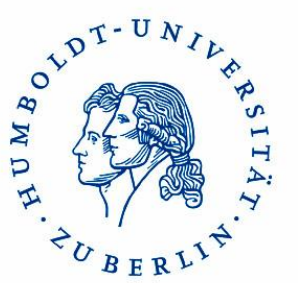
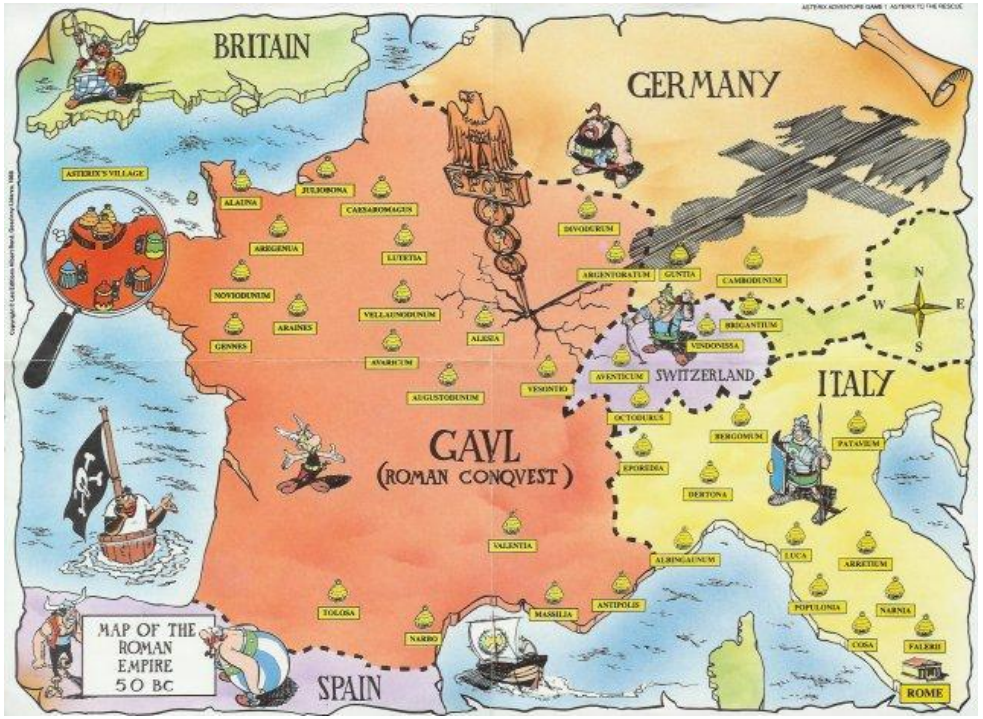
Cross-lingual Candidate Search for Biomedical Concept Normalization

Roland Roller, Madeleine Kittner, Dirk Weissenborn, Ulf Leser



Historic Background

https://www.ancient.eu/image/3372/



https://progressivegeographies.com/2015/07/12/jeremy-crampton-on-maps-permissions-and-asterix/

Background

- Nowadays: Greek and Latin rooted medical words can be found across many languages

English	German	Spanish	French	Swedish	Russian
carcinoma	Karzinom	carcinoma	carcinome	Karcinom	KARTSINOMA
Neurasthenia	Neurasthenie	neurastenia	Neurasthnie	Neurasteni	NEVRASTENIIA
Dioxins	Dioxine	Dioxinas	Dioxines	Dioxiner	DIOKSINY
Leukoplakia	Leukoplakie	Leucoplaquia	Leucoplasie	Leukoplaki	LEUKOPLAKIJA

NLP Pipeline

Pat. hat viel Durst. Appetit gut. Stuhlgang normal.

NLP Pipeline

Pat. hat viel Durst. Appetit gut. Stuhlgang normal.

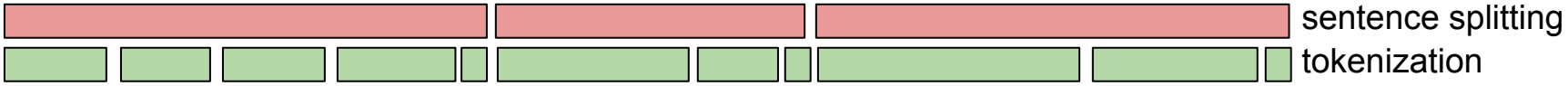
Patient is very thirsty. Good appetit. Bowel movement normal.

NLP Pipeline

Pat. hat viel Durst. Appetit gut. Stuhlgang normal.

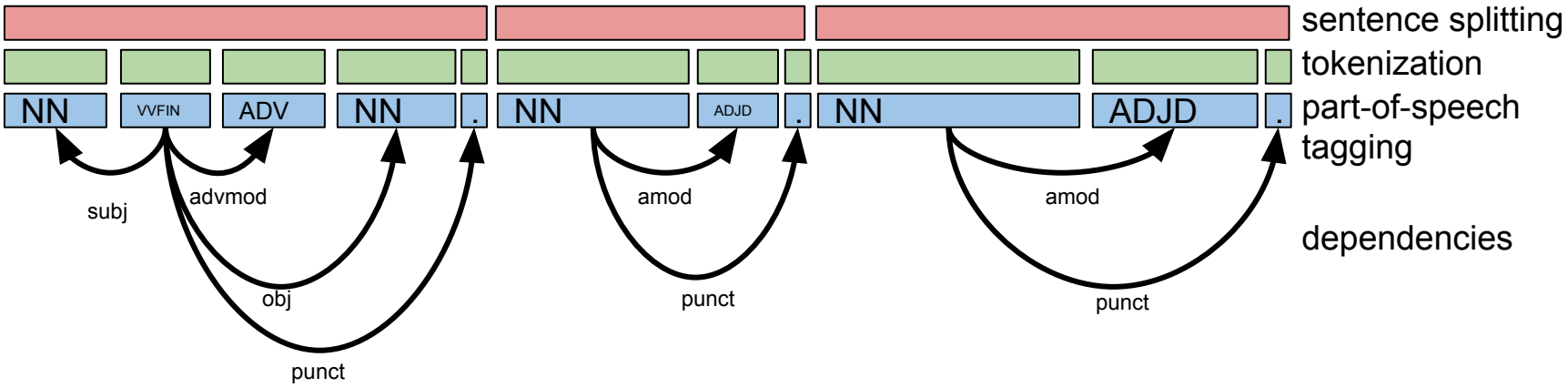
NLP Pipeline

Pat. hat viel Durst. Appetit gut. Stuhlgang normal.

