

Building and Mining Knowledge graphs

(KEN4256)

Lecture 3: Constructing a Knowledge Graph from Unstructured Data



Maastricht University

Institute of Data Science

© 2024 by Michel Dumontier and the Institute of Data Science at Maastricht University is licensed under Attribution 4.0 International
To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, even for commercial purposes.

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.

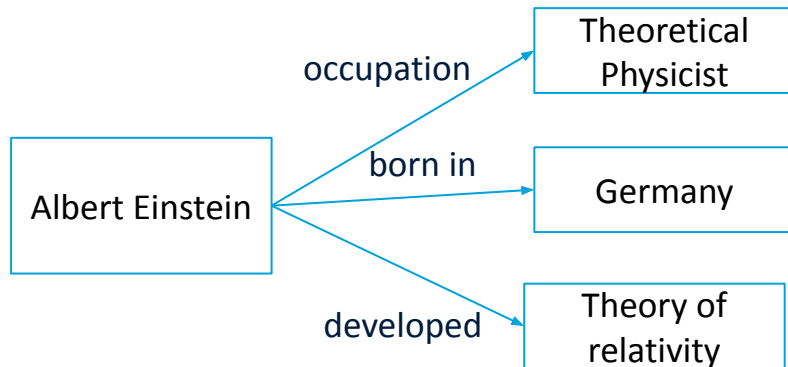
Text is not easily processable by machines in the sense that the program will fully understand what is stated.

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", "builder", 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

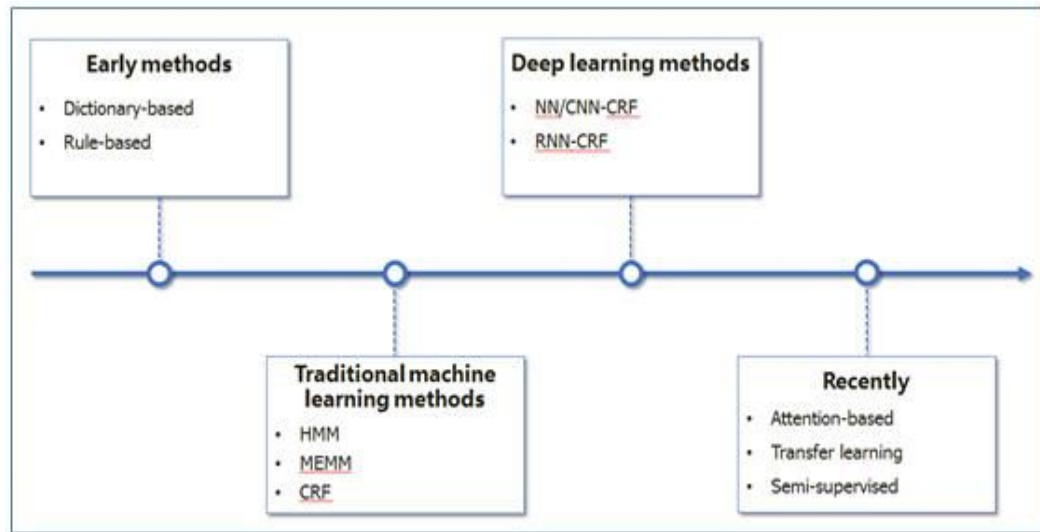


Image: <https://www.programmersought.com/article/48534259085/>

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”

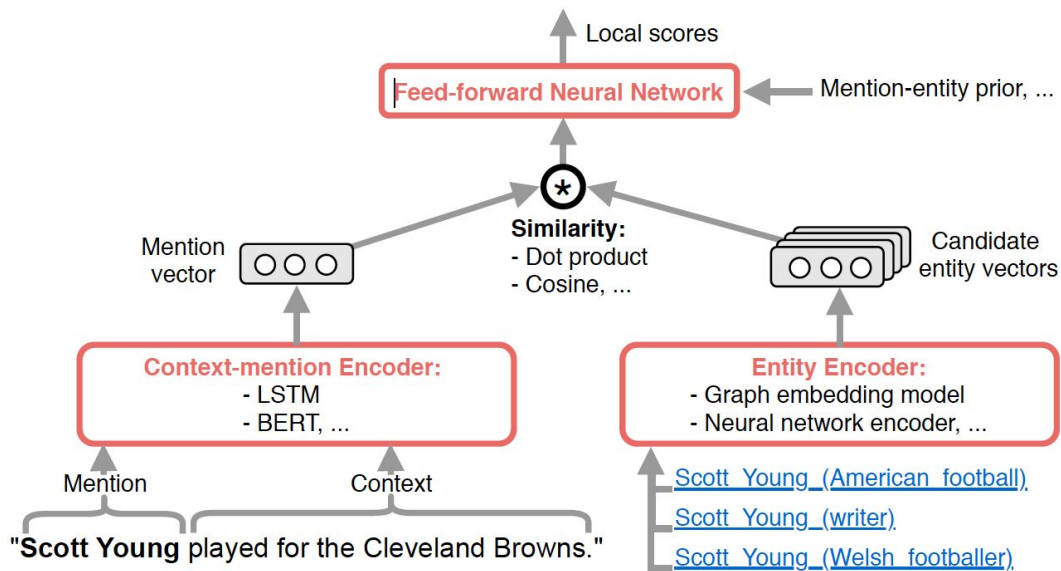
↓
.../wiki/**Pink_Floyd**
.../wiki/Floyd_(name)
.../wiki/Floyd,_Iowa

