Medical Language Processing

Pierre Zweigenbaum

LIMSI-CNRS, Orsay, France & ERTIM-INALCO, Paris, France

NETTAB 2011 Tutorial
October 12, 2011,
Collegio Ghislieri, Pavia, Italy

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Information Repositories in the Biomedical Domain

Which documents contain interesting medical information/knowledge?

Different types of resources, such as textbooks, prescribing resources, and laboratory handbooks, have different types of information and are thus suited to different types of questions.

(Ely et al., BMJ 2000)

- MEDLINE and Pubmed
- Drug Information
- Clinical Practice Guidelines
- Hospital Information Systems
- Specialized News Feeds

MEDLINE and Pubmed

Database of scientific articles in the Biomedical domain

- About 5,200 journals in 37 languages
- Over 16 million citations (2008)
- Free, online access through PubMed portal since 1996
- Long tradition of *search strategies*

`http://www.ncbi.nlm.nih.gov/pubmed/`

PubMed Access to the Scientic Literature

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Search PubMed for [ ] Go Clear

Advanced Search

Limits | Preview/Index | History | Clipboard | Details

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: Singapore Med J. 2009 Jun;50(6):581-3.

SMJ FREE full text article at www.sma.org.sg/smj

Link

New influenza A (H1N1) 2009 in Singapore: the first ten adult imported cases.

Liang M, **Lye DC**, **Chen MI**, **Chow A**, **Krishnan P**, **Seow E**, **Leo YS**.

Department of Infectious Diseases, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore.

INTRODUCTION: Since late March 2009, a novel influenza H1N1 strain emerged in humans in Mexico and the United States. It has rapidly spread to many countries on different continents, prompting unprecedented activation of pandemic preparedness plans. Singapore has adopted a containment strategy with active screening of febrile travellers with respiratory symptoms from affected countries since April 27, 2009. METHODS: All cases with new influenza A (H1N1) confirmed on polymerase chain reaction assay on combined nasal and throat swabs and who were admitted to the Communicable Disease Centre, were included in a prospective evaluation of clinical characteristics of new influenza A (H1N1). RESULTS: From May 26 to June 3, 2009, there were ten patients with a mean age of 27.6 years, seven of whom were female. All but one travelled from the United States, six of whom travelled from New York; the last one travelled from the Philippines. Clinical illness developed within a mean of 1.4 days after arrival in Singapore, and presentation to the emergency department at a mean of 2.7 days from illness onset. Fever occurred in 90 percent, cough 70 percent, coryza 40 percent, sore throat and myalgia/arthralgia 30 percent; none had diarrhoea. The fever lasted a mean of 2.1 days. All were treated with oseltamivir. The clinical course was uncomplicated in all cases. CONCLUSION: Clinical features of new influenza A (H1N1) appeared mild, and ran an uncomplicated course in immunocompetent patients.

PMID: 19551309 [PubMed - in process]

Related articles

- Update: swine influenza A (H1N1) infections--California and Texas, April 2([MMWR Morb Mortal Wkly Rep. 2009
- Novel influenza A (H1N1) virus infections in three pregnan women - United Sta [MMWR Morb Mortal Wkly Rep. 2009
- Infections with oseltamivir-resistant influenza A(H1N1) virus in the United States. [JAMA. 2009
- Review Influenza and the pandemic threat. [Singapore Med J. 2006
- Review [Influenza--always present among us] [Med Pregl. 200

» See reviews... | » See all

Patient Drug Information

- Oseltamivir (Tamiflu®) Oseltamivir is used to treat some types of influenza infection ('flu') in adults and children (older than 1 year of age) who have had

Source: AHFS Consumer Medication Informatio

Recent Activity

Link to Some Full-Text Journal Articles

Depending on journal publisher and article

New influenza A (H1N1) 2009 in Singapore: the first ten adult imported cases

Liang M, Lye D C, Chen M I, Chow A, Krishnan P, Seow E, Leo Y S

ABSTRACT

Introduction: Since late March 2009, a novel influenza H1N1 strain emerged in humans in Mexico and the United States. It has rapidly spread to many countries on different continents, prompting unprecedented activation of pandemic preparedness plans. Singapore has adopted a containment strategy with active screening of febrile travellers with respiratory symptoms from affected countries since April 27, 2009.

Methods: All cases with new influenza A (H1N1) confirmed on polymerase chain reaction assay on combined nasal and throat swabs and who were admitted to the Communicable Disease Centre, were included in a prospective evaluation of clinical characteristics of new influenza A (H1N1).

Results: From May 26 to June 3, 2009, there were ten patients with a mean age of 27.6 years, seven of whom were female. All but one travelled from the United States, six of whom travelled from New York; the last one travelled from the Philippines. Clinical illness developed within a mean of 1.4 days after arrival in Singapore, and presentation to the emergency department at a mean of 2.7 days from illness onset. Fever occurred in 90 percent, cough 70 percent, coryza 40 percent, sore throat and myalgia/arthralgia 30 percent; none had diarrhoea. The fever lasted a mean of 2.1 days. All were treated with oseltamivir. The clinical course was uncomplicated in all cases.

Conclusion: Clinical features of new influenza A (H1N1) appeared mild, and ran an uncomplicated course in immunocompetent patients.

(H1N1) 2009 was notified to the World Health Organisation (WHO), it has spread to 74 countries, with 29,669 cumulative cases and 145 deaths with a case-fatality ratio of 0.5% (as of June 12, 2009).[1] As of June 10, 2009, local transmission was noted in 20 countries with death in four countries. In Asia, cases have been reported from Japan, China, South Korea, Taiwan, Hong Kong and India, and the Southeast Asian countries of Singapore, Malaysia, Thailand, Vietnam and the Philippines. In Asia and Southeast Asia, local transmission has thus far been documented in Japan, China, South Korea, Taiwan, India, Singapore, Thailand, Vietnam and the Philippines. Singapore started screening febrile travellers with respiratory symptoms from affected countries for the new influenza A (H1N1) since April 27, 2009, and the first case was detected on May 26, 2009. We present the epidemiology, clinical illness and treatment outcome of the first ten cases, diagnosed and treated at Communicable Disease Centre (CDC) 2, Tan Tock Seng Hospital (TTSH), Singapore, which is the designated screening centre for new influenza A (H1N1).

METHODS

All confirmed cases of new influenza A (H1N1) treated at TTSH, Singapore, from April 27, 2009 and who were admitted to the Communicable Disease Centre were included in a prospective evaluation of the clinical and virological characteristics of their infection. Baseline and daily clinical data, including demographical, travel and exposure history, comorbidity, symptoms and signs, were collected until discharge. On admission, all patients had full blood count, renal and liver functions, and C-reactive protein assayed, as well as chest radiography done. Virological study included serial sampling of the nose and throat to examine viral shedding. Laboratory diagnosis of new influenza A (H1N1) was made by probe-based polymerase chain reaction (PCR) on combined nasal and throat swabs and confirmed with sequencing of the M gene

Department of Infectious Diseases, Tan Tock Seng Hospital, 11 Jalan Tan Tock Seng, Singapore 308433

Liang M, MBBS
Medical Officer

Lye DC, MBBS, FRACP
Consultant

Leo YS, MBBS, MMed, FRCP
Senior Consultant and Head

Department of Clinical Epidemiology

Chen MI, MBBS, PhD
Registrar

Chow A, MBBS, MPH
Consultant

Department of Laboratory Medicine

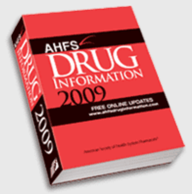
Krishnan P, MBBS, DTM&H, FRCPath
Senior Consultant and Head

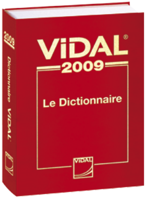
Department of Emergency Medicine

Seow E, MBBS

Drug Information

Structured information sources are also available





Clinical Practice Guidelines

- Authoritative, public documents
- Recommendations for best practices
- Based on a systematic review of current evidence

Clinical Practice Guidelines: Examples

Different types of documents

Long text documents (mostly)

3. Renal-specific aspects of a pandemic

a) Specific risks to renal patients of influenza infection and its complications

In conventional influenza infection, patients at increased risk of complications are considered to be: those aged 65 years or older; long stay residential care home residents; and those with: chronic respiratory diseases; chronic heart disease; chronic kidney disease, nephrotic syndrome and established renal failure; chronic liver disease; diabetes, and immuno-compromised patients. It should be noted that for kidney patients, it is hard to find good published evidence that renal patients are indeed at such increased risk.

The particular complications associated with high morbidity and mortality are influenza pneumonitis and secondary bacterial pneumonia caused by *Streptococcus pneumoniae* or *Haemophilus influenzae*, or sometimes *Staphylococcus aureus* or other healthcare-associated organisms.

Influenza virus infection has also been associated with worsening in the clinical condition of patients with a range of existing medical conditions, such as heart failure, diabetes, coronary heart disease, asthma and chronic obstructive pulmonary disease (COPD).

Finally, patients with pre-existing CKD are at risk of pre-renal exacerbation through pyrexia, poor fluid intake from anorexia and sore throat, diarrhoea (which has been reported in a high proportion of avian flu sufferers), and NSAIDs used by patients for treatment of myalgias and headaches.

Thus renal patients, many of whom have the above listed comorbidities or risk factors, are likely to be more at risk of serious morbidity and mortality during a pandemic. This will result in additional and perhaps disproportionate pressure on renal units where the skills for caring for these patients are concentrated.