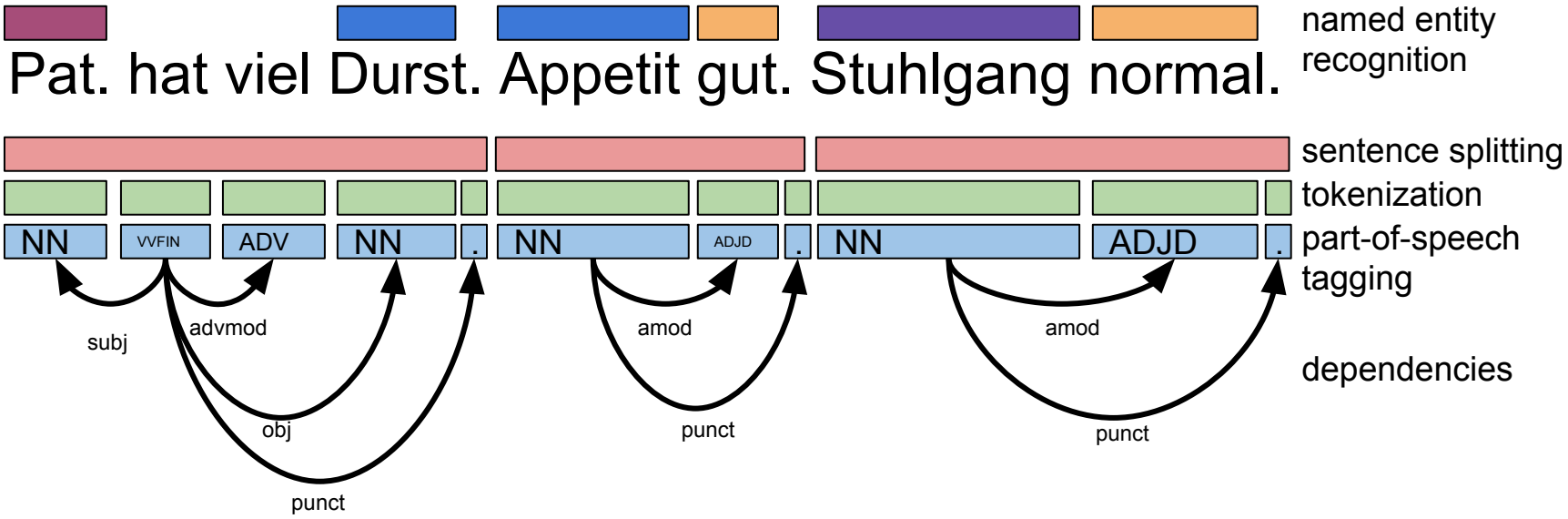


NLP Pipeline

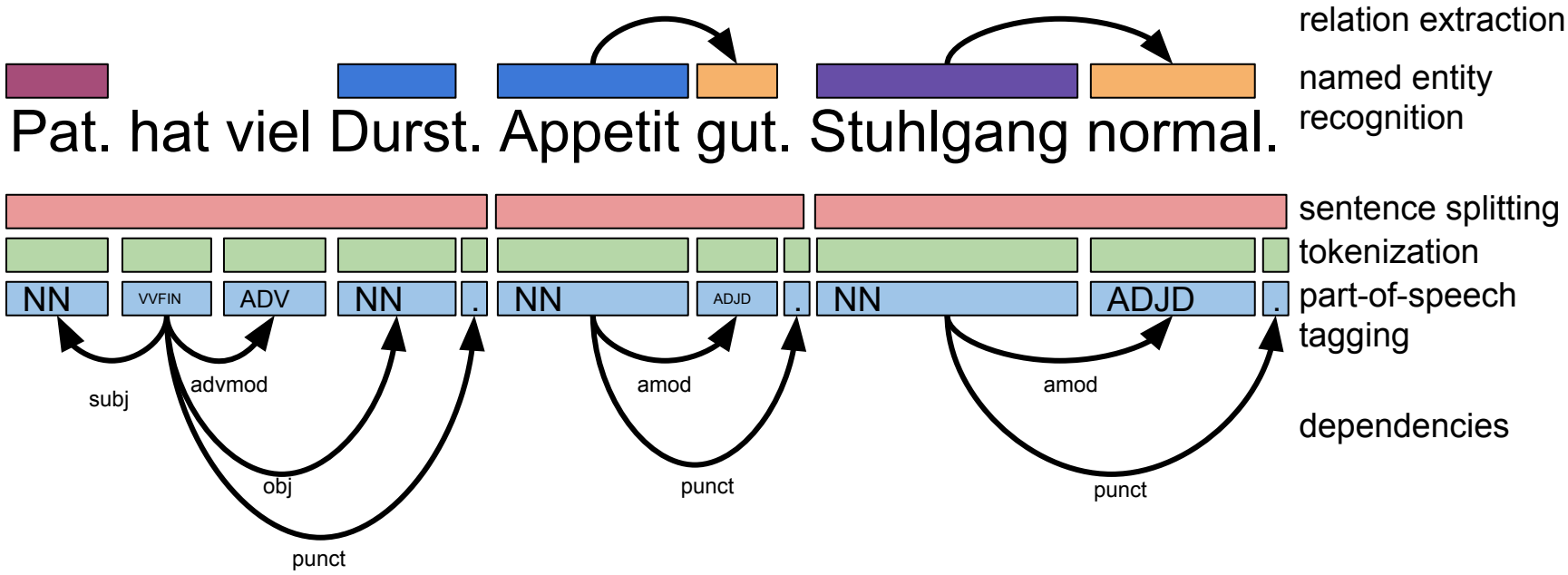
Pat. hat viel Durst. Appetit gut. Stuhlgang normal.



