Towards clinical language understanding

Simon Šuster, Stéphan Tulkens and Walter Daelemans

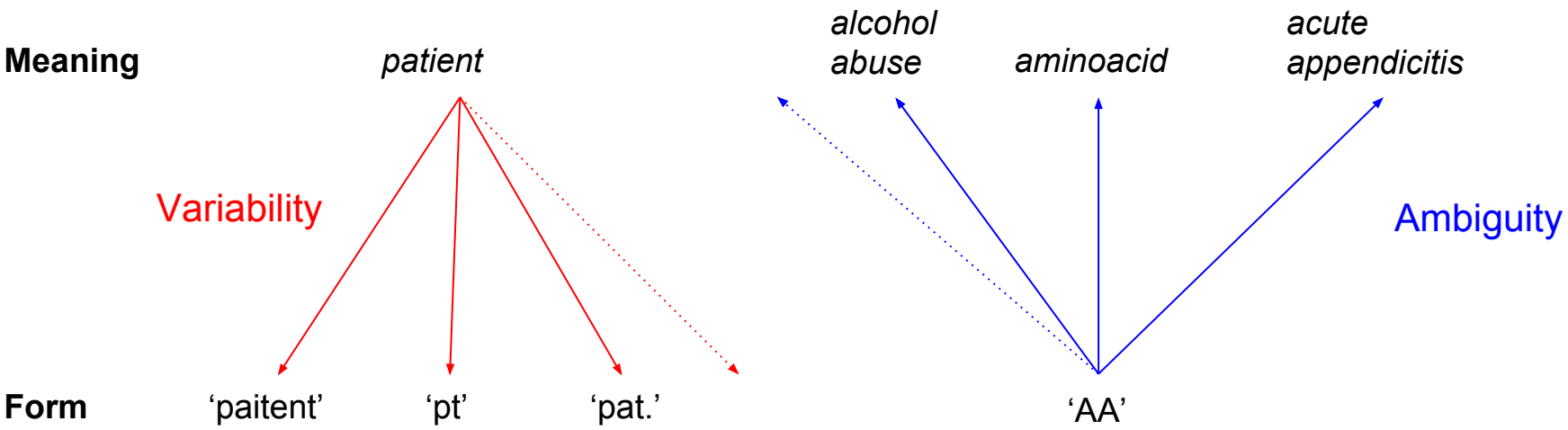
ATILA, October 2016

Part I:
Clinical NLP

Goals of clinical NLP

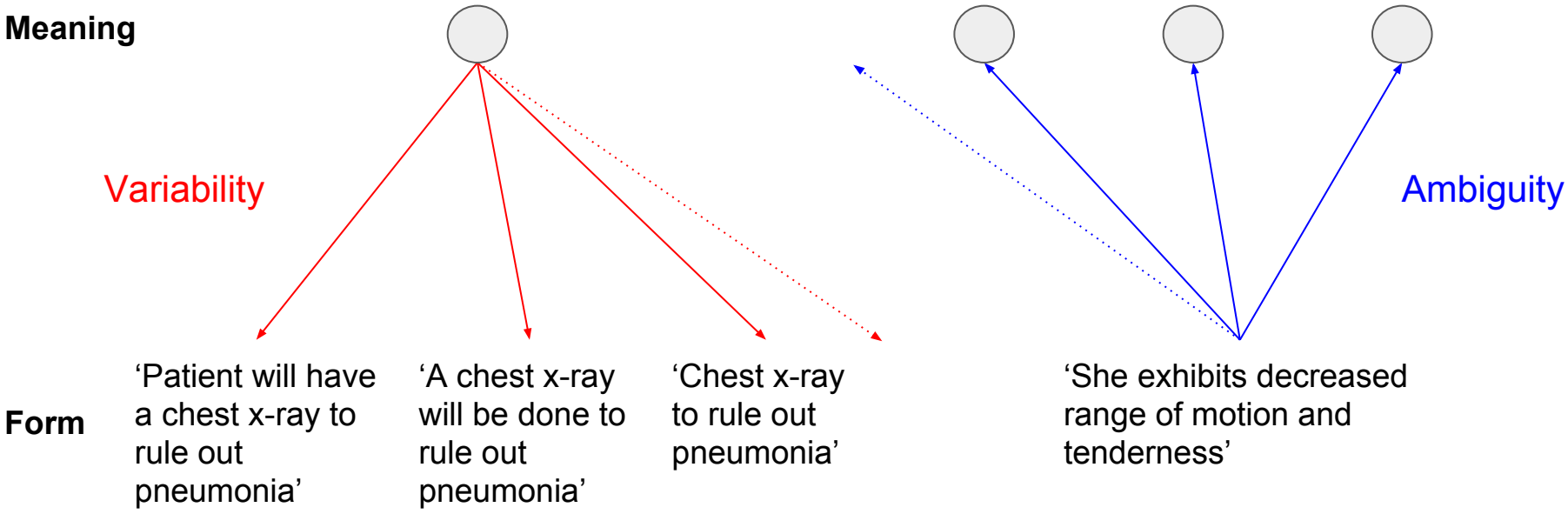
- Assisting in disease diagnosis (clinical decision support)
- Finding new linkages between symptoms, drugs, diseases and patient attributes
- Recruiting patients for clinical trials
- Personalized medicine
- Insights into population health