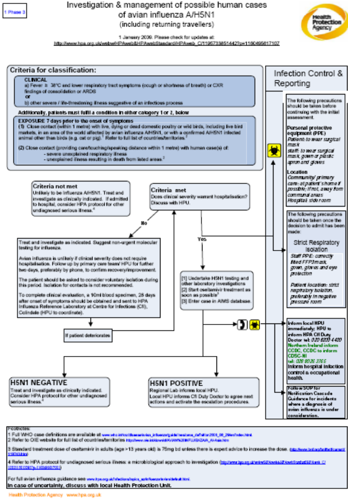
b) Staffing issues

Up to 50% of the workforce may require time off work at some stage over the entire period of the pandemic. Staff absence from work will be not just due to personal influenza infection, but also to provide care for dependants (whether ill relatives, or children as a result of likely school closures), family bereavement, other psychosocial impacts, fear of infection and/or practical difficulties in getting to work. At the peak of the pandemic, between 15% and 20% of staff may be absent at any one time.

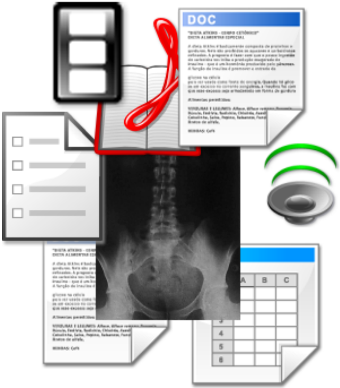
All hospital doctors, whatever their base specialty, are likely to be involved in the care of patients with influenza. Nephrologists (because they have general skills) will need to be prepared to help out in other clinical areas where possible.

As all elective in-patient and non-urgent out-patient activity will be cancelled during the pandemic, new working patterns and responsibilities will need to be brought in to place to cope with the demands of the acute in-patient workload.

Schemata, algorithms

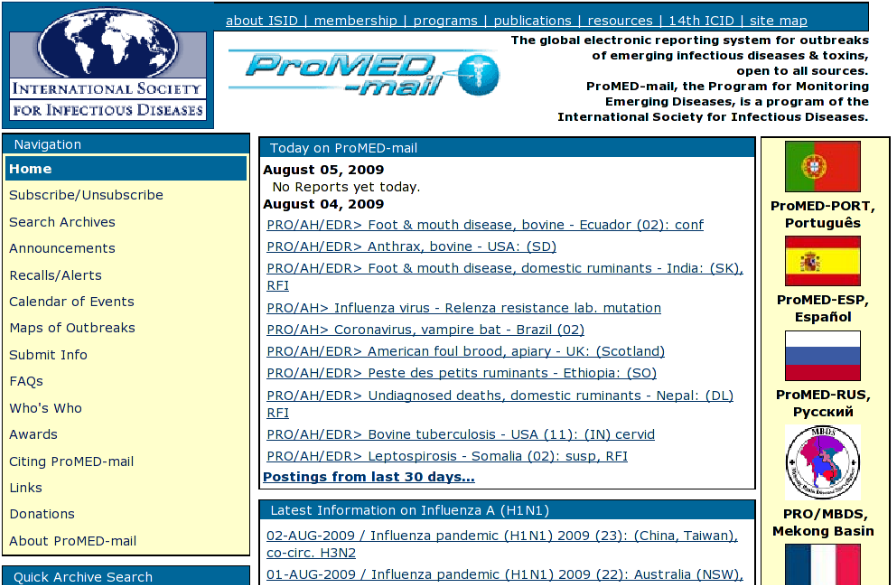


Hospital Information Systems



- Millions of patient records
- Include a large number of text documents
- Restricted access

News Feeds Specialized in Health Information



about ISID | membership | programs | publications | resources | 14th ICID | site map

INTERNATIONAL SOCIETY FOR INFECTIOUS DISEASES

ProMED-mail

The global electronic reporting system for outbreaks of emerging infectious diseases & toxins, open to all sources. ProMED-mail, the Program for Monitoring Emerging Diseases, is a program of the International Society for Infectious Diseases.

Navigation

- **Home**
- Subscribe/Unsubscribe
- Search Archives
- Announcements
- Recalls/Alerts
- Calendar of Events
- Maps of Outbreaks
- Submit Info
- FAQs
- Who's Who
- Awards
- Citing ProMED-mail
- Links
- Donations
- About ProMED-mail

Quick Archive Search

Today on ProMED-mail

August 05, 2009

No Reports yet today.

August 04, 2009

PRO/AH/EDR> Foot & mouth disease, bovine - Ecuador (02): conf

PRO/AH/EDR> Anthrax, bovine - USA: (SD)

PRO/AH/EDR> Foot & mouth disease, domestic ruminants - India: (SK), RFI

PRO/AH> Influenza virus - Relenza resistance lab. mutation

PRO/AH> Coronavirus, vampire bat - Brazil (02)

PRO/AH/EDR> American foul brood, apiary - UK: (Scotland)

PRO/AH/EDR> Peste des petits ruminants - Ethiopia: (SO)

PRO/AH/EDR> Undiagnosed deaths, domestic ruminants - Nepal: (DL) RFI

PRO/AH/EDR> Bovine tuberculosis - USA (11): (IN) cervid

PRO/AH/EDR> Leptospirosis - Somalia (02): susp, RFI

Postings from last 30 days...

Latest Information on Influenza A (H1N1)

02-AUG-2009 / Influenza pandemic (H1N1) 2009 (23): (China, Taiwan), co-circ. H3N2

01-AUG-2009 / Influenza pandemic (H1N1) 2009 (22): Australia (NSW),

ProMED-PORT, Português

ProMED-ESP, Español

ProMED-RUS, Русский

PRO/MBDS, Mekong Basin

Specialized News Feeds

Archive Number 20090730.2673

Published Date 30-JUL-2009

Subject PRO/EDR> Malaria, autochthonous - Singapore

MALARIA, AUTOCHTHONOUS - SINGAPORE
****
A ProMED-mail post
<http://www.promedmail.org>
ProMED-mail is a program of the
International Society for Infectious Diseases
<http://www.isid.org>

[1]
Date: Wed 29 Jul 2009
Source: The Strait Times [edited]
<http://www.straitstimes.com/Breaking%2BNews/Singapore/Story/STIStory_409802.html>
and Ministry of Health, Singapore [edited]
<http://www.moh.gov.sg/mohcorp/pressreleases.aspx?id=22682>

Outbreak of suspected vivax malaria continues to spread in Singapore
----
The Ministry of Health (MOH) is currently
investigating a 3rd malaria cluster involving 4
cases of suspected local transmission, near a row
of shophouses located at the junction of
Sembawang Road and Admiralty Road East. The 1st
case is a 24-year-old Singaporean woman who works
in the area. Her illness started on 30 Jun 2009
and she was admitted to hospital on 20 Jul 2009.

- Texts in Biomedicine
- Introduction to NLP
  - Morphology: from characters to words
  - Syntax: part-of-speech tagging, sentence parsing
  - Semantics: entities, semantic roles and relations
- Types of Methods
  - Knowledge-based methods
  - Machine-learning-based methods
  - Hybrid methods
  - Dependence on language-specific resources
- Tasks and methods in biomedical NLP
  - Expert-based method: Extraction of prescription information
  - Expert-based method: De-identification
  - Data-driven methods for medical entity recognition
  - Normalization, co-reference
  - Detection of medical relations: binary relations

From bytes to characters: Character encoding

426568C3A76574277320737396E64726F6D6520696E20544E462DCEB10A

426568C3A765742773 73796E64726F6D65 696E 544E462DCEB1

B e h C3A7e t ' s s y n d r o m e i n T N F - CEB1

B e h ç e t ' s s y n d r o m e i n T N F - α

Behçet's syndrome in TNF-α

- A string of bytes
- Definition of space → "tokens"
- ASCII → characters
- Unicode standard for characters → (many) more characters
- In plain characters

From bytes to characters: Character encoding

```
426568C3A76574277320737796E64726F6D6520696E20544E462DCEB10A
426568C3A765742773  73796E64726F6D65  696E  544E462DCEB1
B e h C3A7e t ' s   s y n d r o m e   i n   T N F - CEB1
B e h ç   e t ' s   s y n d r o m e   i n   T N F - α
```

Behçet's syndrome in TNF-α

- A string of bytes
- Definition of space $\longrightarrow$ "tokens"
- ASCII $\longrightarrow$ characters
- Unicode standard for characters $\longrightarrow$ (many) more characters
- In plain characters

From bytes to characters: Character encoding

```
426568C3A7657427732073796E64726F6D6520696E20544E462DCEB10A
426568C3A765742773  73796E64726F6D65  696E  544E462DCEB1
B e h C3A7e t ’ s   s y n d r o m e   i n   T N F - CEB1
B e h ç   e t ’ s   s y n d r o m e   i n   T N F - α
```

Behçet’s syndrome in TNF-α

- A string of bytes
- Definition of space $\longrightarrow$ “tokens”
- ASCII $\longrightarrow$ characters
- Unicode standard for characters $\longrightarrow$ (many) more characters
- In plain characters

From bytes to characters: Character encoding

```
426568C3A765742773207379 6E64726F6D6520696E20544E462DCEB10A
```

426568C3A7657427732073796E64726F6D6520696E20544E462DCEB10A

426568C3A765742773 73796E64726F6D65 696E 544E462DCEB1

B e h C3A7e t ’ s s y n d r o m e i n T N F - CEB1

B e h ç e t ’ s s y n d r o m e i n T N F - α

Behçet’s syndrome in TNF-α

- A string of bytes
- Definition of space → “tokens”
- ASCII → characters
- Unicode standard for characters → (many) more characters
- In plain characters

From bytes to characters: Character encoding

```
426568C3A7657427732073796E64726F6D6520696E20544E462DCEB10A
426568C3A765742773  73796E64726F6D65  696E  544E462DCEB1
B e h C3A7e t ’ s   s y n d r o m e   i n   T N F - CEB1
```

B e h ç e t ’ s   s y n d r o m e   i n   T N F - α

Behçet’s syndrome in TNF-α

- A string of bytes
- Definition of space $\longrightarrow$ “tokens”
- ASCII $\longrightarrow$ characters
- Unicode standard for characters $\longrightarrow$ (many) more characters
- In plain characters

Token and word segmentation

What is a word?

