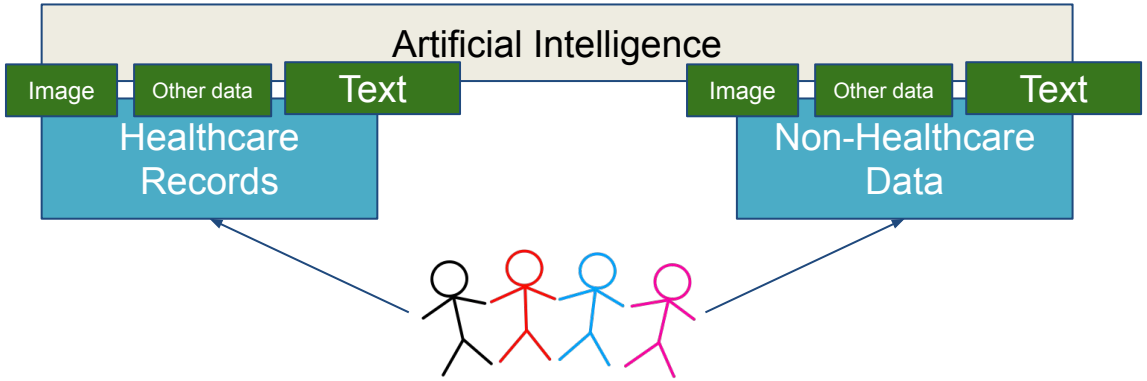
Final Remarks

- Peculiar advantages of a non-healthcare dataset
- Several possibilities of applications



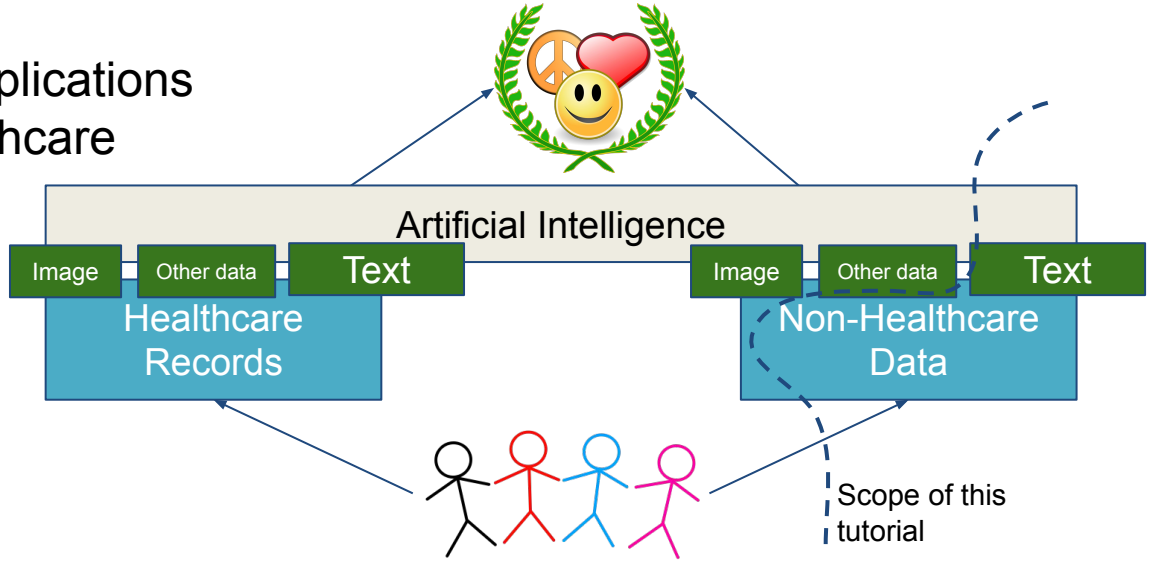
Final Remarks

- Peculiar advantages of a non-healthcare dataset
- Several possibilities of applications
- A piece in the ‘AI for healthcare puzzle’



Final Remarks

- Peculiar advantages of a non-healthcare dataset
- Several possibilities of applications
- A piece in the ‘AI for healthcare puzzle’



THANK YOU

Aditya Joshi
CSIRO Data61

Contact:
aditya.joshi@csiro.au

www.data61.csiro.au