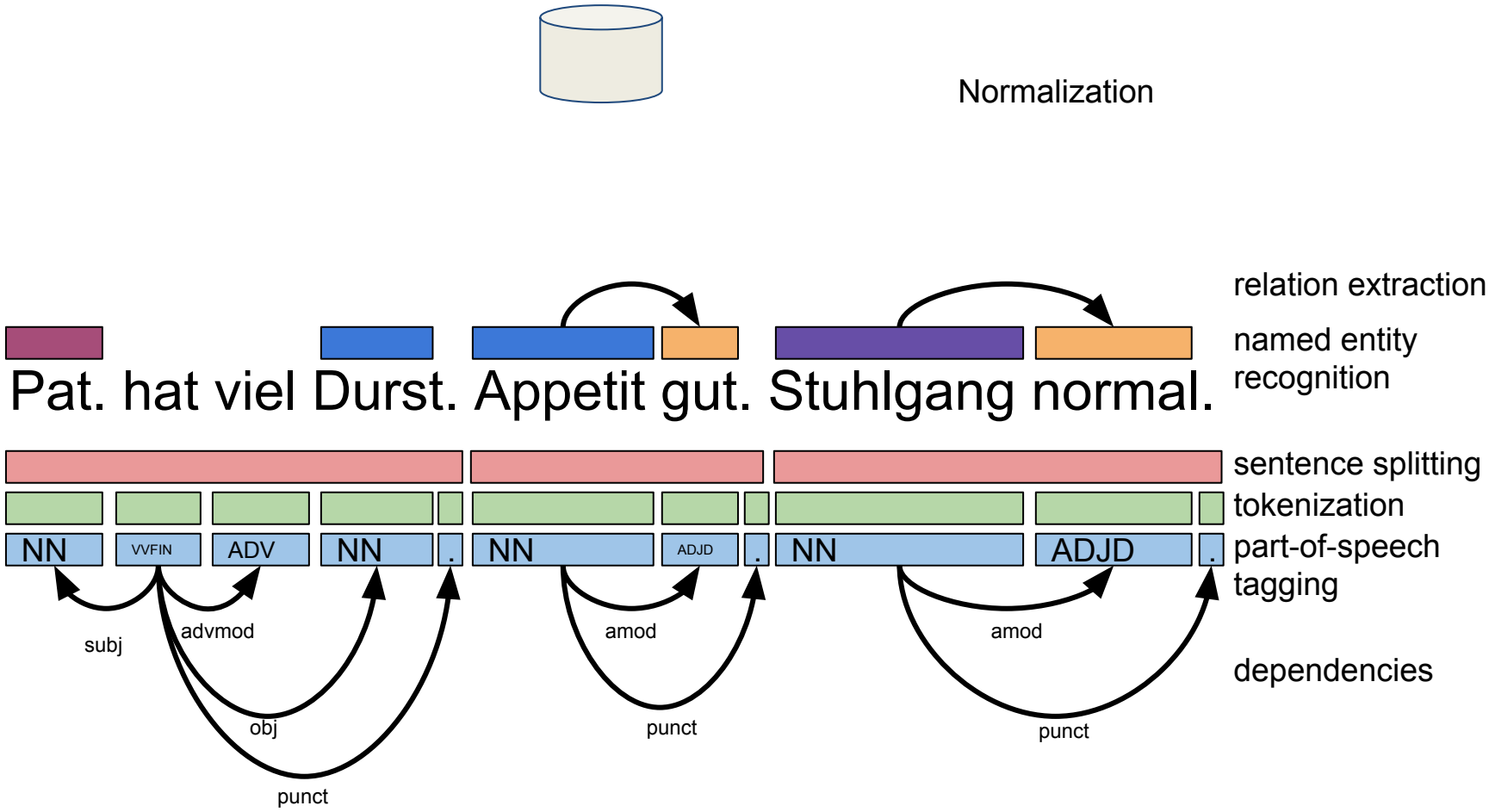
NLP Pipeline



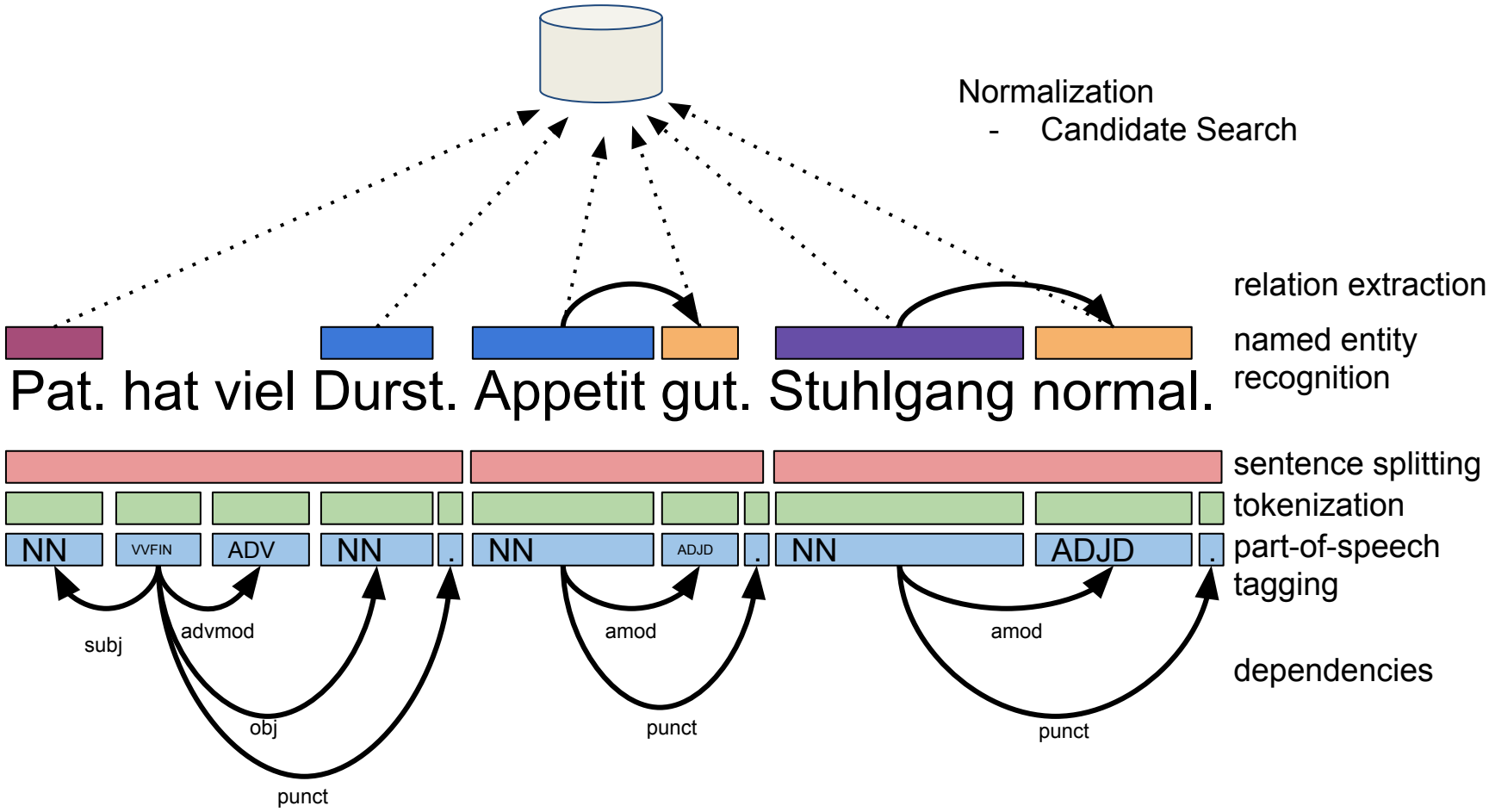
NLP Pipeline



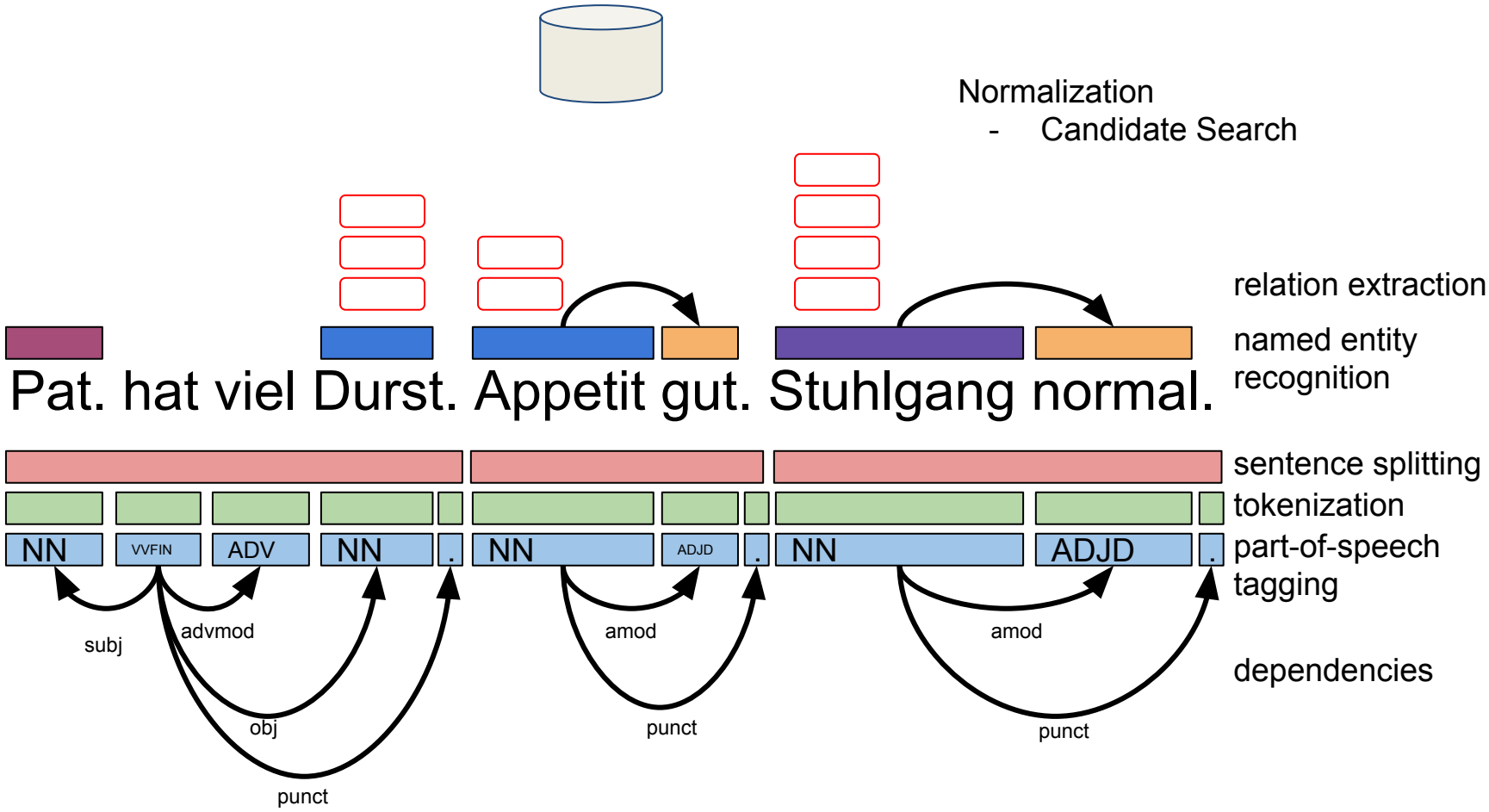
NLP Pipeline



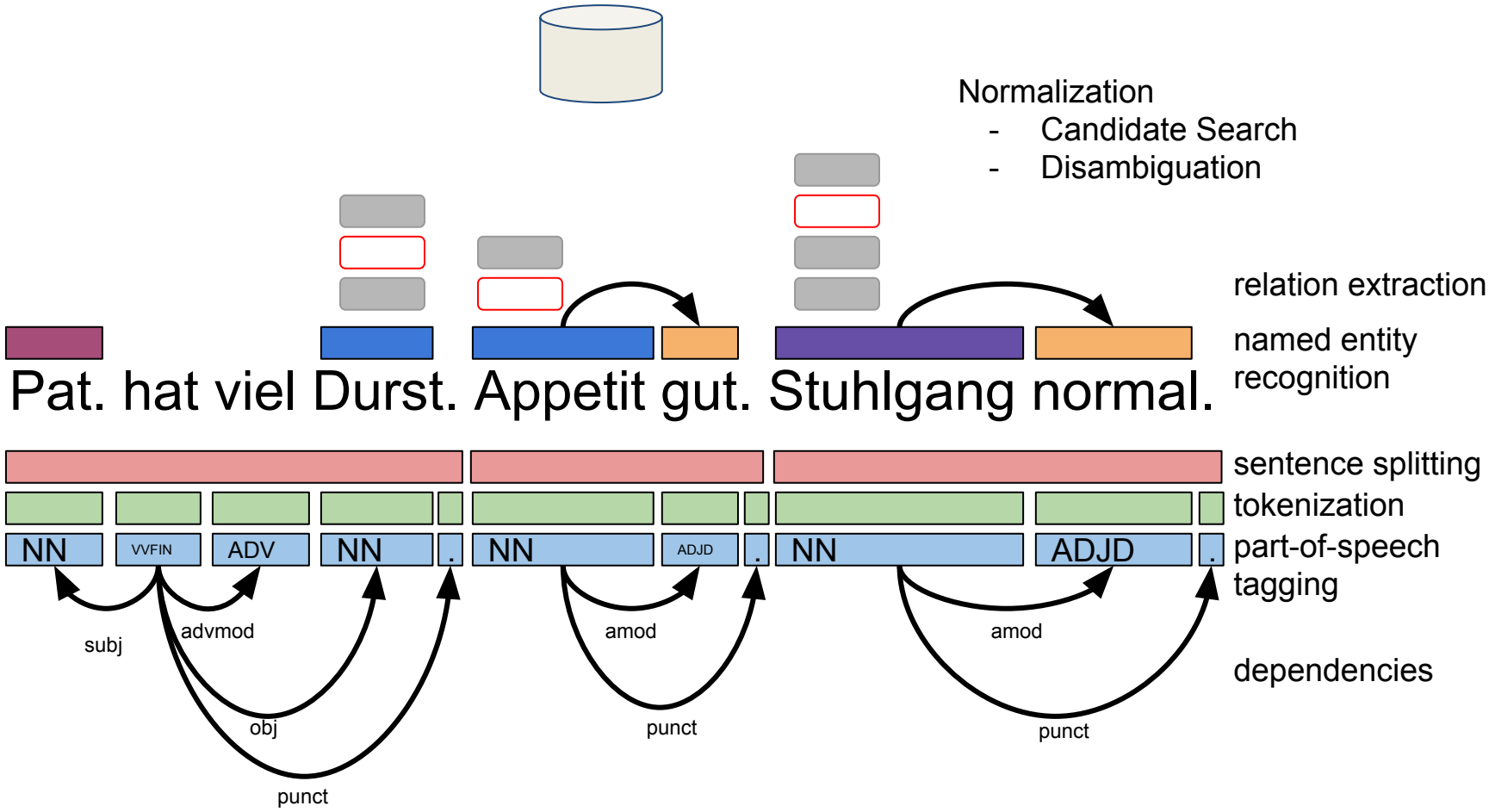
NLP Pipeline



NLP Pipeline



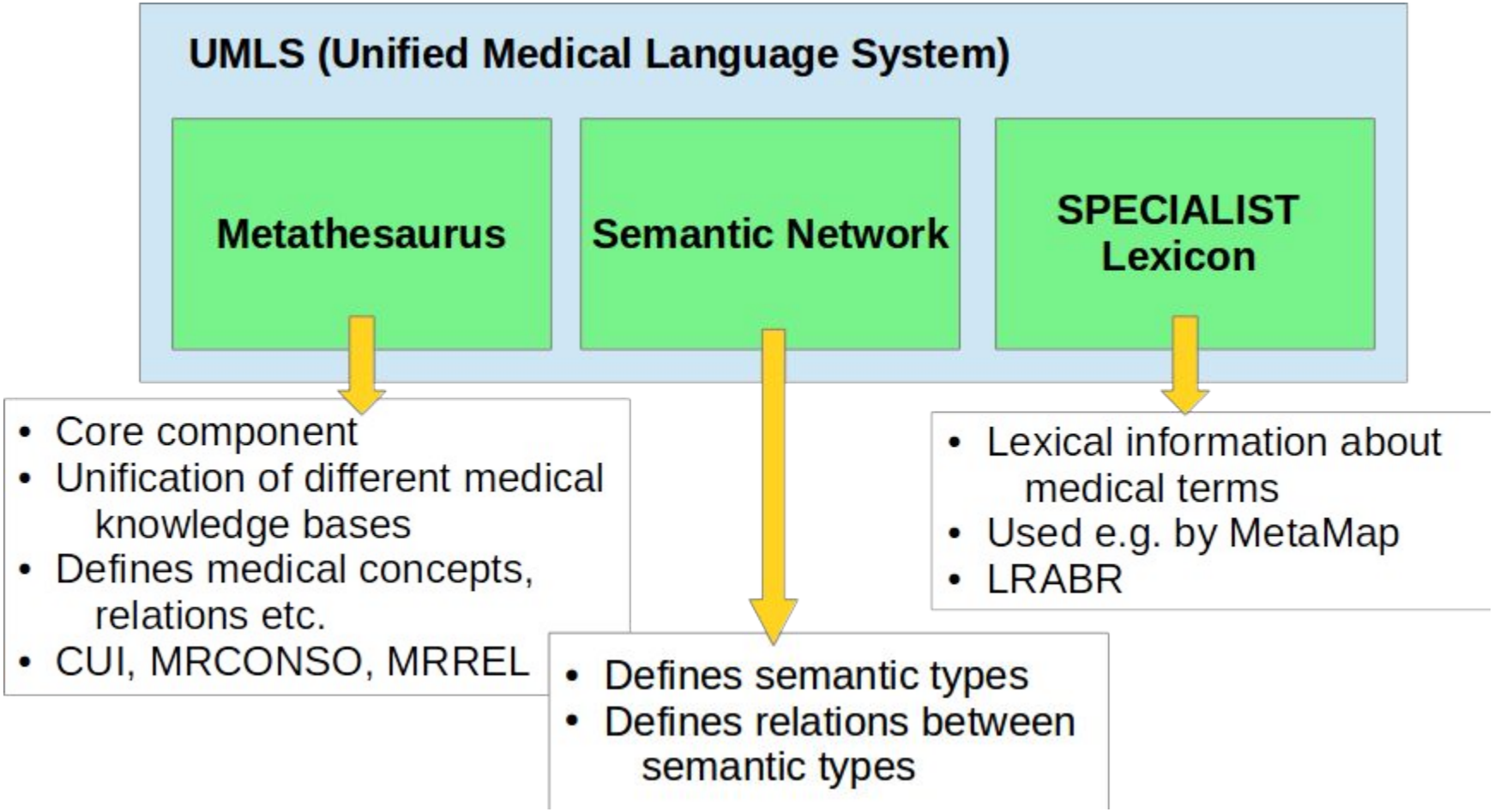
NLP Pipeline



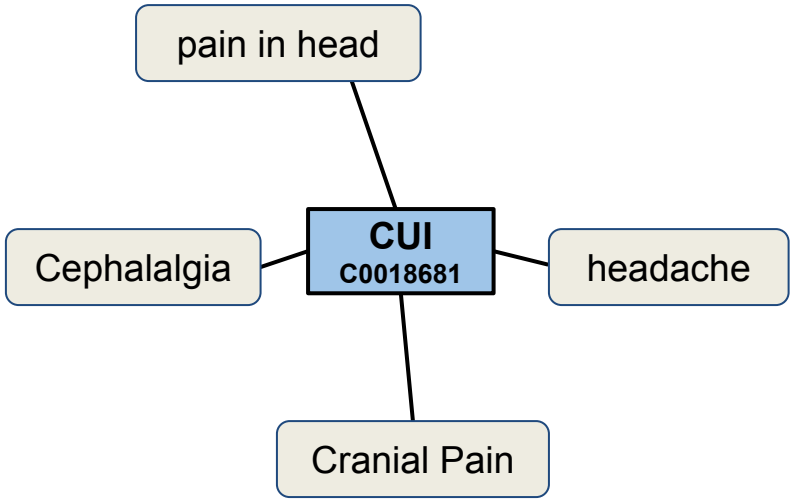
Normalization

- The same concept can be expressed in many different ways (Abbreviations)
- Distinct identification of concepts
  - e.g. Information access
- Examples of ambiguous terms:
  - **cold:** temperature, common cold, chronic obstructive lung disease, cold therapy…