↓
.../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 Vuitton—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)

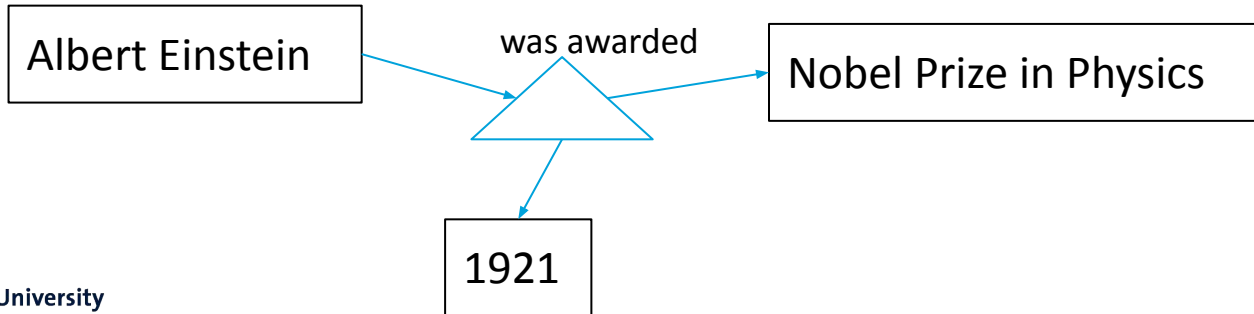
Pattern Name	Example
<i>such as</i>	... works by authors <i>such as</i> Herric, Goldsmith, and Shakespear ...
<i>or other</i>	Bruises, wounds, broken bones, <i>or other</i> injuries ...
<i>and other</i>	... temples, treasures, <i>and other</i> Civic Buildings, ...
<i>including</i>	All common law countries <i>including</i> Canada and England ...
<i>especially</i>	Most European countries <i>especially</i> 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
 - [StanfordNER](#)
 - [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

<http://dbpedia.org/resource/Berlin>
[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](#).

AgroPortal & BioPortal



Annotator

The AgroPortal Annotator processes text submitted by users, recognizes relevant ontology terms in the text and returns the annotations to the user. Use the interface below to submit sample text to get ontology-based annotations. Hover the mouse pointer on any button to see what it does.

A guide to nutrient budgeting on organic farms

insert sample text

Show advanced options >>

Get annotations

```
[{
  "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 is a **free, open-source library** for advanced **Natural Language Processing (NLP)** in Python.

<https://spacy.io>

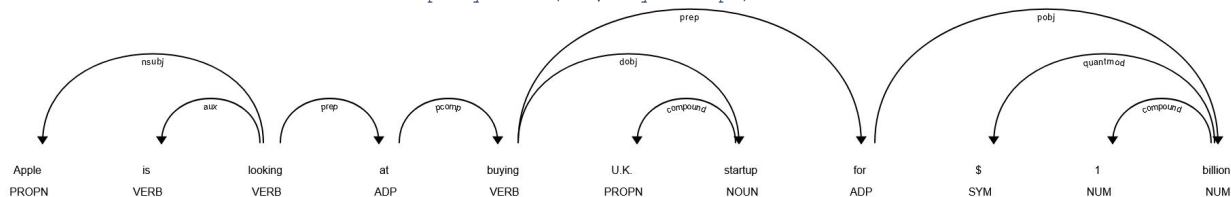
```
!python -m spacy download en_core_web_sm