Behçet's syndrome in TNF-α

Segmentation into "tokens" (tokenization)

Behçet
's
syndrome
in
TNF
-
α

Segmentation into "words" (word segmentation)

Behçet's syndrome
in
TNF-α

Token and word segmentation

What is a word?

Behçet's syndrome in TNF-α

Segmentation into "tokens" (tokenization)

Behçet
's
syndrome
in
TNF
-
α

Segmentation into "words" (word segmentation)

Behçet's syndrome
in
TNF-α

Morphology

- Words may have variable forms (inflection):
  - *be/is/are, cherry/cherries, phenomenon/phenomena*
- Different words may have form and meaning relationships:
  - Derived words: *abdomen/abdominal, eye/ocular, hear/audition*
  - Compound words: *inflammation+liver = hepatitis*

Morphological processing: Tasks

- Lemmatization: reduce inflected form to canonical form
  - *is/are* $\longrightarrow$ *be*, *cherries* $\longrightarrow$ *cherry*, *phenomena* $\longrightarrow$ *phenomenon*
- Morphological analysis: given derived or compound word, compute base / stem(s)
  - Derived words: *abdominal* $\longrightarrow$ *abdomen*, *ocular* $\longrightarrow$ *eye*, *audition* $\longrightarrow$ *hear*
  - Compound words: *hepatitis* $\longrightarrow$ *inflammation+liver*

Morphological processing: Methods

- Approximate methods:
  - Approximate string matching:
    - distance(*abdomen*, *abdominal*) = 2/9
  - Stemming: remove inflection marks and suffixes (*e.g.*, Snowball)
    - cherries $\longrightarrow$ cherri
    - cherry $\longrightarrow$ cherri
    - abdominal$\longrightarrow$ abdomin
    - abdomen $\longrightarrow$ abdomen
- Knowledge-based methods:
  - Morphological analysis (*e.g.*, lvg, MetaMap; Dérif)

Syntax: Part-of-speech and ambiguity

Part-of-speech: a first level of syntactic categorization

Part-of-speech (lexical category): Noun, Verb, Adjective, Adverb, Preposition, etc.

Issue: Many words are ambiguous

Time flies like an arrow.
time: noun, verb
flies: verb, noun
like: preposition, verb, noun

The patient presented with a chief complaint

patient: noun, adjective
presented: verb (preterit), verb (past participle)
chief: adjective, noun

Syntax: Part-of-speech and ambiguity

Part-of-speech: a first level of syntactic categorization

Part-of-speech (lexical category): Noun, Verb, Adjective, Adverb, Preposition, etc.

Issue: Many words are ambiguous

Time flies like an arrow.
time: noun, verb
flies: verb, noun
like: preposition, verb, noun

The patient presented with a chief complaint

patient: noun, adjective
presented: verb (preterit), verb (past participle)
chief: adjective, noun

Syntax: Part-of-speech and ambiguity

Part-of-speech: a first level of syntactic categorization

Part-of-speech (lexical category): Noun, Verb, Adjective, Adverb, Preposition, etc.

Issue: Many words are ambiguous

Time flies like an arrow.
time: noun, verb
flies: verb, noun
like: preposition, verb, noun

The patient presented with a chief complaint

patient: noun, adjective
presented: verb (preterit), verb (past participle)
chief: adjective, noun

Syntax: Part-of-speech tagging

Goal: determine the correct part-of-speech of each word in the context of the current sentence

Usual hypothesis: this can be done by looking at a limited context around each word

Methods:

- Based on linguistic knowledge: rules, transducers
- Data-driven: HMM, MaxEnt, CRF, etc.

Tools: TreeTagger, Banner, MedPost...

Example output:

the	DT
patient	NN
presented	VBD
with	IN
a	DT
chief	JJ
complaint	NN

Syntax: Part-of-speech tagging

Goal: determine the correct part-of-speech of each word in the context of the current sentence

Usual hypothesis: this can be done by looking at a limited context around each word

Methods:

- Based on linguistic knowledge: rules, transducers
- Data-driven: HMM, MaxEnt, CRF, etc.

Tools: TreeTagger, Banner, MedPost...

Example output:

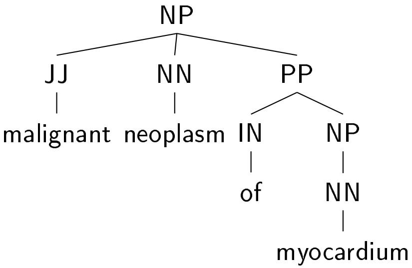
the	DT
patient	NN
presented	VBD
with	IN
a	DT
chief	JJ
complaint	NN

Syntax: Syntactic structure

Structural relations within a sentence

Representation:

- Constituent tree



- Dependency tree

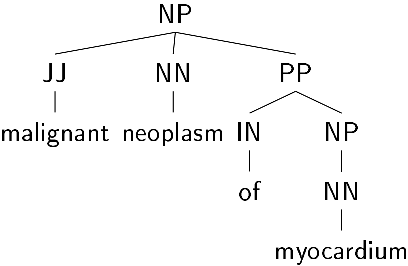
And also: Grammatical relations: subject, object, modifier...

Syntax: Syntactic structure

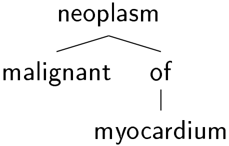
Structural relations within a sentence

Representation:

- Constituent tree



- Dependency tree



And also: Grammatical relations: subject, object, modifier. . .

Syntax: Parsing

Goal: determine the correct syntactic structure of a sentence (or sentence fragment)

Example output:

(NP (JJ malignant) (NN neoplasm) (PP (IN of) (NN neoplasm)))

Overall architecture: parser (= engine) + grammar (= knowledge)

Grammar models: finite state automaton / regular rules, recursive FSA / context-free rules, unification grammar, LFG, HPSG...

Grammar development methods:

- Expert-based: Linguist's knowledge
- Data-driven: PCFG, PCFG-LA, reranking...

Tools: Stanford parser, Berkeley parser, GENIA parser, ...

Syntax: Parsing

Goal: determine the correct syntactic structure of a sentence (or sentence fragment)

Example output:

(NP (JJ malignant) (NN neoplasm) (PP (IN of) (NN neoplasm)))

Overall architecture: parser (= engine) + grammar (= knowledge)

Grammar models: finite state automaton / regular rules, recursive FSA / context-free rules, unification grammar, LFG, HPSG...

Grammar development methods:

- Expert-based: Linguist's knowledge
- Data-driven: PCFG, PCFG-LA, reranking...

Tools: Stanford parser, Berkeley parser, GENIA parser, ...

Syntax: Parsing

Goal: determine the correct syntactic structure of a sentence (or sentence fragment)

Example output:

(NP (JJ malignant) (NN neoplasm) (PP (IN of) (NN neoplasm)))

Overall architecture: parser (= engine) + grammar (= knowledge)

Grammar models: finite state automaton / regular rules, recursive FSA / context-free rules, unification grammar, LFG, HPSG...

Grammar development methods:

- Expert-based: Linguist's knowledge
- Data-driven: PCFG, PCFG-LA, reranking...

Tools: Stanford parser, Berkeley parser, GENIA parser, ...

Syntax: Parsing

Goal: determine the correct syntactic structure of a sentence (or sentence fragment)

Example output:

(NP (JJ malignant) (NN neoplasm) (PP (IN of) (NN neoplasm)))

Overall architecture: parser (= engine) + grammar (= knowledge)

Grammar models: finite state automaton / regular rules, recursive FSA / context-free rules, unification grammar, LFG, HPSG...

Grammar development methods:

- Expert-based: Linguist's knowledge
- Data-driven: PCFG, PCFG-LA, reranking...

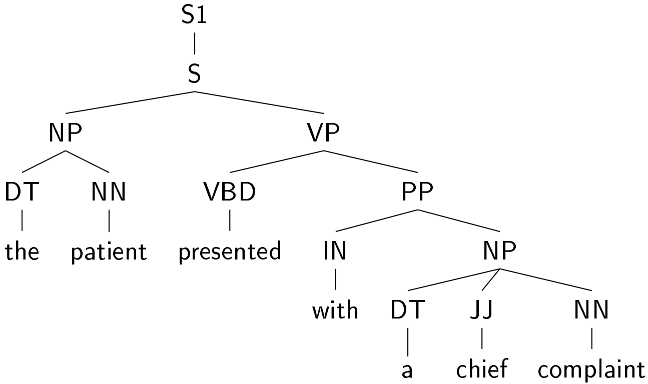
Tools: Stanford parser, Berkeley parser, GENIA parser, ...

