

# The challenges in concept detection for clinical texts

Simon Šuster

CLiPS, University of Antwerp  
Antwerp University Hospital

ACCUMULATE meeting, 12/2016

## WHAT DO WE MEAN BY CLINICAL CONCEPT DETECTION

Finding clinical terms in free-text reports, specifying their spans and types

*Patient* described *pain* as being burning in nature , occasionally *colicky* but he never had *constipation*, *obstipation* or *abdominal distension*.

*constipation* denotes a clinically relevant entity

# WHAT DO WE MEAN BY CLINICAL CONCEPT DETECTION

We can distinguish between two subproblems:

- Named entity recognition (NER)

- Concept disambiguation

# WHAT DO WE MEAN BY CLINICAL CONCEPT DETECTION

We can distinguish between two subproblems:

- Named entity recognition (NER)
  - find term location and corresponding semantic type
  - sequence labeling problem, with number of categories being small
  - labels are e.g. problem, treatment, test
  - **constipation** → e.g. problem
- Concept disambiguation

# WHAT DO WE MEAN BY CLINICAL CONCEPT DETECTION

We can distinguish between two subproblems:

- Named entity recognition (NER)
  - find term location and corresponding semantic type
  - sequence labeling problem, with number of categories being small
  - labels are e.g. problem, treatment, test
  - **constipation** → e.g. problem
- Concept disambiguation
  - choose the correct sense (link to an ontology)
  - number of "labels" is much larger compared to NER
  - labels can be all concept identifiers in an ontology
  - **constipation** → 14760008
  - knowing the id, finding the semantic type is straightforward

## WHY DETECT CONCEPTS

- Extracted concepts give a condensed summary of a report
- Extracted concepts are useful for downstream tasks we'd like to perform (relation extraction and temporal information extraction)
- They are ultimately important for applications like
  - diagnosis explanation
  - modeling of disease progression
  - analysis of treatment effectiveness
  - simplification of reports for patient use

## WHAT PRECISELY COUNTS AS A CONCEPT

Suppose you want to annotate a text corpus with concepts, what will be the operational definition of a concept?

# WHAT PRECISELY COUNTS AS A CONCEPT

Suppose you want to annotate a text corpus with concepts, what will be the operational definition of a concept?

## Inclusiveness:

- ❖ What is best?
  - ❖ early **intrauterine insult** of inferior mesenteric artery
  - ❖ **early intrauterine insult** of inferior mesenteric artery
  - ❖ early **intrauterine insult** of **inferior mesenteric artery**
  - ❖ **early intrauterine insult of inferior mesenteric artery**
- ❖ How far from the node do we go (left/right modification, determiners)?
- ❖ Concepts need to be intuitive and learnable



# HOW PRECISE SHOULD THE CATEGORIES BE (NER)

## Granularity:

- i2b2-2010 has only 3 labels: **treatment, test, problem**
- UMLS has 133 semantic types:  
e.g. **mental process, vitamin, sign or symptom, food**
- UMLS has 15 semantic groups:  
**activities & behaviors, anatomy, chemicals & drugs, concepts & ideas, devices, disorders, genes & molecular sequences, geographic areas, living beings, objects, occupations, organizations, phenomena, physiology, procedures**

## HOW DO WE LINK TO AN ONTOLOGY

At disambiguation time, we need to address inclusiveness (as in NER):

- diabetic                      monitoring  
Diab. Mellitus (disorder)    Monitoring-action  
73211009                      360152008
- diabetic monitoring  
Diabetic monitoring (regime/therapy)  
170742000

## AMBIGUITY

Many terms are ambiguous (esp. when looked at in isolation):

*The nasopharynx appeared normal with patent ET on both sides.*

## AMBIGUITY

Many terms are ambiguous (esp. when looked at in isolation):

*The nasopharynx appeared normal with patent ET on both sides.*

ET is a highly ambiguous entity in an ontology (SNOMED-CT):

- Endotracheal Tube (instrument), 26412008
- Embryo Transfer (procedure), 75456002
- Exercise Tolerance (obs. entity), 248243004
- Eustachian Tube (body structure), 91207004
- ...