Word-level dichotomy

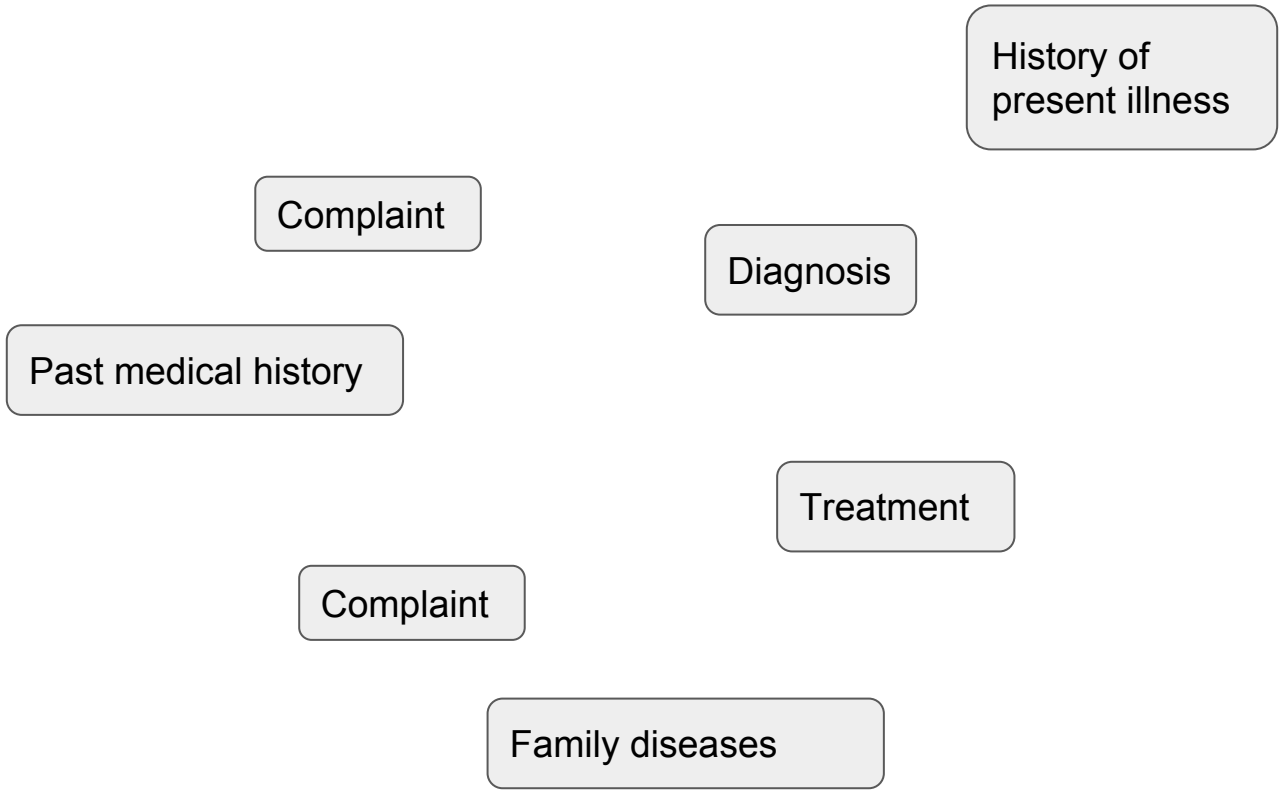


Adapted from http://l2r.cs.illinois.edu/%7Edanr/Talks/Roth-Making-Sense-02-2013.pptx

Sentence-level dichotomy



Discourse level:
Can we reliably identify the internal structure?



The ecology of clinical NLP research

- Models should be highly accurate to be useful
  - But accuracy requires lots of annotated data
- Little annotated data, mostly English
- Difficult access to unannotated data. Why?
  - Risk of disclosing personal information
  - Risk of disclosing hospital practices
  - Clinicians may lack trust
  - Clinicians may fear losing their unique role
- Divide between biomedical and NLP communities

Part II: Accumulate

Accumulate project (2016–2019, SBO-IWT)

Develop technology for analysis of free-text clinical reports in English and Dutch

The role of CLiPS

- **Word-level (terminology extraction)**
  - Developing techniques for normalizing the reports — variability
  - Recognizing and disambiguating concepts — ambiguity+variability
- **Sentence-level (event structure)**
  - Predicate-argument semantics / relation extraction — ambiguity+variability
  - Negation, modality, quantification