Syntactic structure: Example output

- Linear form

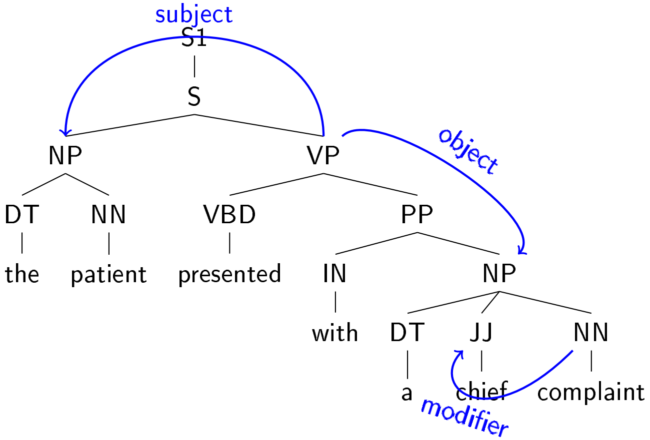
```
(S1 (S (NP (DT the) (NN patient))
  (VP (VBD presented)
    (PP (IN with) (NN (DT a) (JJ chief) (NN complaint))))))
```

- Constituent tree

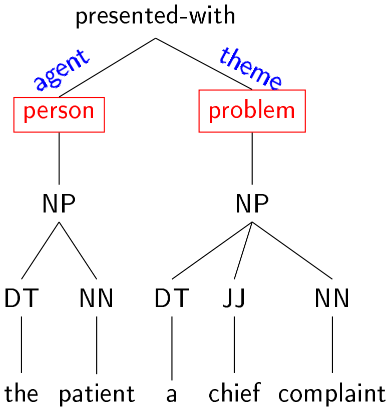


Syntactic structure: Grammatical relations

- Constituent tree + grammatical relations



Semantics: entities, semantic roles



Semantics: entities and relations

Test reveals Problem

On physical examination , the patient is afebrile .

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Knowledge-based methods

Human-knowledge-driven methods

- Human experts formalize and encode their knowledge

  - Knowledge on language: linguists
  - Knowledge on domain: domain experts

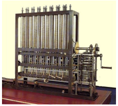
$\longrightarrow$ generally reliable results (if enough time provided)

$\longrightarrow$ time consuming

$\longrightarrow$ requires expensive expertise

Machine-learning-based methods

Data-driven methods



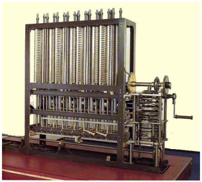
- **Machine-learning approaches:**
  - Start from examples of what must be annotated and how: a (large) annotated corpus
    - Representation of problem: features, structure
  - Learn how to determine the right annotation for unseen input

$\longrightarrow$ needed expertise for corpus annotation generally less expensive

$\longrightarrow$ provides a fast path to results once large enough corpora have been annotated

$\longrightarrow$ may keep a strong dependence on the training corpus

Hybrid methods



- Run both then combine results
- Pre: Use knowledge to prepare a better representation of the problem
- Post: Use knowledge to correct errors of machine-learning-based system

Dependence on language-specific resources

Need to find or develop language- and domain-related knowledge bases

- Lexicons
  - General lexicons
  - Specific lexicons, e.g. Verb classes (Verbnet)[1] or sentiment information SentiWordNet [2]
- Terminologies and Thesauri
  - General language: WordNet[3]
  - Medical language: Unified Medical Language System (UMLS)[4]
- Importance of "Language Resources" (and Evaluation)
  - LREC international conference
  - LRE journal

[1] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[2] http://sentiwordnet.isti.cnr.it/
[3] http://wordnet.princeton.edu/
[4] http://www.nlm.nih.gov/research/umls/

UMLS: Unified Medical Language System

Objective

Facilitate search and integration of information from multiple electronic sources of biomedical information.

- Metathesaurus : includes and links over 100 biomedical terminologies
- 2.2 million concepts, 7.2 million distinct terms (2010AA)
- Freely distributed resource (but observe rights restrictions)
  http://www.nlm.nih.gov/research/umls/

→ A must in any biomedical language processing work ...

→ ... including information extraction

UMLS: Unified Medical Language System

Objective

Facilitate search and integration of information from multiple electronic sources of biomedical information.

- Metathesaurus : includes and links over 100 biomedical terminologies
- 2.2 million concepts, 7.2 million distinct terms (2010AA)
- Freely distributed resource (but observe rights restrictions)
  http://www.nlm.nih.gov/research/umls/

→ A must in any biomedical language processing work . . .

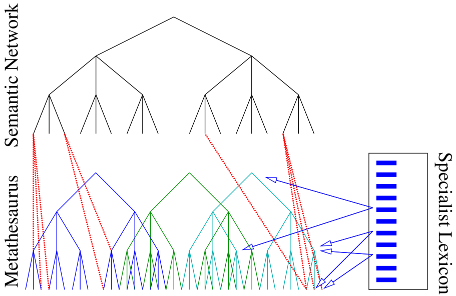
→ . . . including information extraction

UMLS Components

Metathesaurus: Systematized union of over a hundred biomedical terminologies

Semantic Network: Unifying structuring superimposed over these terminologies

Specialist Lexicon: Morphosyntactic information about (biomedical) terms



Languages in the UMLS Metathesaurus

Languages: UMLS 2010AA

Language	*Unique strings*
ENG	5193854
SPA	1046877
JPN	210847
DUT	184722
FRE	156049
GER	150522
POR	119097
RUS	104321
ITA	102385
CZE	97667
	.../...

Language	*Unique strings*
.../...	
SWE	26209
FIN	25385
KOR	10951
SCR	8228
LAV	1391
DAN	697
NOR	697
HUN	684
BAQ	675
HEB	485

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Named entities: names, dates, locations, instances of biomedical concepts

This is to notify you that your patient, Gianni Di Maggio, arrived in the Emergency Department at Pavia's Hospital on Oct 12, 2011. The patient presented with a chief complaint of shortness of breath and a dry non-productive cough.

- Entities: patient, doctor, date, hospital, ward, medical problem, test, treatment...

Named entities: names, dates, locations, instances of biomedical concepts

This is to notify you that your patient, Gianni Di Maggio, arrived in the Emergency Department at Pavia's Hospital on Oct 12, 2011. The patient presented with a chief complaint of shortness of breath and a dry non-productive cough.

- Entities: patient, doctor, date, hospital, ward, medical problem, test, treatment. . .

Named entities: names, dates, locations, instances of biomedical concepts

This is to notify you that your patient, Gianni Di Maggio, arrived in the Emergency Department at Pavia's Hospital on Oct 12, 2011. The patient presented with a chief complaint of shortness of breath and a dry non-productive cough.

- Entities: patient, doctor, date, hospital, ward, medical problem, test, treatment. . .

An example of human-knowledge-based methods

The COKAINE System (Louise Deléger, Cyril Grouin, JAMIA 2010)

COrpus- and
Knowledge-based
Automatic
INformation
Extraction

Drug Prescriptions in Patient Records

Find the medications; Find the details of drug prescriptions

A patient record (excerpt)

PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.

DISCHARGE MEDICATIONS: Captopril 50 mg p.o. q.i.d.; Isordil 20 mg p.o. t.i.d.; Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of Lasix that day; Lotrimin 1% cream topical b.i.d.; and digoxin 0.25 mg p.o. q. day.

DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Drug Prescriptions in Patient Records

Drug: Dosage + Mode of administration + Frequency + Duration + Reason

Detected Information

PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
DISCHARGE MEDICATIONS: Captopril 50 mg p.o. q.i.d.; Isordil 20 mg p.o. t.i.d.; Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of Lasix that day; Lotrimin 1% cream topical b.i.d.; and digoxin 0.25 mg p.o. q. day.
DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

How to Detect a Drug Prescription?

'Simple' Way: Lexicons + Patterns

- Which words reveal drug prescriptions?
  - Lexicons of drug names
  - Lexicons of dosage units
  - Lists of abbreviations
- Which forms of expressions reveal drug prescriptions?
  - Hand-designed patterns (regular expressions)
  - E.g., form of a dosage: <number> <dosage_unit>

Lexicons

Several types of lexicons are needed

1. **drug lexicons** to detect drug names
2. **sign and symptom lexicons** to identify the reason why a given medication was prescribed
3. **lists of abbreviations and expressions** to extract drug-related information: dosage, mode of administration, frequency, duration

Shopping list

Lexicons

Several types of lexicons are needed

1. **drug lexicons** to detect drug names
2. **sign and symptom lexicons** to identify the reason why a given medication was prescribed
3. **lists of abbreviations and expressions** to extract drug-related information: dosage, mode of administration, frequency, duration

Shopping list

Where to Find Drug Lexicons

1.Drug lexicons

- Two lists created to detect drug names:
  - **FDA + RxList websites**: 8,923 drug names (generic and trade names) $\longrightarrow$ maximize precision
  - **UMLS Metathesaurus** (only for "*Clinical Drug*" and "*Pharmacologic Substance*" semantic types): 180,089 terms (after cleaning of data) $\longrightarrow$ maximize recall
- Combined with a list of 102 *therapeutic classes*.

Where to Find Sign and Symptoms Lexicons

2.Sign and symptoms lexicons

- Three lexicons created from the UMLS:
  - "*Signs and Symptoms*" semantic type: 19,718 entries → maximize recall
  - "*MetaMap NLP View*" flagged terms (useful for NLP): 9,027 terms → maximize precision
  - "*Disorders*" semantic type: much too noisy in our first experiments, gave it up
- These lexicons are used to identify the reason why a given medication was prescribed.