# VARIABILITY

Eustachian canal structure (body structure) → 91207004

- Several terms could be used to mean the same thing
- Eustachian tube, auditory tube and pharyngotympanic tube are listed as synonyms in SNOMED-CT
- We can be reasonably confident to resolve those with simple look-up

# VARIABILITY

What about these variants:

- **tube**: underspecified, only obvious from the context
- **eustachian tube**: change in case
- **ET**: acronym
- **eust tube**: abbreviation
- **tuba auditiva**: Latin
- **eusthacian tube**: misspelling

## IMPLEMENTED OR IN DEVELOPMENT

- Lesk-like concept disambiguation [Tulkens et al., 2016]
  - Using word embeddings and UMLS definitions
- NER without labeled data
  - Using word embeddings, a chunker to detect noun phrases and UMLS
  - Mapping between UMLS semantic groups and the i2b2-defined categories
- Deep-neural NER using i2b2-2010
  - Fast and accurate, but how good does it generalize?
- Unsupervised spell-checking
  - Using a canonical-form lexicon and word embeddings with character n-gram information

## IMPLEMENTED OR IN DEVELOPMENT

- Lesk-like concept disambiguation [Tulkens et al., 2016]
  - Using word embeddings and UMLS definitions
- NER without labeled data
  - Using word embeddings, a chunker to detect noun phrases and UMLS
  - Mapping between UMLS semantic groups and the i2b2-defined categories
- Deep-neural NER using i2b2-2010
  - Fast and accurate, but how good does it generalize?
- Unsupervised spell-checking
  - Using a canonical-form lexicon and word embeddings with character n-gram information

**Currently, these work for English only!**



## WHAT ABOUT DUTCH

In SNOMED-CT, Dutch translations are sometimes available

- ❖ 22k validated translations, 37k non-validated translations, 6k validated synonyms\*
- ❖ For comparison, English version contains >300k concepts and 1M synonyms

---

\* Source: <http://zorgict.be/congres/wp-content/uploads/2016/06/Sessie-N1-Arabella-DHave-SNOMED-CT-vinger-aan-de-pols.pdf>

## WHAT ABOUT DUTCH

In SNOMED-CT, Dutch translations are sometimes available

- 22k validated translations, 37k non-validated translations, 6k validated synonyms\*
- For comparison, English version contains >300k concepts and 1M synonyms
- For **Eustachian tube**, no translation yet
  - **buis van Eustachius**
  - **eustachiusbuis**: premodification
  - **oortrompet**: synonym
  - **tuba Eustachii**: Latin origin

How can we map **buis van Eustachius** (or its alternatives) in text to 91207004?

---

\* Source: <http://zorgict.be/congres/wp-content/uploads/2016/06/Sessie-N1-Arabella-DHave-SNOMED-CT-vinger-aan-de-pols.pdf>

# WHAT ARE THE OPTIONS

## 1 Machine translation

- free Web-based translation can work fine [Schulz et al., 2013]

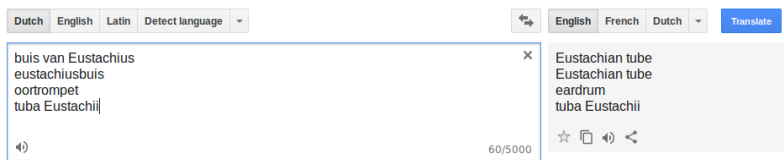
The screenshot shows a web-based translation interface. At the top, there are language selection buttons: 'Dutch', 'English', 'Latin', and 'Detect language'. On the right, there are buttons for 'English', 'French', and 'Dutch', along with a 'Translate' button. The input text on the left is 'buis van Eustachius', 'eustachiusbuis', 'oortrompet', and 'tuba Eustachii'. The output text on the right is 'Eustachian tube', 'Eustachian tube', 'eardrum', and 'tuba Eustachii'. There are also icons for a star, a document, a speaker, and a share icon at the bottom right of the output area.

Source (Dutch)	Target (English)
buis van Eustachius	Eustachian tube
eustachiusbuis	Eustachian tube
oortrompet	eardrum
tuba Eustachii	tuba Eustachii

# WHAT ARE THE OPTIONS

## 1 Machine translation

- free Web-based translation can work fine [Schulz et al., 2013]



The screenshot shows a web-based translation interface. At the top, there are language selection buttons: 'Dutch', 'English', 'Latin', and 'Detect language'. On the right, there are buttons for 'English', 'French', and 'Dutch', along with a 'Translate' button. The input field on the left contains the Dutch text: 'buis van Eustachius', 'eustachiusbuis', 'oortrompet', and 'tuba Eustachii'. The output field on the right shows the English translations: 'Eustachian tube', 'Eustachian tube', 'eardrum', and 'tuba Eustachii'. There are also icons for a star, a document, a speaker, and a back arrow, and a character count '60/5000'.

