Building and Mining Knowledge graphs

(KEN4256)

Lecture 3: Constructing a Knowledge Graph from Unstructured Data

Maastricht University
Institute of Data Science

id: KEN4256_L3
version: 1.2024.0
created: February 2, 2019
last modified: March 26, 2024
published on: March 26, 2024

Introduction

Knowledge in the form of natural language (e.g. text) offers a rich source of information for answering questions.

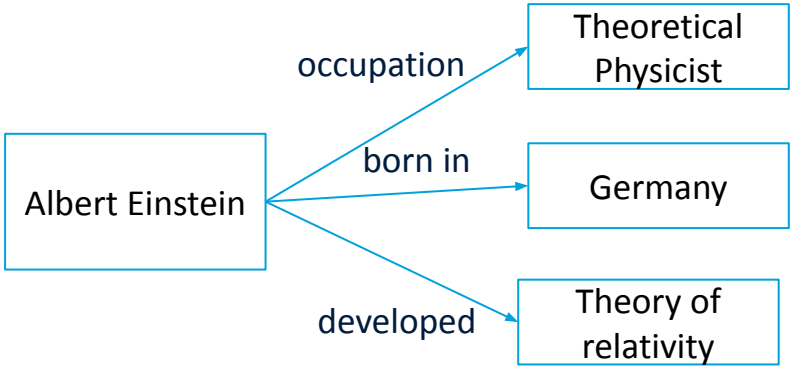
Text is not easily processable by machines in the sense that the program will fully understand what is stated.

Natural language processing (NLP) can be used to make the semantic content of text processable by machines.

From text to (RDF) graph

Albert Einstein was a **German**-born **theoretical physicist** who developed the **theory of relativity**.

1. recognize the entities and relations (named entity recognition & relation extraction)
2. map to identifiers (entity/relation linking)
3. project to target schema



NLP

(classic) NLP involves a pipeline with some or all of the steps:

- **tokenisation**: segment text into words, punctuation, tokens
- **normalisation**: transforms abbreviations, slang, upper/lower case, etc. into a standard form
- **stopword** removal. remove common words e.g. the, or, of
- **stemming & lemmatisation**. replace words such as "build", "builds", "build", "buildler", and "building" with the same word (and stem) "build"
- **parts of speech (PoS) tagging**. assign word types to tokens such as nouns, verbs, adjectives, and adverbs.
- **dependency parsing**: find relations between parts of speech

Named Entity Recognition

NER is a task of assigning one of pre-defined types to each word/word phrase in the text.

Albert Einstein was a German-born theoretical physicist who is best known for developing the theory of relativity. In 1905, he was awarded a PhD by the University of Zurich and received the 1921 Nobel Prize in Physics "for his services to theoretical physics.

[PER Albert Einstein] was a [LOC German]-born theoretical physicist who is best known for developing the theory of relativity. In [TIME 1905], he was awarded a PhD by the [ORG University of Zurich]and received the [TIME 1921] Nobel Prize in Physics "for his services to theoretical physics.