  - **blood pressure:** Arterial Blood Pressures (Finding), Blood pressure (Organism Function), taking blood pressure (Health Care Activity)

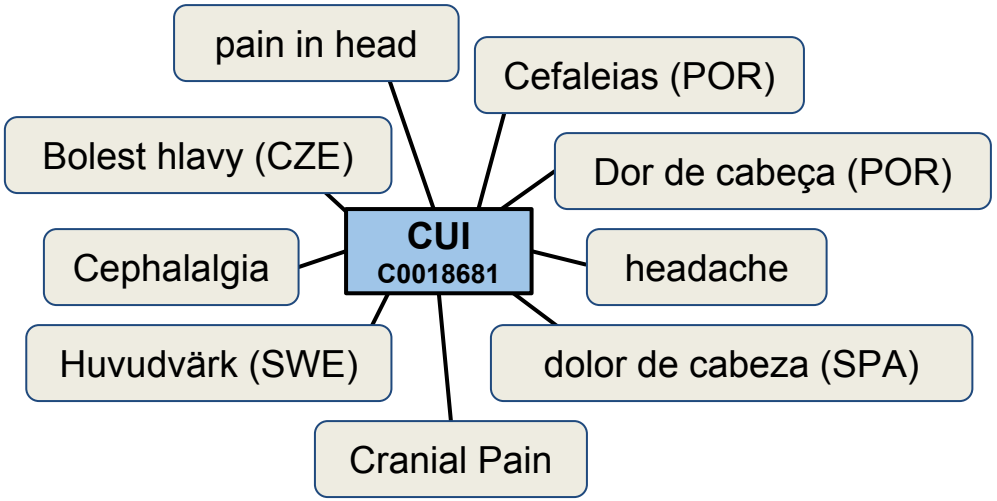
UMLS


UMLS (Unified Medical Language System)
Metathesaurus
Semantic Network
SPECIALIST Lexicon
• Core component
• Unification of different medical knowledge bases
• Defines medical concepts, relations etc.
• CUI, MRCONSO, MRREL
• Defines semantic types
• Defines relations between semantic types
• Lexical information about medical terms
• Used e.g. by MetaMap
• LRABR

UMLS


pain in head
Cephalalgia
CUI
C0018681
headache
Cranial Pain

UMLS


pain in head
Cefaleias (POR)
Bolest hlavy (CZE)
Dor de cabeça (POR)
CUI
C0018681
Cephalalgia
headache
Huvudvärk (SWE)
dolor de cabeza (SPA)
Cranial Pain

UMLS

UMLS includes 25 languages
~ 70% English
~ 10% Spanish
~ 3% French
~ 2% German

UMLS

UMLS includes 25 languages
~ 70% English
~ 10% Spanish
~ 3% French
~ 2% German

Non-English Normalization against UMLS:
Ambiguity is not the main problem!
Task: how to find the right concept

UMLS

UMLS includes 25 languages
~ 70% English
~ 10% Spanish
~ 3% French
~ 2% German

First idea: use Google Translate or Bing Translator to translate unknown terms to increase recall
BUT:

- services not for free if you start extensive tests
- sending clinical data via Internet might be not that what you want

UMLS

UMLS includes 25 languages
~ 70% English
~ 10% Spanish
~ 3% French
~ 2% German

First idea: use Google Translate or Bing Translator to translate unknown terms to increase recall
BUT:

- services not for free if you start extensive tests
- sending clinical data via Internet might be not that what you want

UMLS

CUI	English	German	Spanish	French	Swedish	Russian
C0007097	carcinoma	Karzinom	carcinoma	carcinome	Karcinom	KARTSINOMA
C0027804	Neurasthenia	Neurasthenie	neurastenia	Neurasthnie	Neurasteni	NEVRASTENIIA
C0012503	Dioxins	Dioxine	Dioxinas	Dioxines	Dioxiner	DIOKSINY
C0023531	Leukoplakia	Leukoplakie	Leucoplaquia	Leucoplasie	Leukoplaki	LEUKOPLAKIJA

Table 1. Similar words of different languages in UMLS linked by the same CUI

UMLS includes 25 languages

~ 70% English

~ 10% Spanish

~ 3% French

~ 2% German

UMLS

CUI	English	German	Spanish	French	Swedish	Russian
C0007097	carcinoma	Karzinom	carcinoma	carcinome	Karcinom	KARTSINOMA
C0027804	Neurasthenia	Neurasthenie	neurastenia	Neurasthnie	Neurasteni	NEVRASTENIIA
C0012503	Dioxins	Dioxine	Dioxinas	Dioxines	Dioxiner	DIOKSINY
C0023531	Leukoplakia	Leukoplakie	Leucoplaquia	Leucoplasie	Leukoplaki	LEUKOPLAKIJA

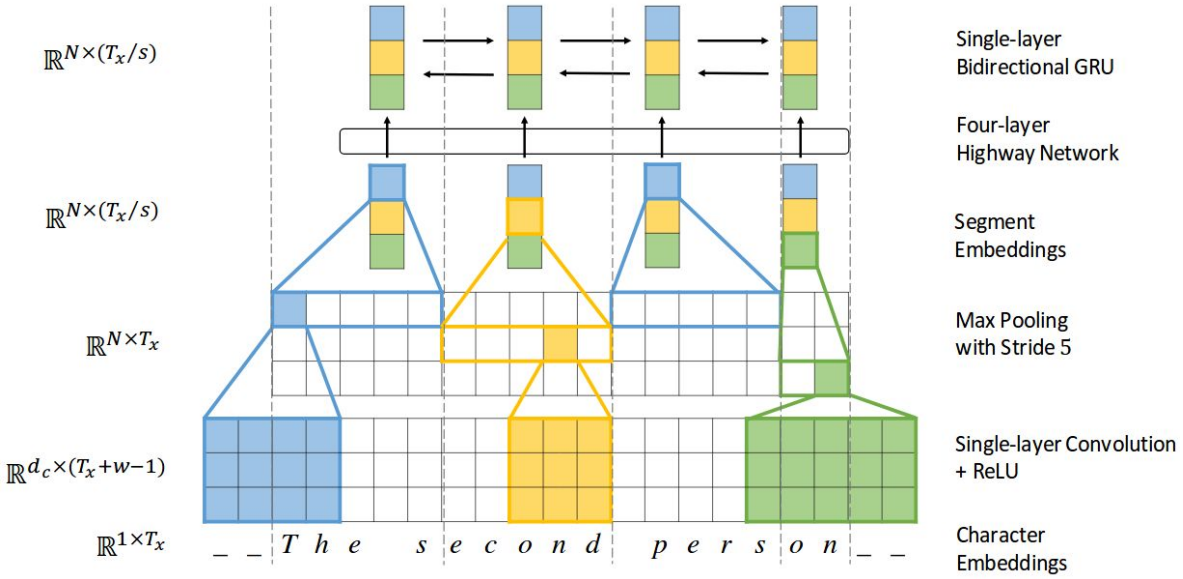
Table 1. Similar words of different languages in UMLS linked by the same CUI

UMLS includes 25 languages
~ 70% English
~ 10% Spanish
~ 3% French
~ 2% German

Baseline idea: learn to convert Latin-/Greek-rooted words

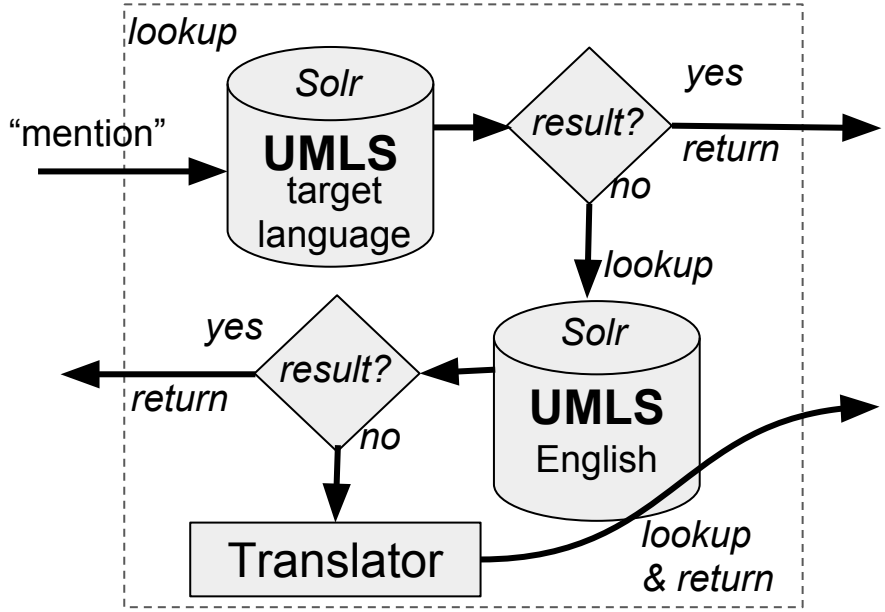
Neural Translation Model

- Character-based neural translation model based on Lee et al., (2016)
- Training data:
  - Parallel data of UMLS
  - FreeDict dictionary