Accumulate project (2016–2019, SBO-IWT)

Develop technology for analysis of free-text clinical reports in English and Dutch

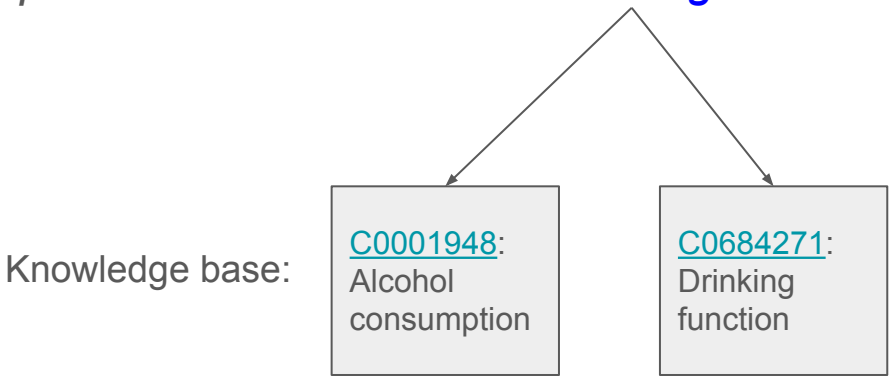
The role of CLiPS

- **Word-level (terminology extraction)**
  - Developing techniques for normalizing the reports — variability
  - Recognizing and disambiguating concepts — ambiguity+variability
- **Sentence-level (event structure)**
  - Predicate-argument semantics / relation extraction — ambiguity+variability
  - Negation, modality, quantification

Concept disambiguation (Tulkens et al. ’16, BioNLP)

Example: drinking

“366 class 1 and 2 pupils completed a questionnaire about their drinking habits”



Idea:

- Choose the sense whose KB definition is the most similar to the word’s current neighborhood
- Similarly to the Simplified Lesk algorithm for WSD

Unified Medical Language System ®

UMLS Terminology Services

Metathesaurus Browser

Welcome back, simchy

Search | Tree | Recent Searches

Term ○ CUI ○ Code

Drinking [Go]

Release: 2015AB

Search Type: Word

Source: All Sources
AIR
ALT
AOD
AOT

Search Results (553)
[: 1 - 25 :]

C0001948 Alcohol consumption
C0684271 Drinking function
C0001962 Ethanol
C0001967 Alcoholic Beverages
C0013124 Drinking behavior processes
C0085762 Alcohol abuse
C0349097 Mental and behavioral disorders due to use
C0425332 Drinks wine

Basic View | Report View | Raw View

Concept: [C0684271] Drinking function

Semantic Types

Organism Function [T040]

Definitions

ICF | Taking hold of a drink, bringing it to the mouth, and consuming the drink in culturally acceptable ways, mixing, stirring and pouring liquids for drinking, opening bottles and cans, drinking through a straw or drinking running water such as from a tap or a spring; feeding from the breast.

ICF-CY | Indicating need for, and taking hold of a drink, bringing it to the mouth and consuming the drink in culturally acceptable ways; mixing, stirring and pouring liquids for drinking, opening bottles and cans, drinking through a straw or drinking running water, such as from a tap or a spring; feeding from the breast.

MSH | The consumption of liquids.

MSHCZE | Spotřeba tekutin.

Atoms (46) string [AUI / RSAB / TTY / Code]

drinking [A18641616/CHV/PT/0000043974]

drinking [A14256958/GO/ET/GO:0007631]

Procedure

1. Train biomedical embeddings
2. Based on the embeddings and the UMLS thesaurus, represent each concept *s* with a vector $\mathbf{v}_s$:

> $\mathbf{v}_s$: is the average of definition vectors $\mathbf{d}_s$
> $\mathbf{d}_s$: is the sum over vectors of all words in the definition

3. For every occurrence of an ambiguous word *w* in a document, sum the vectors of context words
4. Average these summed vectors into $\mathbf{x}_w$
5. Choose the highest-scoring concept: argmax_s cosine($\mathbf{v}_s$, $\mathbf{x}_w$)

Evaluation

MSH-WSD dataset

- ~200 ambiguous terms (each with 2–5 concepts)
- ~38k Medline® abstracts

Accuracy of our method:

- 0.84 (only attested concepts)
- 0.75 (all UMLS concepts for a term)

Other findings

Results vary depending on:

- source of training text for word embeddings
- chosen term

Disambiguation difficult when the definitions for concepts are similar

We outperform methods that (like us) don’t use relational KB information

This talk

Part I:

- Variability and ambiguity in clinical NLP
- Challenges of the clinical domain

Part II:

- Accumulate project
- Concept disambiguation with a Lesk-like algorithm and word embeddings
- http://github.com/clips/yarn
- Stéphan Tulkens, Simon Šuster and Walter Daelemans. *Using Distributed Representations to Disambiguate Biomedical and Clinical Concepts.* BioNLP'16