Other, Smaller Lexicons

3.Lists of abbreviations and expressions

- Elements used in drug-related information
- Typographic variants taken into account
- Each entry is associated with the type of information it denotes:
  - *mg* ⟶ dosage
  - *sliding scale* ⟶ dosage
  - *iv* ⟶ mode of administration 
  - *intramuscular* ⟶ mode of administration
  - *qd* ⟶ frequency 

  - *prn* ⟶ frequency
  - *week* ⟶ duration

Algorithm: Drugs, then the Rest

A two-step strategy

1. **Identification of drug names** based upon an exact match from drug lexicons;
2. **Identification of related information** based on regular expressions and lexicon look-up.

Inherently encodes dependency between drug and attached information

Algorithm: More Detail

General algorithm

- **Segment text into sentences:**
  - section titles: *MEDICATIONS ON ADMISSION*, *ALLERGIES*
  - typographical clues: full stops (not those in abbreviations or numbers) and section separation (line of stars)
- **Identify drug names**
- **Segment sentences into drug portions** (each drug name starts a new portion)

  $\longrightarrow$ hypothesis: related information often follows drug name.
- **Identify related information:**
  - search inside each drug portion for associated information
  - search extended to sequence closely preceding drug name

Algorithm: Initial Text

Original text

[line] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT
[line] 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary
[line] function test 8/27/97 , cardiac catheterization 0/20/97.
[line] DISCHARGE MEDICATIONS: Captopril 50 mg p.o. q.i.d.; Isordil 20 mg
[line] p.o. t.i.d.; Lasix 40 mg p.o. q. day with
[line] instructions that if his weight increased by three to four pounds ,
[line] he should take 80 mg of Lasix that day; Lotrimin 1% cream topical
[line] b.i.d.; and digoxin 0.25 mg p.o. q. day.
[line] DIET: He was also discharged on a 2 gram sodium diet with 2 liter
[line] fluid restriction.

Algorithm: Sentence Segmentation

Text segmentation in sentences

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS: Captopril 50 mg p.o. q.i.d.; Isordil 20 mg p.o. t.i.d.; Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of Lasix that day; Lotrimin 1% cream topical b.i.d.; and digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Drug name Identification

Drug name identification

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS: Captopril 50 mg p.o. q.i.d.; Isordil 20 mg p.o. t.i.d.; Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of Lasix that day; Lotrimin 1% cream topical b.i.d.; and digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Segmentation into Drug Portions

Segmentation of sentences into drug portions

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS:
[portion] Captopril 50 mg p.o. q.i.d.;
[portion] Isordil 20 mg p.o. t.i.d.;
[portion] Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of
[portion] Lasix that day;
[portion] Lotrimin 1% cream topical b.i.d.; and
[portion] digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Dosage

Related information extraction inside each portion

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS:
[portion] Captopril 50 mg p.o. q.i.d.;
[portion] Isordil 20 mg p.o. t.i.d.;
[portion] Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of
[portion] Lasix that day;
[portion] Lotrimin 1% cream topical b.i.d.; and
[portion] digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Mode of Administration

Related information extraction inside each portion

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS:
[portion] Captopril 50 mg p.o. q.i.d.;
[portion] Isordil 20 mg p.o. t.i.d.;
[portion] Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of
[portion] Lasix that day;
[portion] Lotrimin 1% cream topical b.i.d.; and
[portion] digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Frequency

Related information extraction inside each portion

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.
[sentence] DISCHARGE MEDICATIONS:
[portion] Captopril 50 mg p.o. q.i.d.;
[portion] Isordil 20 mg p.o. t.i.d.;
[portion] Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of
[portion] Lasix that day;
[portion] Lotrimin 1% cream topical b.i.d.; and
[portion] digoxin 0.25 mg p.o. q. day.
[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Algorithm: Look Left if Needed

Related information extraction from preceding portion

[sentence] PROCEDURES: He underwent an echocardiogram on 8/27/97 , ETT 8/27/97 , abdominal ultrasound 8/27/97 , pulmonary function test 8/27/97 , cardiac catheterization 0/20/97.

[sentence] DISCHARGE MEDICATIONS:

[portion] Captopril 50 mg p.o. q.i.d.;

[portion] Isordil 20 mg p.o. t.i.d.;

[portion] Lasix 40 mg p.o. q. day with instructions that if his weight increased by three to four pounds , he should take 80 mg of

[portion] Lasix that day;

[portion] Lotrimin 1% cream topical b.i.d.; and

[portion] digoxin 0.25 mg p.o. q. day.

[sentence] DIET: He was also discharged on a 2 gram sodium diet with 2 liter fluid restriction.

Expert-based: How to Transfer to Another Language?

Could that work for Italian patient reports?

How we did for French (project Akenaton)

- Domain: cardiology
- Prepare similar lexicons: go shopping again
- Adapt patterns (regular expressions)

Expert-based: How to Transfer to Another Language?

Could that work for Italian patient reports?

How we did for French (project Akenaton)

- Domain: cardiology
- Prepare similar lexicons: go shopping again
- Adapt patterns (regular expressions)

Medication Extraction: French Example

Mon Cher Confrère,

[. . .]

Actuellement, sous Flécaïne 1 cp matin et soir et Préviscan, le patient est totalement asymptomatique. D'autre part, l'hypertension artérielle semble bien équilibrée par l'Aprovel 300, 1 par jour.

. . .

Au total, comme Monsieur <Nom patient> est actuellement peu symptomatique, je continuerai le même traitement sous la forme de Flécaïne 1 cp matin et soir en plus de l'Aprovel 300, 1 par jour. Par contre, je diminuerai progressivement le Préviscan et je le remplacerai par Kardégic 160 mg/24 h chez ce patient présentant une insuffisance aortique très modérée et une minime insuffisance mitrale sur prolapsus de la grande valve.

[. . .]

Bien confraternellement.

Medication Extraction: French Example

Mon Cher Confrère,
[. . .]
Actuellement, sous Flécaïne 1 cp matin et soir et Préviscan, le patient est totalement asymptomatique. D'autre part, l'hypertension artérielle semble bien équilibrée par l'Aprovel 300, 1 par jour.
. . .
Au total, comme Monsieur <Nom patient> est actuellement peu symptomatique, je continuerai le même traitement sous la forme de Flécaïne 1 cp matin et soir en plus de l'Aprovel 300, 1 par jour. Par contre, je diminuerai progressivement le Préviscan et je le remplacerai par Kardégic 160 mg/24 h chez ce patient présentant une insuffisance aortique très modérée et une minime insuffisance mitrale sur prolapsus de la grande valve.
[. . .]
Bien confraternellement.

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- **Expert-based method: De-identification**
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

De-identification: An instance of entity recognition

Constraints on clinical documents

- Clinical documents can be used outside of the patient care course only after removal of any mark identifying the patient,
- Real need to anonymize these documents before distribution for research or for publication (case study),
- Anonymization is a strong constraint which creates a scarcity of available clinical corpora.

Anonymization rules

Importance of Protected Health Information

USA: 18 identifiers that should be anonymized in order to allow PHI documents distribution (HIPAA)

- first name, last name, age (over 90 y.o.);
- address, place name;
- phone & fax numbers, e-mails, URL, IP address;
- social security number, medical record number, health plan beneficiary number, account number, certificate/license numbers, vehicle identifiers;
- serial numbers and device identifiers;
- biometric identifiers, including voice and finger prints;
- any other unique identifying number, characteristic (tatoos, scars) or code.

De-identification systems

i2b2 anonymization challenge from clinical data (2007)

- machine-learning methods based on features obtained from NLP tools;
- F-measure over 0.98 for the best systems.

Example: Anonymizing French clinical reports

Medina (MEDical INformation Anonymization) (Grouin et al., MIE 2009)

Three-stage Process

1. Pre-anonymization (in hospital): use data from hospital information system
   - Scan text for patient's name and birth date
2. Main stage: apply lexicons and regular expressions
3. Post stage: study the neighbourhood of already anonymized words

anonymized first name> <capitalized word not in common dictionary> →
<first name> <last name>

Medina, main stage: Expert-based method

1. Linguistic resources
   - French dictionary: 251,306 inflected forms
   - Proper names: 23,079 first names, 12,994 last names, 247 countries, 30,748 towns, 3,296 drug names, 1,993 hospitals, 108 doctor names (from the training corpus), pacemaker tradenames
   - Black list: terms which must not be anonymized (disease names composed of first name and last name: *Emery-Dreifuss*)
2. Trigger words
   - For person names: *Docteur, Dr, DR, Madame, Melle*, etc.
   - For hospital names: *CHU, CHR, Clinique, Hôpital*, etc.
3. Regular expressions

Example anonymization

Original text (fake sentence)

J'ai examiné en consultation
Madame Dupont Michèle,
née le 13.1.1943,
âgée de 62 ans,
pour le contrôle annuel de
son stimulateur double chambre.

Example anonymization

First stage (in hospital)

J'ai examiné en consultation
Madame <marital_patient_name/> Michèle,
née le 13.1.1943,
âgée de 62 ans,
pour le contrôle annuel
de son stimulateur double chambre.

Example anonymization

Second and third stages

J'ai examiné en consultation
Madame <last_name/> <first_name/>,
née le <date/>,
âgée de <age/>,
pour le contrôle annuel
de son stimulateur double chambre.