NER Methods

Early Methods
- Dictionary-based
- Rule-based

Traditional Machine Learning
- HMM / CRF / MEMM

Deep Learning
- CNN/RNN CRF
- Attention-based
- Transfer Learning

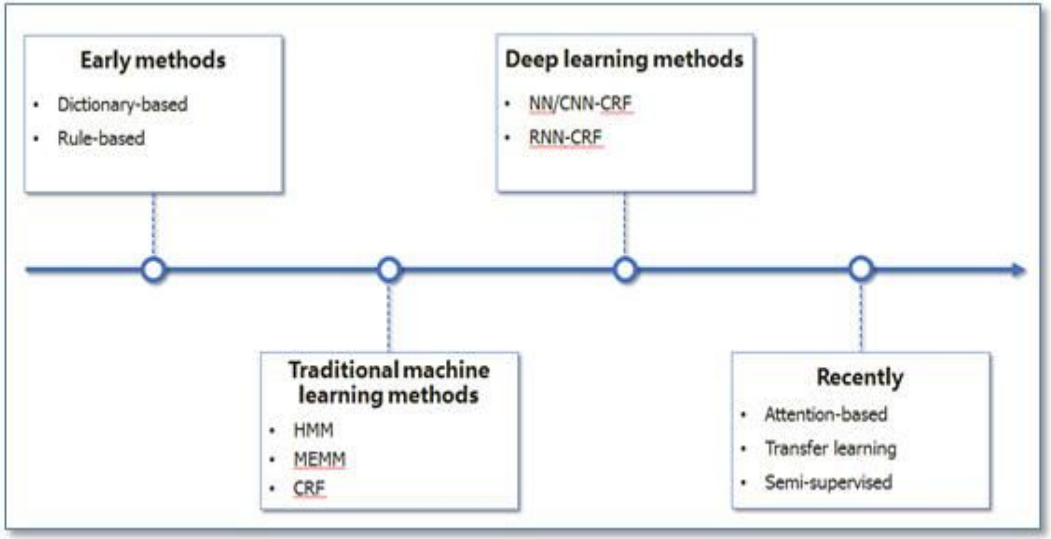


Image: https://www.programmersought.com/article/48534259085/

Approaches to NER

- Dictionary-based
  - use a pre-defined dictionary of terms to find exact matches in the text.
  - include synonyms and misspellings to increase recall
- Rule-based
  - Express the extraction rules in a formal rule language
    - Regular expressions, such as address (city + province + country…)
    - References to dictionary
    - Invoke custom extractors

Approaches to NER

- Language Models
  - Task-independent training
    - Train the model on the domain of interest
  - Task-dependent training
    - Introduce special tags in the input

BERT : Bidirectional Encoder Representations from Transformers

- **Bidirectional** - look back and forward in a sentence to understand the meaning
- **Transformers** - The Transformer reads entire sequences of tokens at once. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. The attention mechanism allows for learning contextual relations between words.
- **(Pre-trained) contextualized word embeddings** - Encode words based on their meaning/context.

in 2018, BERT showed state of the art performance for a number of tasks such as natural language inference, sentiment analysis, question answering, paraphrase detection, linguistic acceptability

Named Entity Linking

NEL is the task of linking entity mentions with their corresponding objects in a target database/ontology.

"Floyd revolutionized rock with the Wall"

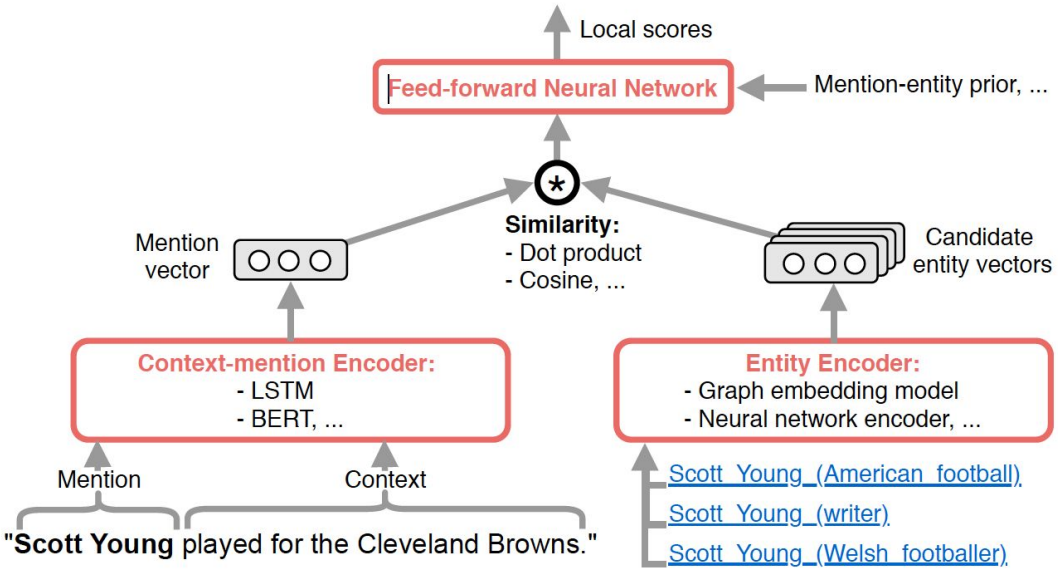
.../wiki/Pink_Floyd
.../wiki/Floyd_(name)
.../wiki/Floyd,_Iowa

.../wiki/Rock_(geology)
.../wiki/The_Rock
.../wiki/Rock_Music

.../wiki/Defensive_Wall
.../wiki/Berlin_Wall
.../wiki/The_Wall_(album)

NEL Methods

- Syntactic
- Semantic
- Contextual
- Embedding



Challenges in NER and NEL

- Ambiguity
  - Louis Vouitton—can be company, person, or product
- Training data
  - Data is usually small and incomplete
- Domain-specific Variations
  - Michael Jordan -> MJ, Michael Jeffrey Jordan
- Many different forms of an entity
  - Need to have a lexicon

Relation Extraction

RE is the task of identifying relationships between entity mentions.