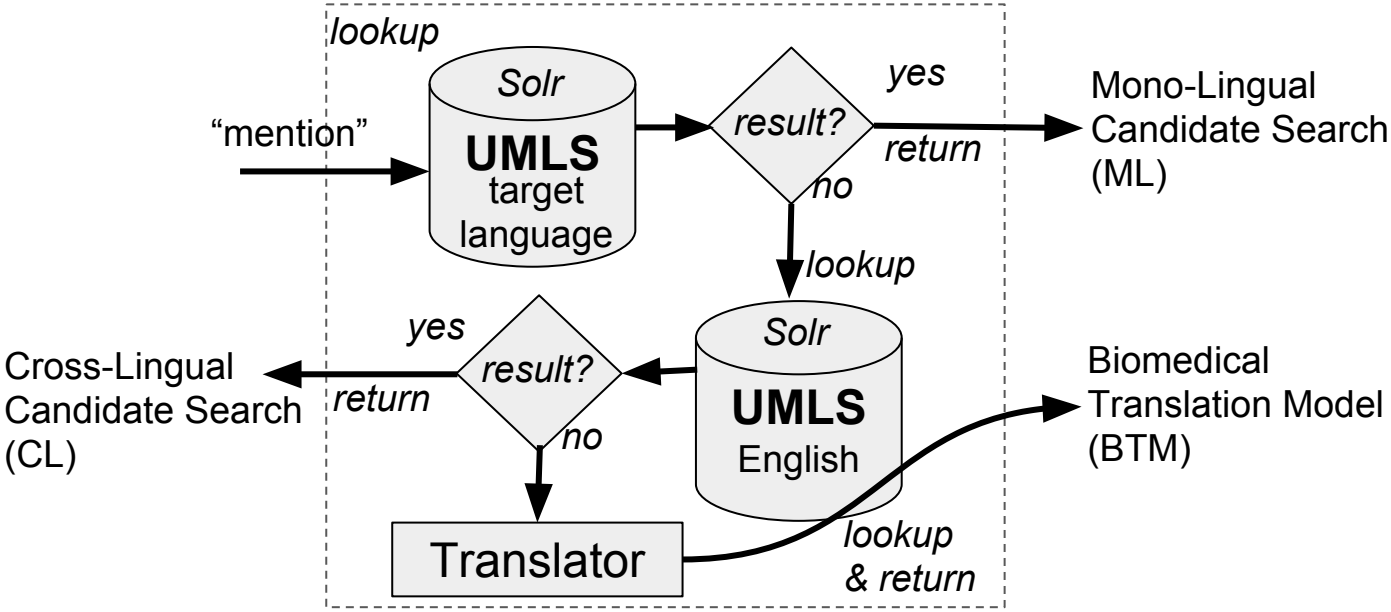
Concept Normalization

- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



Concept Normalization

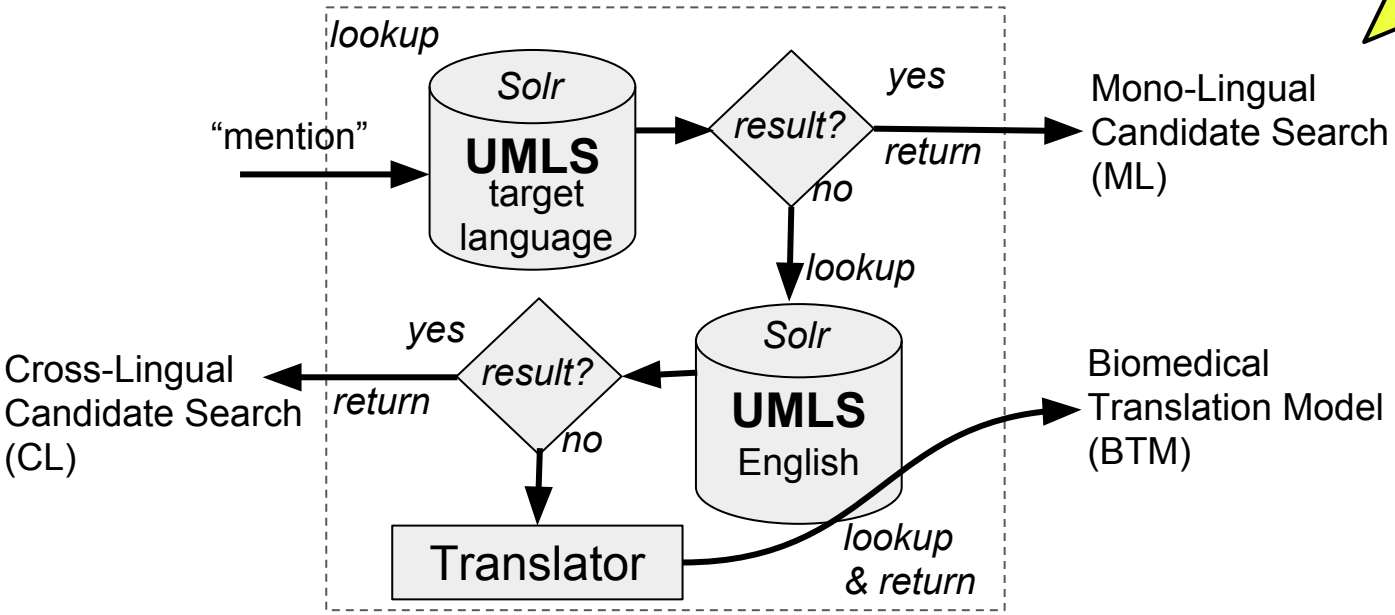
- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



Concept Normalization

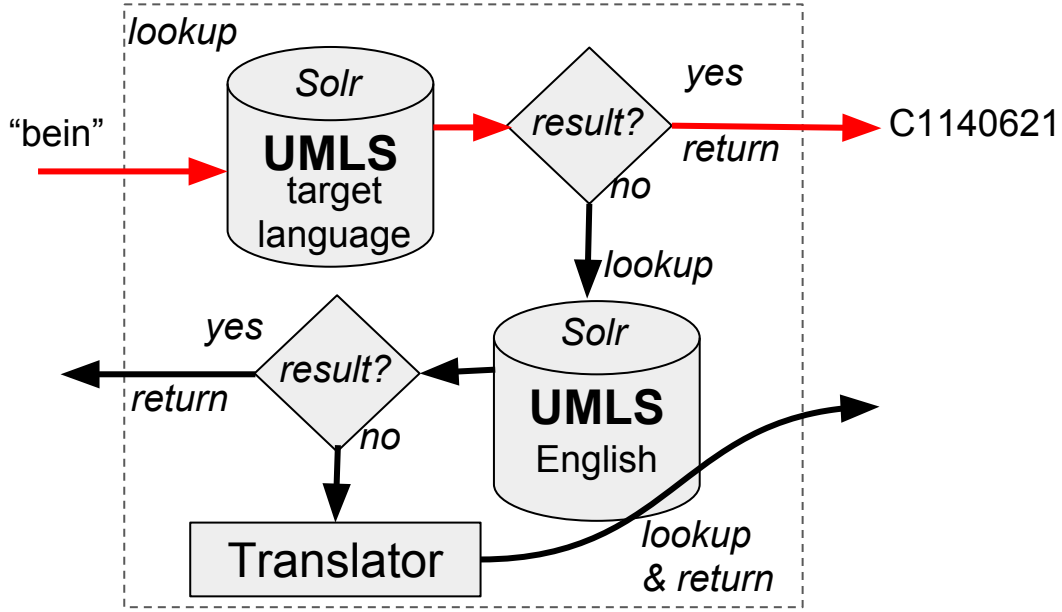
- Normalization often two steps:
  - **Candidate Search** (increase rec
  - Disambiguation (solving ambigu

Note: A good optimization of Solr can already help!