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Architectures for Medical Entity Recognition

Adrenal-sparing surgery for hereditary phaeochromocytoma

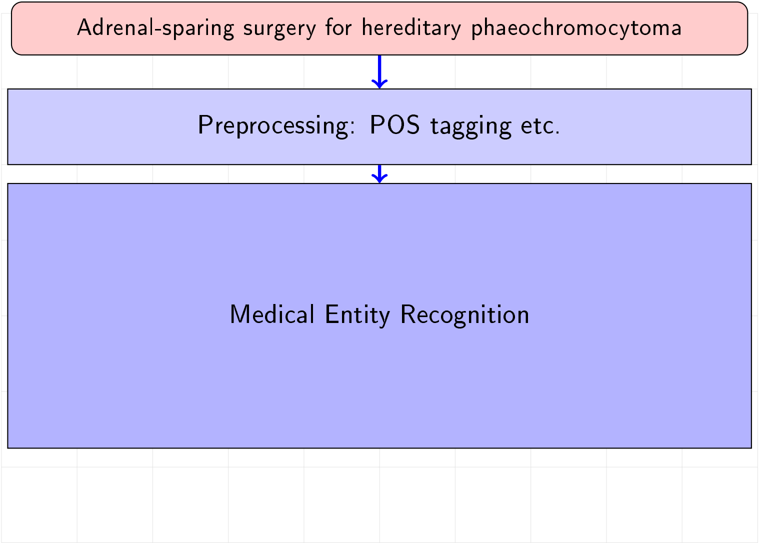
Architectures for Medical Entity Recognition

Adrenal-sparing surgery for hereditary phaeochromocytoma

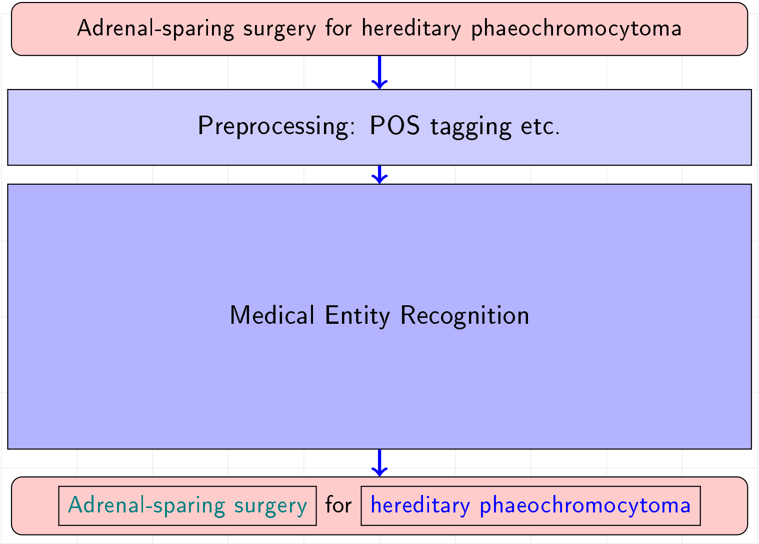
$\downarrow$

Preprocessing: POS tagging etc.

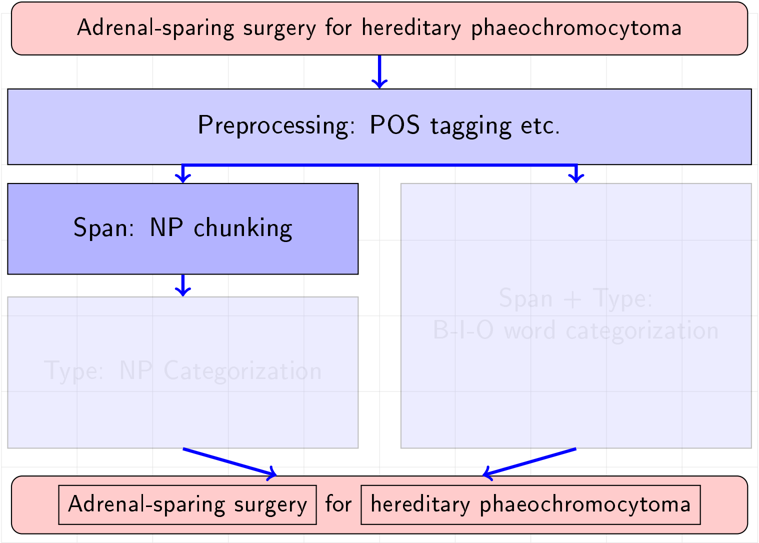
Architectures for Medical Entity Recognition



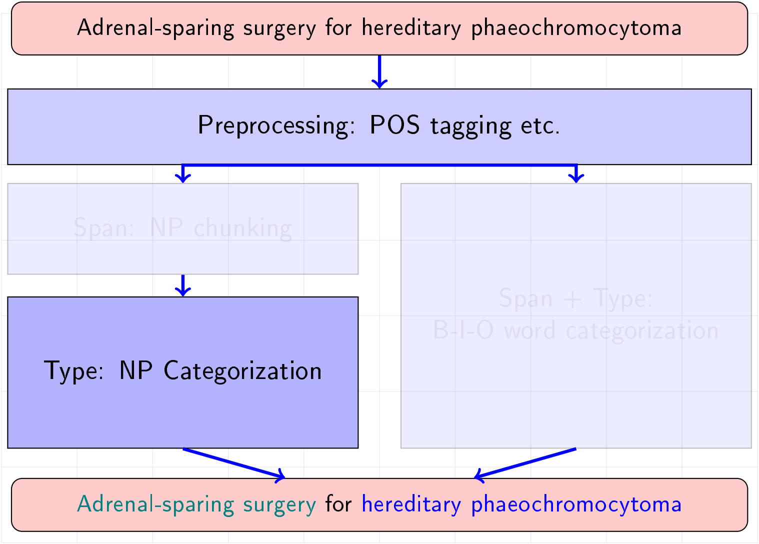
Architectures for Medical Entity Recognition



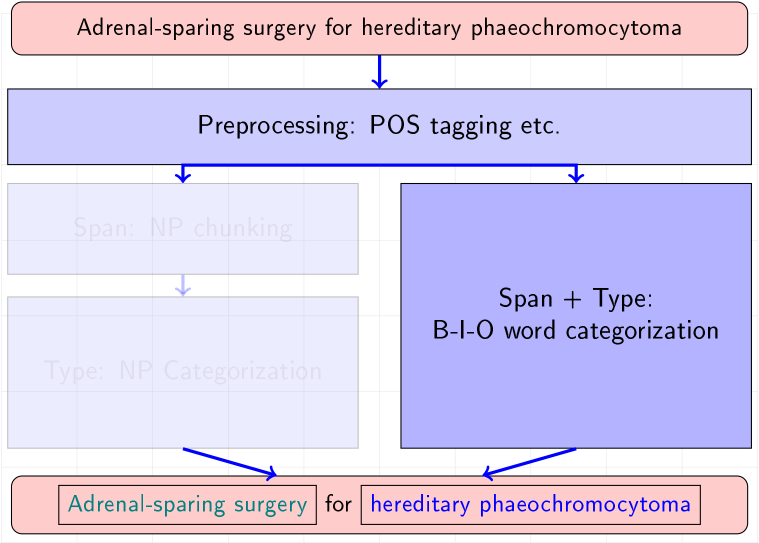
Architectures for Medical Entity Recognition



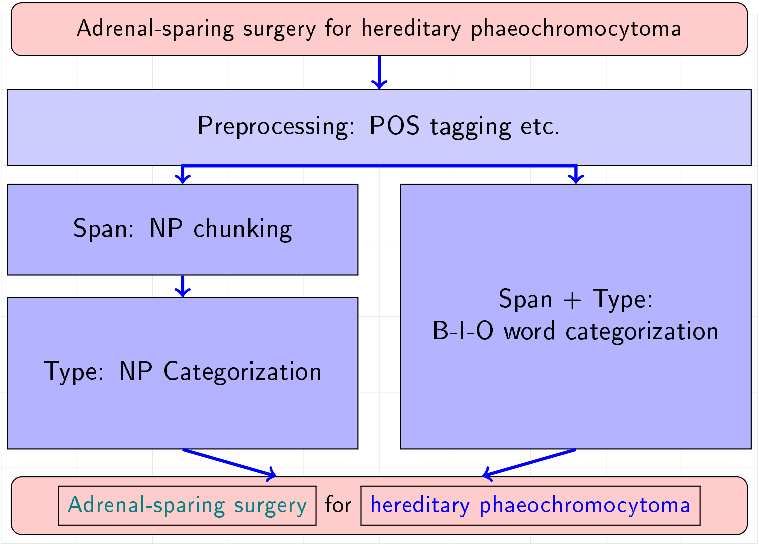
Architectures for Medical Entity Recognition



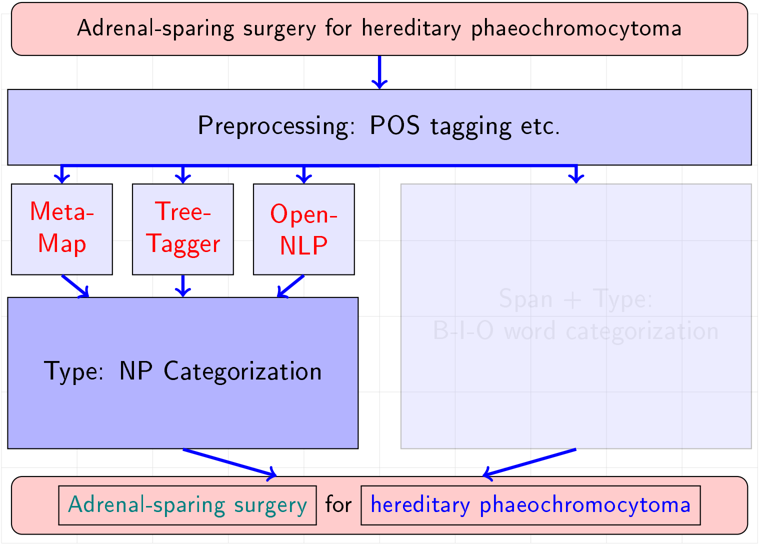
Architectures for Medical Entity Recognition