Albert Einstein was a German-born theoretical physicist who is best known for developing the theory of relativity. In 1905, he was awarded a PhD by the University of Zurich and received the 1921 Nobel Prize in Physics "for his services to theoretical physics.

- Albert Einstein ***born in*** Germany
- Albert Einstein ***occupation*** Theoretical physicist
- Theoretical physicist ***branch of*** Physics
- ……

Approaches to Relation Extraction

- Syntactic patterns (or rule-based)
  - To discover pattern for a new relation, collect several examples of that relation
  - Look for generalities to discover new patterns

The Netherlands has many well-known universities, such as Maastricht University, having strong research capabilities.

Even though we have never heard of Maastricht University, but we can extract that it is a kind of well-known university

Approaches to Relation Extraction

- Syntactic patterns (or rule-based) (known as Hearst Pattern)

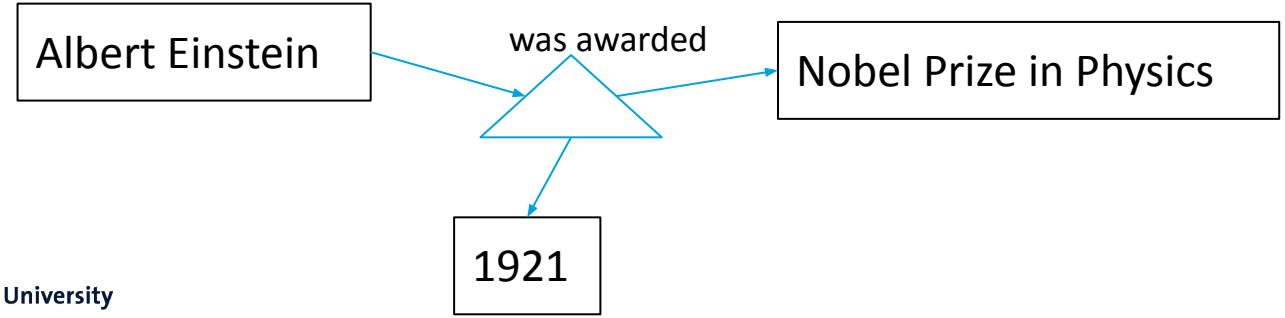
Pattern Name	Example
such as	... works by authors *such as* Herric, Goldsmith, and Shakespear ...
or other	Bruises, wounds, broken bones, *or other* injuries ...
and other	... temples, treasuries, *and other* Civic Buildings, ...
including	All common law countries *including* Canada and England ...
especially	Most European countries *especially* France, England, and Spain, ...

Approaches to Relation Extraction

- Supervised learning
  - Requires a huge amount of training data
  - We can use syntactic patterns to generate training data
  - We can write approximate labelling functions

Challenges in Relation Extraction

- Open information extraction
  - Does not rely on a designed set of relations
  - Can be difficult to use/understand the relations
- N-ary relation
  - N-ary relations involve more than two entities
  - Requires complex methods and high-quality data



NLP Tools

- Many tools for information extraction have been released.
  - Named Entity Recognition
    - StandfordNER
    - OpenNLP
    - SpaCy
    - NLTK
  - Named Entity Linking
    - DBPedia Spotlight
    - CoNEREL
  - Relation Extraction
    - CiceroLite
    - FOX
    - Open Calais