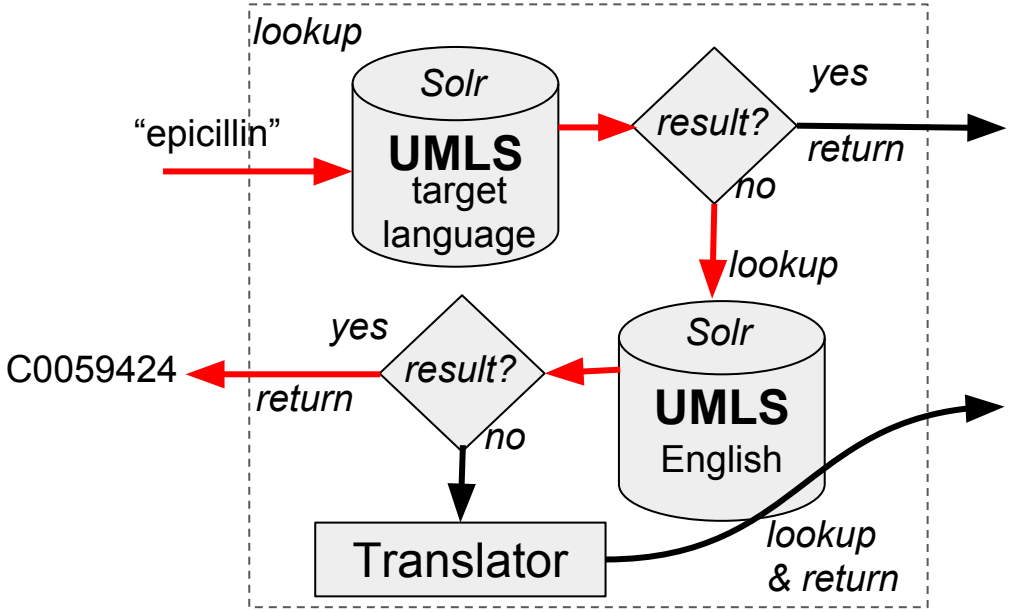
Concept Normalization

- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



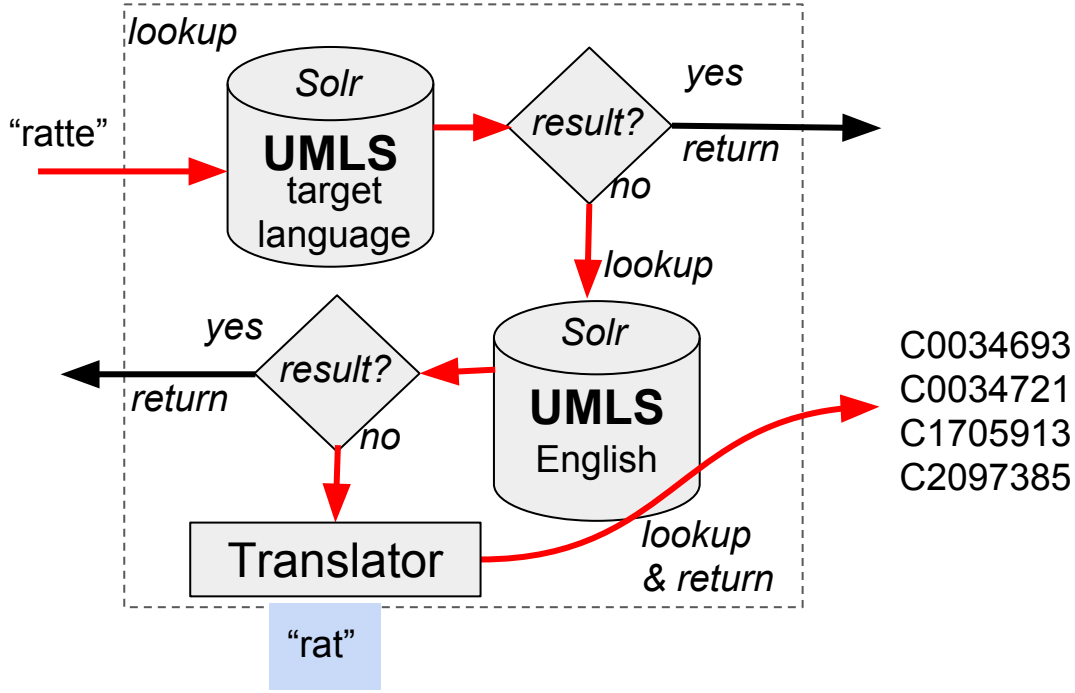
Concept Normalization

- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



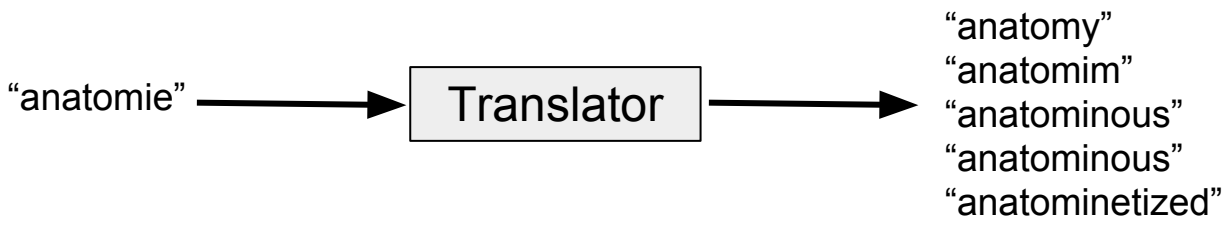
Concept Normalization

- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



Concept Normalization

- Normalization often two steps:
  - **Candidate Search** (increase recall)
  - Disambiguation (solving ambiguity)



Evaluation Data

- **Quaero Corpus** used for CLEF eHealth 2015 Task 1b (Neveol et al., 2015) and 2016 Task 2 (Neveol et al., 2016)
  - French Medline titles and EMEA abstracts
- **Mantra**
  - Medline titles, EMEA abstracts and EPO patents for GER, SPA, FRE, DUT
  - much smaller than Quaero
  - comparison to Google Translate & Bing Translator: manual translation by our students

Results

- Comparison to the best system of CLEF

	Medline			EMEA		
Method	P	R	F1	P	R	F1
ML	0.831	0.575	0.680	0.911	0.632	0.746
CL	**0.834**	0.611	0.705	0.919	0.764	0.834
BTM	0.831	**0.661**	**0.736**	0.909	0.772	0.835
Erasmus	0.805	0.575	0.671	**1.000**	**0.774**	**0.872**

Evaluation: CLEF eHealth 2015

	Medline			EMEA		
Method	P	R	F1	P	R	F1
ML	**0.800**	0.594	0.682	**0.822**	0.552	0.661
CL	0.786	0.620	0.693	0.808	0.676	0.736
BTM	0.771	**0.663**	**0.713**	0.781	**0.692**	**0.734**
SIBM	0.594	0.515	0.552	0.604	0.463	0.524

Evaluation: CLEF eHealth 2016

Results

	SPA			FRE			DUT			GER		
Method	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ML	**0.799**	0.561	0.659	**0.814**	0.469	0.595	**0.800**	0.357	0.494	**0.833**	0.493	0.620
CL	0.788	0.583	0.670	0.795	0.502	0.615	0.769	0.424	0.546	0.817	0.530	0.643
BTM	0.781	**0.619**	**0.691**	0.780	0.593	0.674	0.725	0.533	0.614	0.771	0.582	0.663
GB	0.790	0.607	0.687	0.794	**0.604**	**0.686**	0.767	**0.560**	**0.648**	0.804	**0.588**	**0.679**

Evaluation against Mantra corpus

Conclusion

- Free Neural Translation Model for Cross-lingual Concept Normalization
- Recall can be increased, but final results also strongly depend on disambiguation
- Usage in clinical context:
  - we have to deal with a special vocabulary (many abbreviations) which can be not covered by translator -> resolution first

Thank you