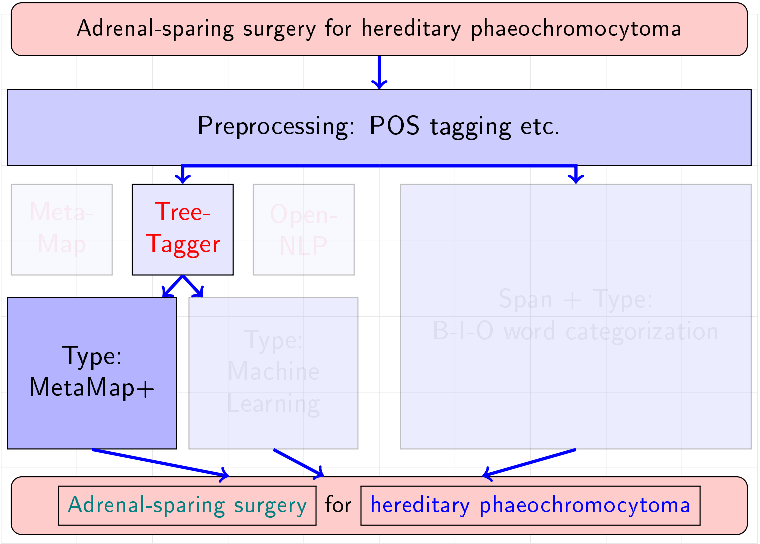
Architectures for Medical Entity Recognition



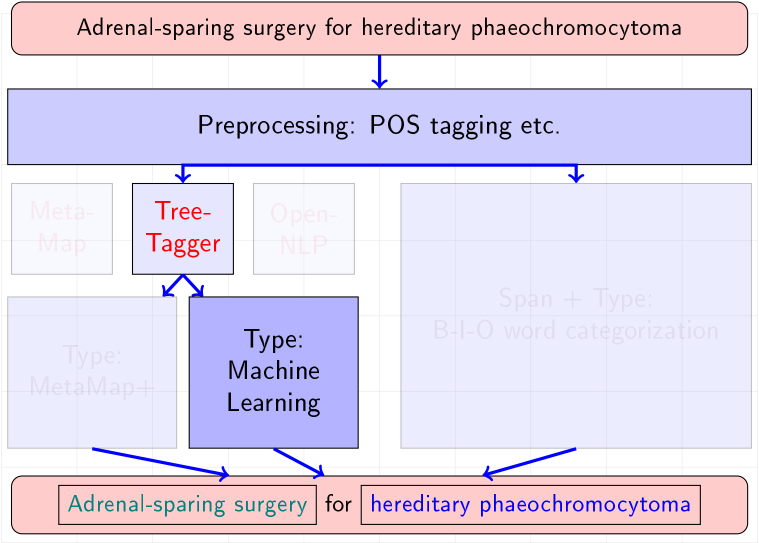
Architectures for Medical Entity Recognition



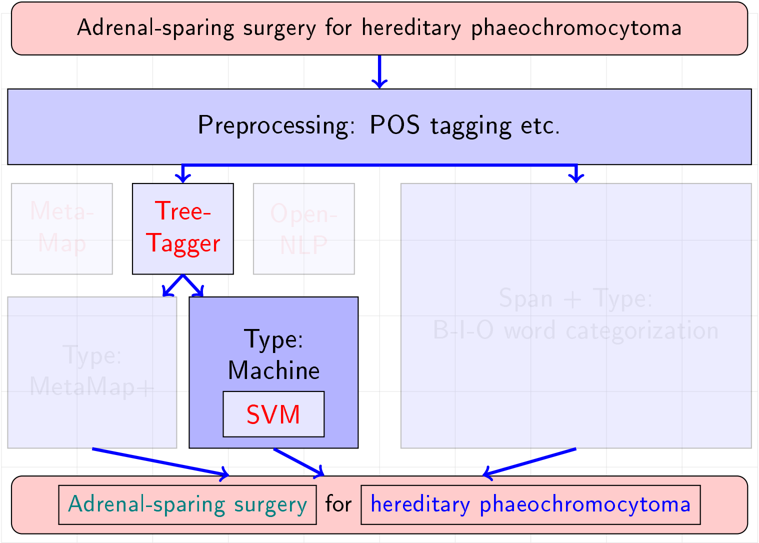
Architectures for Medical Entity Recognition



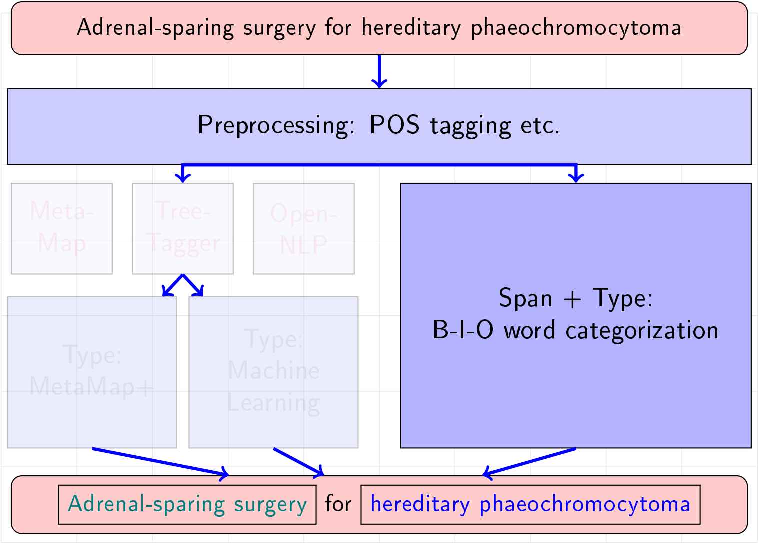
Architectures for Medical Entity Recognition



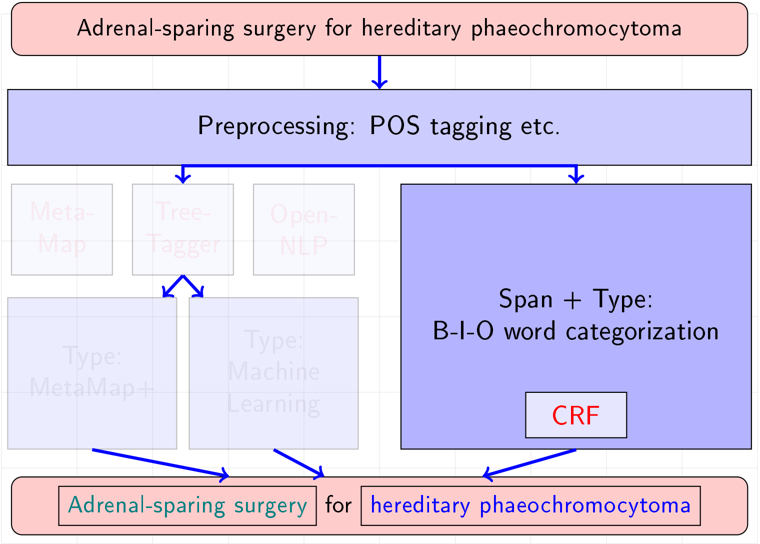
Architectures for Medical Entity Recognition



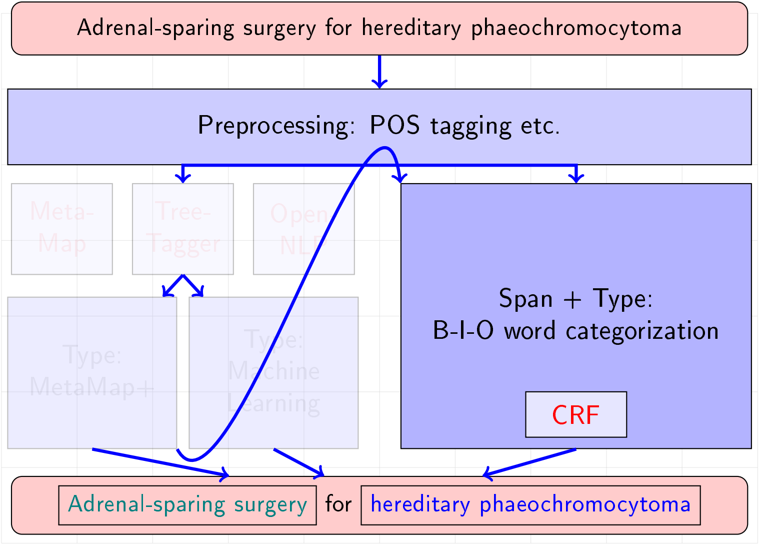
Architectures for Medical Entity Recognition



Architectures for Medical Entity Recognition



Architectures for Medical Entity Recognition



Named Entity Recognition as a Word Categorization Problem

B-I-O notation (Typically used with Conditional Random Fields classifier)

- Reformulation of the problem
  - *Where* are the entities? boundaries
  - *What* are they? type
- Position of word with respect to entity of given type:
  - *Beginning* or *Inside* entity
    (of type *Problem*, of type *TEst*, of type *TReatment*)
  - *Outside* any entity

		problem		problem				
history	of	hypercholesterolemia	and	type	II	diabetes	mellitus	.
O	*O*	*BP*	*O*	*BP*	*IP*	*IP*	*IP*	*O*