DBpedia Spotlight

Confidence: 0.5

Language: English

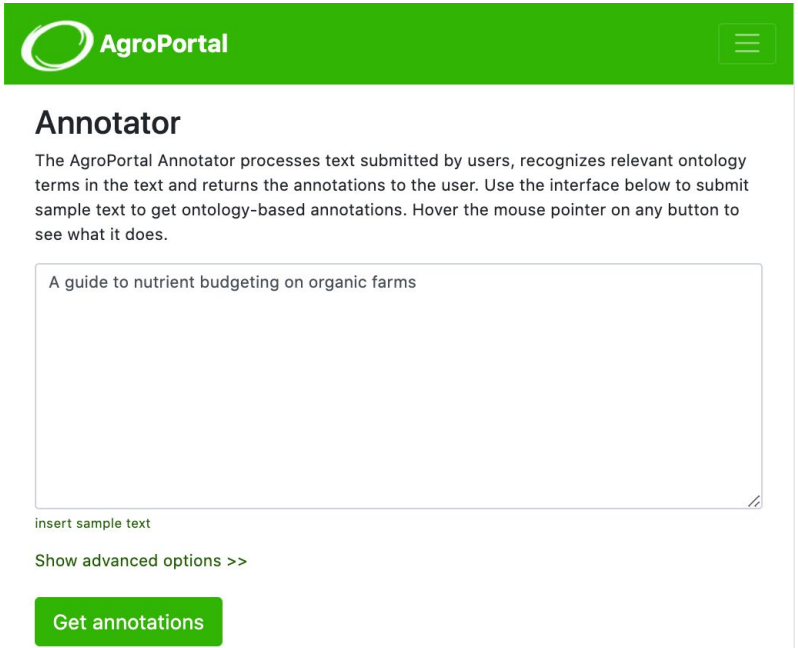
n-best candidates

SELECT TYPES... ANNOTATE

Berlin is the capital and largest city of Germany by both area and population. Its 3,769,495 inhabitants as of 31 December 2019 make it the most-populous city of the European Union, according to population within city limits. The city is also one of Germany's 16 federal states. It is surrounded by the state of Brandenburg, and contiguous with Potsdam, Brandenburg's capital. The two cities are at the center of the Berlin-Brandenburg capital region, which is, with about six million inhabitants and an area of more than 30,000 km2, Germany's third-largest metropolitan region after the Rhine-Ruhr and Rhine-Main regions. Berlin straddles the banks of the River Spree, which flows into the River Havel (a tributary of the River Elbe) in the western borough of Spandau. Among the city's main topographical features are the many lakes in the western and southeastern boroughs formed by the Spree, Havel, and Dahme rivers (the largest of which is Lake Müggelsee). Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. About one-third of the city's area is composed of forests, parks, gardens, rivers, canals and lakes. The city lies in the Central German dialect area, the Berlin dialect being a variant of the Lusatian-New Marchian dialects.

http://dbpedia.org/resource/Berlin

AgroPortal & BioPortal




```
[{
"annotations": [
    {
      "from": 42,
      "to": 46,
      "matchType": "PREF",
      "text": "FARMS"
    }
  ],
  "mappings": [

  ]
},{
  "annotatedClass": {
    "definition": [
      "An area of land which is used for the cultivation of crops or grazing of livestock,
    ],
    "prefLabel": "farm",
    "synonym": [
      "agricultural site",
      "FARM",
      "farmstead",
      "farms",
      "farm",
      "ranch"
    ]
```

spaCy

spaCy is a **free, open-source library** for advanced **Natural Language Processing** (NLP) in Python.
https://spacy.io

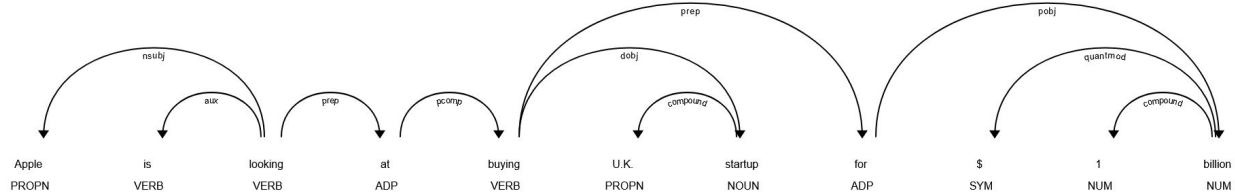
```
!python -m spacy download en_core_web_sm

import spacy
nlp = spacy.load('en_core_web_sm')
doc = nlp("Apple is looking at buying U.K. startup for $1 billion")

for token in doc:
        print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_)
```