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_)
```

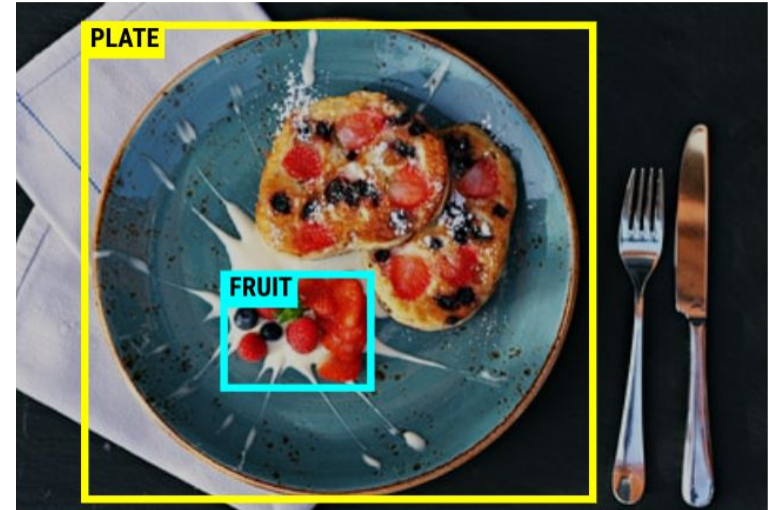
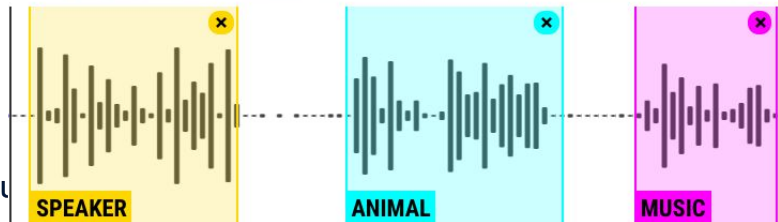
TEXT	LEMMA	POS	TAG	DEP
Apple	apple	PROPN	NNP	nsubj
is	be	AUX	VBZ	aux
looking	look	VERB	VBG	ROOT
at	at	ADP	IN	prep
buying	buy	VERB	VBG	pcomp
U.K.	u.k.	PROPN	NNP	compound
startup	startup	NOUN	NN	dobj
for	for	ADP	IN	prep
\$	\$	SYM	\$	quantmod
1	1	NUM	CD	compound
billion	billion	NUM	CD	pobj

```
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'



PERSON 1

ORG 2

In 1917, former newspaperman **William Gordon** PERSON enlists in **the U.S. Army** ORG . The day before he is to leave Washington, D.C. for the fighting in Europe, he meets socialite **Joel Carter** PERSON . The couple spend the day together. He tells her that, because he once wrote a book on cryptography under a pen name, the army is searching for him to put him to work behind a desk, but he is eager to fight the Germans.

SOURCE: CMU Movie Summary Corpus

Named Entity Recognition

DISEASE 1

alprazolam tablets are indicated for the management of **anxiety disorder** DISEASE a condition corresponding most closely to the apa diagnostic and statistical manual dsm iii r diagnosis of **generalized anxiety disorder** DISEASE or the short term relief of symptoms of anxiety anxiety or tension associated with the stress of everyday life usually does not require treatment with an anxiolytic **generalized anxiety disorder** DISEASE is characterized by unrealistic or excessive anxiety and worry apprehensive expectation about two or more life circumstances for a period of six months or longer during which the person has been bothered more days than not by these concerns at least 6 of the following 18 symptoms are often present in these patients trembling twitching or feeling shaky muscle tension aches or soreness rest smothering s or te sweating



or
te sweating
ness nervous

Named Entity Linking

terminology lithium carbonate extended release tablets are also indicated as a maintenance treatment for individuals with a diagnosis of **bipolar disorder** DISEASE maintenance therapy reduces the frequency of manic episodes and diminishes the intensity of those episodes which may occur typical symptoms of mania include pressure of speech motor hyperactivity reduced need for sleep flight of ideas grandiosity elation poor judgment aggressiveness and possibly hostility when given to a patient experiencing a manic episode lithium may produce a normalization of symptomatology within 1 to 3 weeks

MONDO 0004985: bipolar disorder



A disorder of the brain that causes unusual shifts in mood, energy, activity levels and the ability to carry out day-to-day tasks. Often these moods range and shift from periods of elation and energized behavior to those of hopelessness and depression. [NCIT:C34423]

1



Link not in options

2



Need more context

3







SCORE: 1.00

Relation Extraction

DISEASE_MODIFYING 1

SYMPTOMATIC_RELEIF 2

☒ All relations ☒ All labels ☒ Wrap



1.10

Doxycycline

DRUG

hyclate

tablets

are

indicated for

RELATION_PHRASE

treatment of

RELATION_PHRASE

DISEASE_MODIFYING

DISEASE_MODIFYING

DISEASE_MODIFYING

DISEASE_MODIFYING

Rocky Mountain spotted fever

DISEASE

,

typhus fever

DISEASE

and

the

typhus

group

,

Q fever

DISEASE

,

rickettsial pox

DISEASE

,

and

tick fevers

DISEASE

caused

by

Rickettsiae

Doxycycline

hyclate

tablets

are

✓

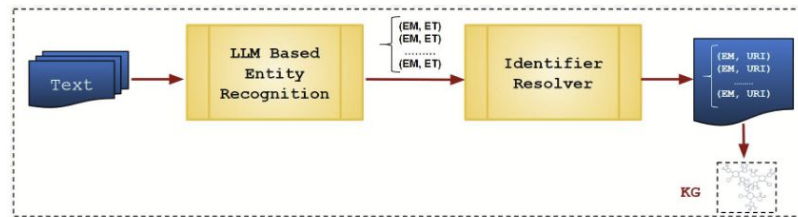
✗

⊘

↩

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?