## 2 Direct lookup: diabetes → diabetes

- Dutch form same as English, or when Latin or English term is used

## 3 String similarity: syndroom → syndrome

# WHAT ARE THE OPTIONS

## 5 Joint-space embeddings (and model transfer)

- Induce word representations over both English and Dutch corpora
- All English and Dutch terms denoting the same concept have a very similar vectorial representation
- Challenging setup and data requirements (alignment or dictionaries?) [Gouws et al., 2014, Hermann and Blunsom, 2014]

# WHAT ARE THE OPTIONS

## 5 Joint-space embeddings (and model transfer)

- Induce word representations over both English and Dutch corpora
- All English and Dutch terms denoting the same concept have a very similar vectorial representation
- Challenging setup and data requirements (alignment or dictionaries?) [Gouws et al., 2014, Hermann and Blunsom, 2014]

## 6 External resources, especially Dutch Wikipedia

- textual descriptions available
- information about the medical domain through "categories"
- infoboxes with medical codes
- interlanguage links

# "OPGEBLAZEN GEVOEL"

WIKIPEDIA  
De vrije encyclopedie

Hoofdpagina  
Vind een artikel  
Vandaag  
Etalage  
Categorieën  
Recente wijzigingen  
Nieuwe artikelen  
Willekeurige pagina

Informatie  
Gebruikersportaal  
Snelcursus  
Hulp en contact  
Donaties

Hulpmiddelen  
Links naar deze pagina  
Verwante wijzigingen  
Bestand uploaden  
Speciale pagina's  
Permanente koppeling  
Paginagegevens  
Wikidata-item  
Deze pagina citeren

## Dyspepsie

(Doorverwezen vanaf Opgeblazen gevoel)



Neem het **voorbehoud bij medische informatie** in acht.  
Raadpleeg bij gezondheidsklachten een arts.

**Dyspepsie** is in de **geneeskunde** de term waarmee een verstoring in het spijsverteringsstelsel wordt bedoeld, waarbij voornamelijk maagklachten optreden.<sup>[1]</sup> Veelal wordt er gesproken over een opgeblazen gevoel in de maagstreek, welke de vertering van voeding onaangenaam maakt.

### Inhoud [verbergen]

- 1 Symptomen
- 2 Oorzaken
  - 2.1 Directe oorzaken
  - 2.2 Indirecte oorzaken
- 3 Specifieke Oorzaken
  - 3.1 Fructose
- 4 Behandeling
- 5 Noot

### Symptomen [ bewerken ]

Ingeval van een opgeblazen gevoel hoeft men niet alleen gas te hebben in de buik. De meest voorkomende andere symptomen zijn:

- Winderigheid
- Gerommel in de buik

### Dyspepsie

#### Coderingen

<b>ICD-10</b>	K30 <span>↗</span>
<b>ICD-9</b>	536.8 <span>↗</span>
<b>DiseasesDB</b>	30831 <span>↗</span>
<b>MeSH</b>	C23.888.821.236 <span>↗</span>

**Portaal** ↗ **Geneeskunde**

Thank you!



## Categorie:Maagaandoening



Pathologie } → Aandoening  
Geneeskunde }  
Inwendige geneeskunde → Gastro-enterologie } → Aandoening van het spijsverteringsstelsel → **Maagaandoening**  
Anatomie → Maag-darmstelsel



De categorie **Maagaandoening** biedt een overzicht van artikelen over aandoeningen van de **maag**.  
De **maag** is een **orgaan** dat behoort tot het spijsverteringsstelsel.

Hulpmiddelen: **Alle categorieën** - Toon bovenliggende categorieboom (png/svg) - Toon onderliggende categorieboom (png/svg) - Zoek artikelen met CatScan

### Artikelen in de categorie "Maagaandoening"

Deze categorie bevat de volgende 12 pagina's, van in totaal 12.

#### D

- [Dyspepsie](#)

#### F

- [Functionele maagklachten](#)

#### G

- [Gastritis](#)

#### H

- [Hematemesis](#)
- [Hernia diaphragmatica](#)

#### M

- [Maagbloeding](#)
- [Maagkanker](#)
- [Maagperforatie](#)

- [Maagslijmvliesirritatie](#)
- [Maagzweer](#)





#### P

- [Pylorushypertrofie](#)

#### Z

- [Syndroom van Zollinger-Ellison](#)

Categorie: Aandoening van het spijsverteringsstelsel

-  Gouws, S., Bengio, Y., and Corrado, G. (2014).  
BilBOWA: Fast Bilingual Distributed Representations without  
Word Alignments.  
*arXiv preprint arXiv:1410.2455.*
-  Hermann, K. M. and Blunsom, P. (2014).  
Multilingual Distributed Representations without Word  
Alignment.  
In *ICLR*.
-  Schulz, S., Bernhardt-Melischnig, J., Kreuzthaler, M., Daumke,  
P., and Boeker, M. (2013).  
Machine vs. human translation of SNOMED CT terms.  
In *MEDINFO World Congress on Medical and Health  
Informatics*.
-  Tulkens, S., Suster, S., and Daelemans, W. (2016).  
Using distributed representations to disambiguate  
biomedical and clinical concepts.  
In *BioNLP*.