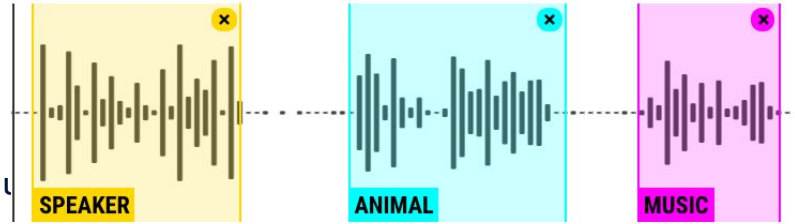
TEXT	LEMMA	POS	TAG	DEP
Apple	apple	PROPN	NNP	nsubj
is	be	AUX	VBZ	aux
looking	look	VERB	VBG	ROOT
at	at	ADP	IN	prep
buying	buy	VERB	VBG	pcomp
U.K.	u.k.	PROPN	NNP	compound
startup	startup	NOUN	NN	dobj
for	for	ADP	IN	prep
$	$	SYM	$	quantmod
1	1	NUM	CD	compound
billion	billion	NUM	CD	pobj

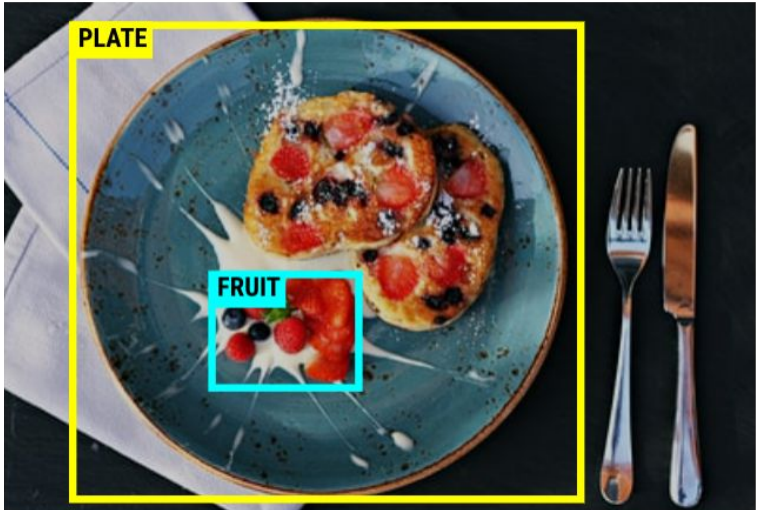
```
from spacy import displacy
displacy.serve(doc, style="dep")
```