Decision among $2n + 1 = 7$ categories
BP, IP, BTE, ITE, BTR, ITR, O

Named Entity Recognition as a Word Categorization Problem

B-I-O notation (Typically used with Conditional Random Fields classifier)

- Reformulation of the problem
  - *Where* are the entities? boundaries
  - *What* are they? type
- Position of word with respect to entity of given type:
  - *Beginning* or *Inside* entity
    (of type *Problem*, of type *TEst*, of type *TReatment*)
  - *Outside* any entity

		problem		problem				
history	of	hypercholesterolemia	and	type	II	diabetes	mellitus	.
O	*O*	*BP*	*O*	*BP*	*IP*	*IP*	*IP*	*O*

Decision among $2n + 1 = 7$ categories
BP, *IP*, *BTE*, *ITE*, *BTR*, *ITR*, *O*

Example features to train a CRF

Token	Part of speech	Target class
But	CC	O
analysts	NNS	B-NP
reckon	VBP	B-VP
underlying	VBG	B-NP
support	NN	I-NP
for	IN	B-PP
sterling	NN	B-NP
has	VBZ	B-VP
been	VBN	I-VP
eroded	VBN	I-VP
by	IN	B-PP
the	DT	B-NP
chancellor	NN	I-NP
's	POS	B-NP
failure	NN	I-NP
...		

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Normalization, co-reference

Briefly , the patient has a history of chronic obstructive pulmonary disease , ethanol abuse , chronic pleural effusions , and chronic renal insufficiency .
He presented to Gaanvantsir on 04-17-92 with abdominal pain and bloody diarrhea .
Workup revealed ischemic bowel secondary to Celiac and SMA stenoses .
The patient underwent an angioplasty of his SMA from 90-20% residual .
The patient was also found to have gram negative rod sepsis with blood cultures times two growing E. coli and B. fragilis

Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Detection of medical relations

What treatment was given to cure the problem? What signs did the tests reveal?

A patient record (excerpt)

He has a recent history of dyspnea on exertion on exertional chest pain which has increased over the last several weeks and is relieved by sublingual nitroglycerin .

On 2016-06-26 , he had a positive exercise tolerance test .

Cardiac catheterization at the end of October revealed a dilated aortic root to 4.4 cm and 80% stenosis of the mid left anterior descending at the bifurcation involving the diagonal branch , 70% stenoses of the left circumflex and oblique marginal artery and 90% stenosis of the posterior descending artery .

His atrioventricular valve gradient was 27 with an AV surface area of .91 .

REVIEW OF SYSTEMS :

Review of systems shows that the patient denies any orthopnea , paroxysmal nocturnal dyspnea .

See HTML version

Some Relations Between Medical Concepts

i2b2/VA 2010 Challenge

Eight types of *relations* between concepts

- problem-treatment
  - treatment improves (*TrIP*), worsens (*TrWP*), causes (*TrCP*) the problem
  - treatment is administered (*TrAP*) or not (*TrNAP*) for the problem
- problem-test
  - test reveals (*TeRP*) or allows a physician to investigate (*TeCP*) the problem
- problem-problem
  - problem indicates another problem (*PIP*)

Relation Identification as a Classification Task

- Given two concepts, decide:
  - whether or not there is a relation between them
  - and if there is a relation, determine which one
- 8 relation types:
  - TrIP, TrWP, TrCP, TrAP, TrNAP
  - TeRP, TeCP
  - PIP

Relation Identification as a Classification Task

Number of possible categories depends on concept types

- problem-treatment: 4 relations + no relation = 5
  - treatment improves (*TrIP*), worsens (*TrWP*), causes (*TrCP*) the problem
  - treatment is administered (*TrAP*) or not (*TrNAP*) for the problem
- problem-test: 2 relations + no relation = 3
  - test reveals (*TeRP*) or allows a physician to investigate (*TeCP*) the problem
- problem-problem: 1 relation + no relation = 2
  - problem indicates another problem (*PIP*)
- test-treatment: 0

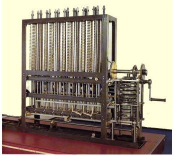
A Hybrid Method for Relation Detection

Supervised classification requires enough training examples

- This was not the case for four relations

1. A hybrid approach: combines machine-learning techniques and linguisitic-pattern matching.
   - Trained an SVM (libsvm tool)
   - Built linguistic patterns manually
2. A supervised learning approach with more linguistic preprocessing

Hybrid, Supervised, and Their Combination



- System R1: Hybrid system:
  1. use hand-designed patterns
     - to identify 4 relations
     - TrIP, TrWP, TrNAP, TeCP
     - few examples in training set
  2. predict the other relation types by supervised classification (SVM)
- System R2: supervised classification from simplified texts.
- System R3: combination of results of systems R1 and R2.

R1: Normalization of Texts

- **Texts are preprocessed and normalized**
  - replace abbreviations with their meanings:
    - *h.o.* $\longrightarrow$ *history of*
    - *p.r.n.* $\longrightarrow$ *as needed*
  - substitute the person's name (**NAME[VVV]), the date (**DATE[Jan 06 2008]), the person's age and other numbers respectively with <NAME>, <DATE>, <AGE> and <NUM>.
  - apply part-of-speech tagger (TreeTagger)

R1: Manually-Designed Relation Patterns

First-level classification, priority over supervised classification

- **Patterns:** Design and tune patterns on the training corpus.
  - Here, kept only the patterns of four relations types as the others did not yield satisfying results:

	Example	Precision	Recall
TrIP	_PB_ (.*) headed by _TX_	0.35	0.45
TrWP	(despite)? _TX_ (.*) no relief of _PB_	0.16	0.79
TrNAP	_TX_ (.*) avoided for _PB_	0.16	0.65
TeCP	_PB_ (.*) _TE_ (.*) recommended	0.08	0.60

R2: Sentence Simplification

Modify sentences before supervised classification

- **Concept substitution:** concepts are substituted with their types (problem, test or treatment), and each sentence is duplicated for each candidate relation.
- **Syntactic analysis** by the Charniak/McClosky self-training parser.
- **Syntactic simplification:** deletion of some syntactic phrases between the candidate concepts.
  - If the concept is at the beginning of the noun phrase, all words after the concept in the noun phrase are deleted.
  - If there is a PP, an ADJP, a CONJP, a WHNP or a CC (followed by a noun phrase) between the concepts, it is replaced with its POS tag (<PP>, <ADJP>, etc.).

R1–R2: Supervised Classification of Relations

Lexical features

Stemming, Part-of-speech, Verb classes

- tokens and stemmed tokens in candidate concepts,
- left and right trigrams (of stemmed tokens) of the two concepts,
- stemmed tokens between them,
- verbs in 3-word window before and after each concept and between them,
- Levin's class of the verbs (coming from VerbNet),
- preposition between concepts,
- headword of concepts (headword is the token after preposition, else it is the last token).

R1–R2: Supervised Classification of Relations

Syntactic features

Part-of-speech

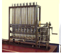
- part-of-speech in a 3-word window to the left and the right of the candidate concepts,
- presence of a preposition,
- presence of a coordination conjunction between concepts.
- punctuation sign.

R1–R2: Supervised Classification of Relations

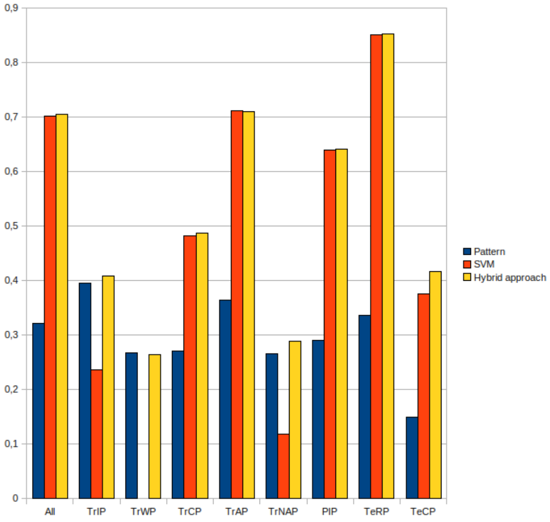
Concept-related features

Knowledge of detected concepts

- order of the candidate concepts
- distance between them (i.e. number of tokens)
- presence of other concepts
- semantic type (from the UMLS) of tokens in a 3-word window to the left and the right of each candidate concept
- type of the concepts (problem, test or treatment)
- normalized title of the section




Hybrid Approach: Contribution of Methods



Texts in Biomedicine

Introduction to NLP

- Morphology: from characters to words
- Syntax: part-of-speech tagging, sentence parsing
- Semantics: entities, semantic roles and relations

Types of Methods

- Knowledge-based methods
- Machine-learning-based methods
- Hybrid methods
- Dependence on language-specific resources

Tasks and methods in biomedical NLP

- Expert-based method: Extraction of prescription information
- Expert-based method: De-identification
- Data-driven methods for medical entity recognition
- Normalization, co-reference
- Detection of medical relations: binary relations

Conclusion

- State of the art can be observed at i2b2 challenges (*e.g.*, de Bruijn *et al.*, JAMIA 2011)

Task	F-measure
Problems, Tests, Treatments	0.85
Factuality of problems	0.94
Relations	0.73

  - Working with millions of features
  - Use of rich word features and of syntactic dependency structures
  - Ensemble classification
  - Self-training