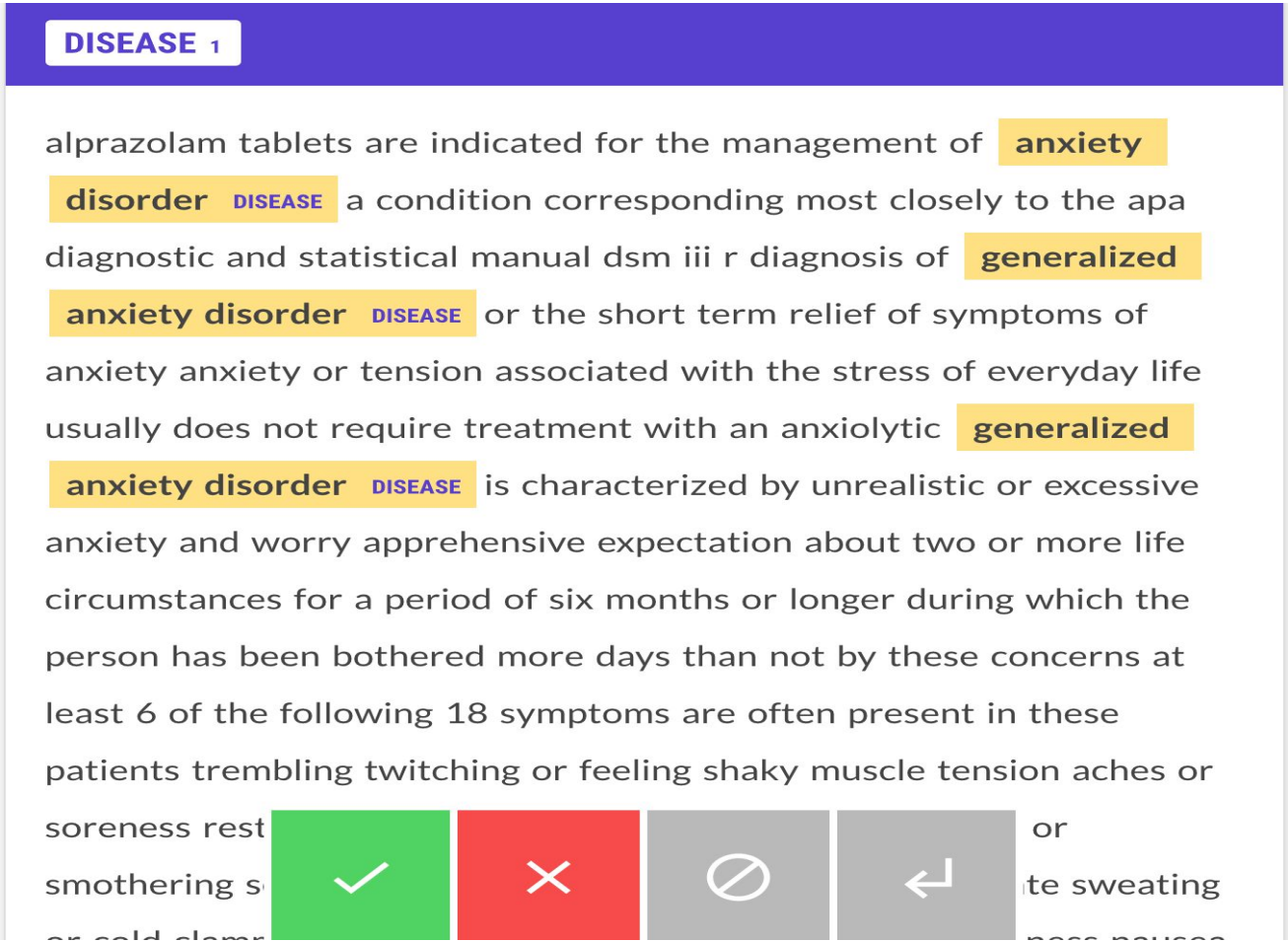
Prodigy

- For creating training and evaluation data for machine learning
- Python library + web UI
- ‘Fast, intuitive and efficient annotation’

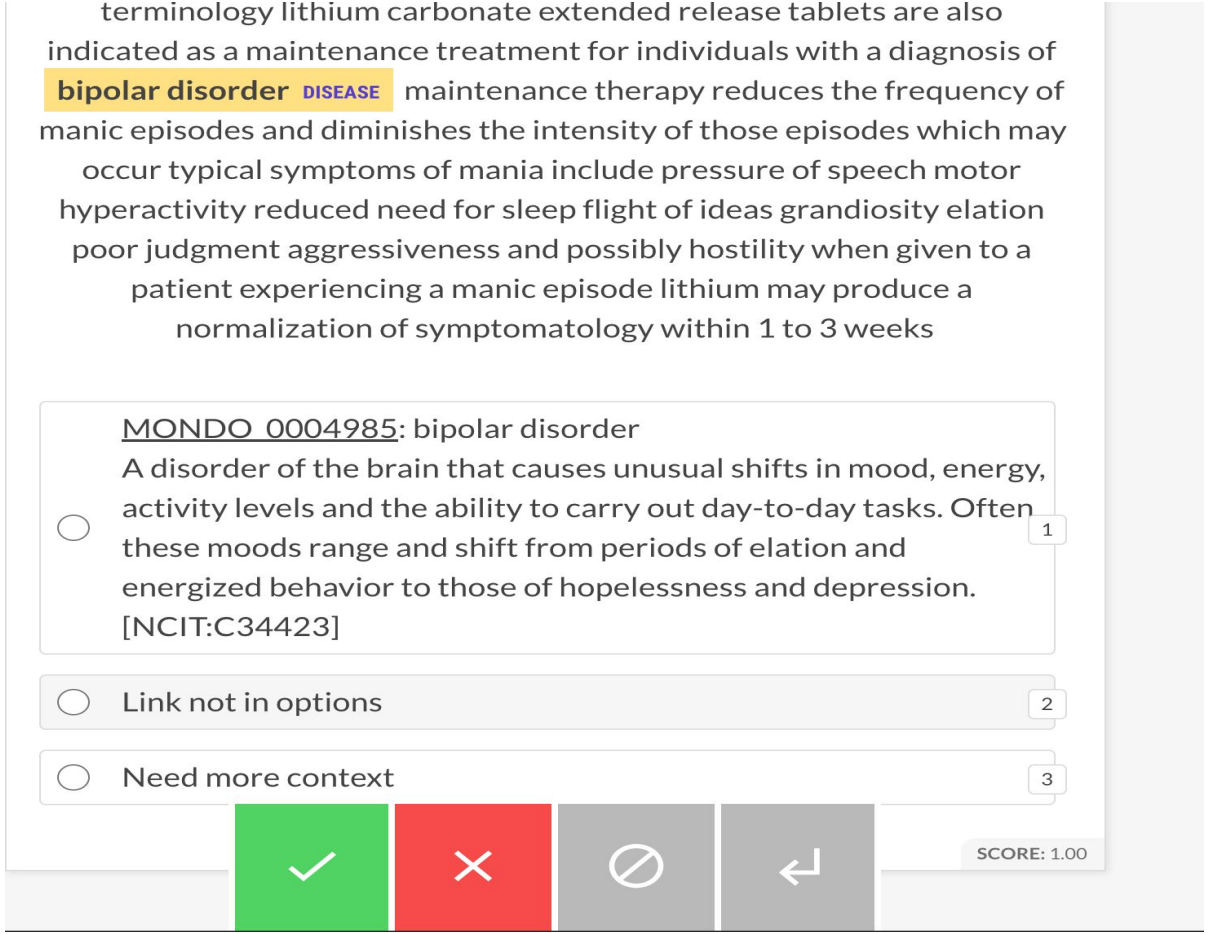




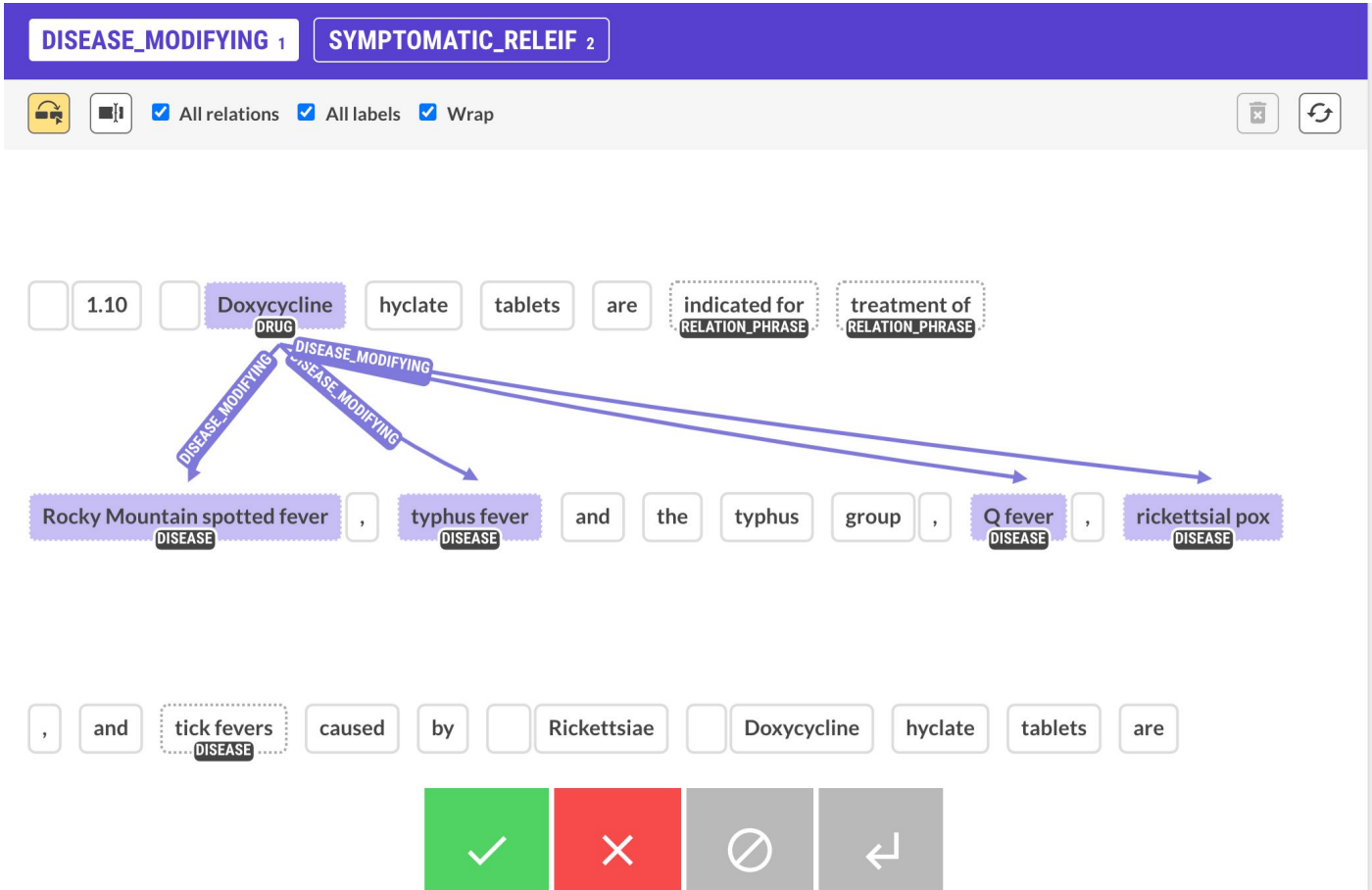
Named Entity Recognition



Named Entity Linking

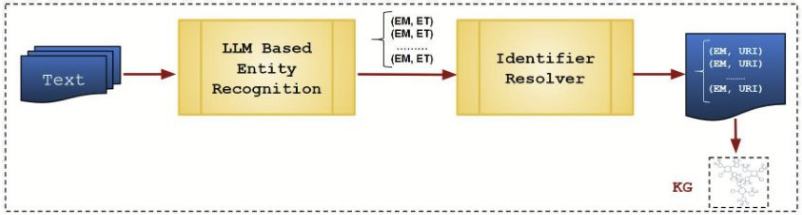


Relation Extraction



LLM-powered graph construction

- prompt an LLM to extract the entities
- use a tool to assign the identifier
- construct the graph




```
Paragraphs (separated by " | "): {" | ".join(newSentences)}
Instructions:
Inside each Paragraph, identify all context words/phrases (allowed context types: "age group",
"co-morbidity", "symptom", "co-therapy", "adjunct therapy", "past therapies", "treatment duration",
"conditional", "co-prescribed medication", "genetics", "temporal aspects") and output a JSON
dictionary with Paragraphs as keys and lists of {{[context type]: [corresponding words/phrases as a
list of values (more than one possible)]}} as values.
Definition of context types:
"conditional" - a statement about when the medication is appropriate to use
"target" - the condition (symptom or illness) that is intended to be treated
"co-prescribed medication"- drugs commonly prescribed together with the given drug (not therapeutic
procedures! -> for that use "co-therapy")
"co-therapy" - procedures or therapies that should be applied in combination with the drug (not
medications or substances! -> for that use "co-prescribed medication")
"co-morbidity" - diseases or conditions that commonly occur together (with a target condition) in
the same patients
"genetics"-  particular genetic strains of a disease
"temporal aspects"-  information which explains at what life stage, disease stage, or treatment
phase a drug should be administered
Other context types are self-explanatory.
Only output the resulting JSON.
```

Table 5: Third prompt evaluation. The precision, recall, and F1-score are provided for each type of context. "Support" represents the number of pairs considered. We omitted the "genetics" context type, as it has support equal to zero.

	Precision	Recall	F1-score	Support
Target	**0.81**	0.79	0.80	214
Symptom	**0.67**	0.70	0.68	66
Age Group	**0.91**	**0.96**	**0.93**	71
Adjunct Therapy	0.39	**1.00**	**0.84**	8
Co-morbidity	0.58	0.28	0.38	25
Treatment Duration	0.59	0.76	0.67	17
Co-therapy	0.73	**1.00**	**0.84**	8
Co-prescribed Medication	0.43	0.83	0.57	12
Conditional	0.56	0.77	0.65	65
Past Therapies	0.14	0.17	0.15	6
Temporal Aspects	0.40	0.67	0.50	3
Micro Average	0.69	0.78	0.73	502
Macro Average	0.56	0.72	0.61	502
Weighted Average	**0.72**	**0.78**	**0.74**	502
Samples Average	0.64	0.64	0.64	502

Summary

- Entity and relation extraction are fundamental problems to creating knowledge graphs from text
- Use of rule-based methods for training data generation that can be fed into pre-trained language models is becoming an increasingly popular paradigm
- Entity linking and resolution will eventually play an important role

Questions?