Knowledge Extraction from text

Relationship Extraction

Oscar RODRÍGUEZ ROCHA
Teach on Mars
oscar.rodriguez@teachonmars.com

IC 2023

References

- **Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition** - Daniel Jurafsky and James H. Martin

- **Information Extraction and Named Entity Recognition** - Elena Cabrio

- **Information Extraction and Named Entity Recognition** - Christopher Manning

- **Introduction to Information Extraction** - Elena Demidova

- **Practical Natural Language Processing** - Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta & Harshit Surana

Course outline

- Intro to NLP
  - NLP in Real World / NLP tasks
  - What is language?
  - Approaches to NLP
- Intro to Information Extraction
  - IE applications / tasks / pipeline
  - Key phrase extraction
  - NER
    - Part-of-Speech Tagging
    - Named Entities and Named Entity Tagging
  - Named Entity Disambiguation and Linking
  - Relationship Extraction
    - REL algorithms
    - POS pattern matching
- Advanced IE tasks: temporal information extraction, event extraction, template filling
- Intro to DBpedia

Course outline

- Intro to NLP
  - NLP in Real World / NLP tasks
  - What is language?
  - Approaches to NLP
- Intro to Information Extraction
  - IE applications / tasks / pipeline
  - Key phrase extraction
  - NER
    - Part-of-Speech Tagging
    - Named Entities and Named Entity Tagging
  - Named Entity Disambiguation and Linking
  - Relationship Extraction
    - REL algorithms
    - POS pattern matching
- Advanced IE tasks: temporal information extraction, event extraction, template filling
- Intro to DBpedia

Relation Extraction

finding and classifying semantic relations among entities mentioned in a text

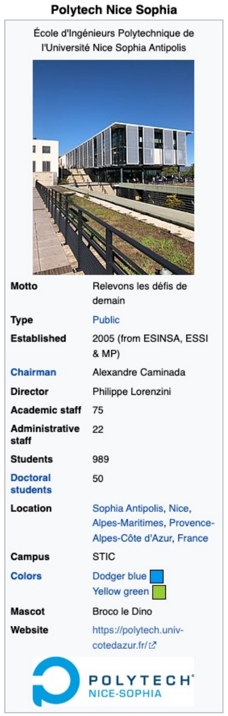
Relation extraction has a close relation to populating a relational database, and *knowledge graphs*, datasets of structured relational knowledge, are a useful way for search engines to present information to users.

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower- cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

Domain	$\mathcal{D} = \{a, b, c, d, e, f, g, h, i\}$
United, UAL, American Airlines, AMR	a, b, c, d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f, g, h, i
Classes	
United, UAL, American, and AMR are organizations	$Org = \{a, b, c, d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver, and San Francisco are places	$Loc = \{f, g, h, i\}$
Relations	
United is a unit of UAL	$PartOf = \{\langle a, b\rangle, \langle c, d\rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c, e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$

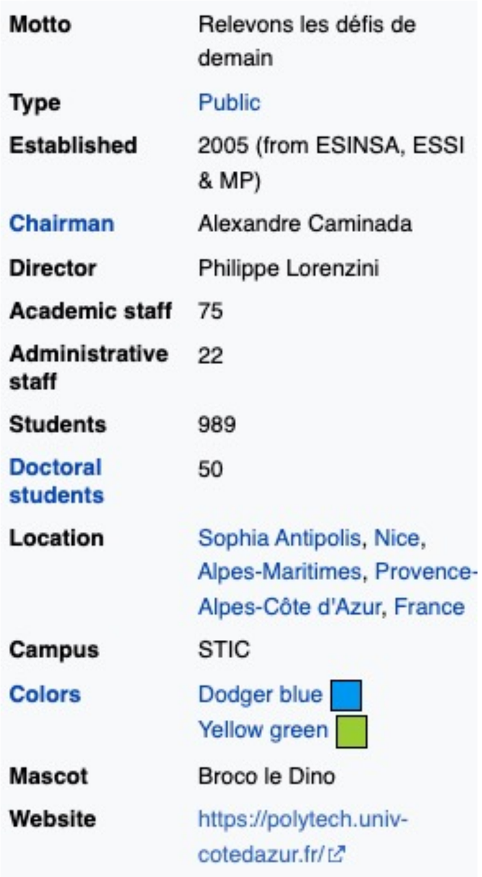
Real examples: Wikipedia



Polytech Nice Sophia

École d'Ingénieurs Polytechnique de l'Université Nice Sophia Antipolis

Motto	Relevons les défis de demain
Type	Public
Established	2005 (from ESINSA, ESSI & MP)
Chairman	Alexandre Caminada
Director	Philippe Lorenzini
Academic staff	75
Administrative staff	22
Students	989
Doctoral students	50
Location	Sophia Antipolis, Nice, Alpes-Maritimes, Provence-Alpes-Côte d'Azur, France
Campus	STIC
Colors	Dodger blue 🟦 Yellow green 🟩
Mascot	Broco le Dino
Website	https://polytech.univ-cotedazur.fr/



Motto	Relevons les défis de demain
Type	Public
Established	2005 (from ESINSA, ESSI & MP)
Chairman	Alexandre Caminada
Director	Philippe Lorenzini
Academic staff	75
Administrative staff	22
Students	989
Doctoral students	50
Location	Sophia Antipolis, Nice, Alpes-Maritimes, Provence-Alpes-Côte d'Azur, France
Campus	STIC
Colors	Dodger blue 🟦 Yellow green 🟩
Mascot	Broco le Dino
Website	https://polytech.univ-cotedazur.fr/

Real examples: DBpedia

DBpedia (Bizer et al., 2009) project aiming to extract structured content from the information created in the Wikipedia project, it contains over a 850 million triples (as of June 2021)

Relation Extraction Algorithms

5 main classes of algorithms

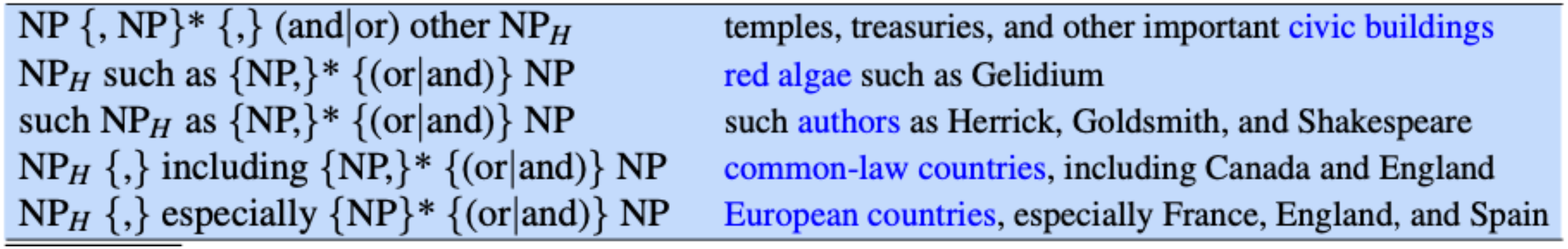
- handwritten patterns,
- supervised machine learning,
- semi-supervised (via bootstrapping)
- semi-supervised (distant supervision), and
- unsupervised.

Handwritten patterns

Hearst (lexico-syntactic patterns) patterns

- Hearst (1992a)
- Earliest and **still common** algorithm for (a hyponym of) relation extraction

Agar is a substance prepared from a mixture of red algae, such as ***Gelidium****, for laboratory or industrial use.*



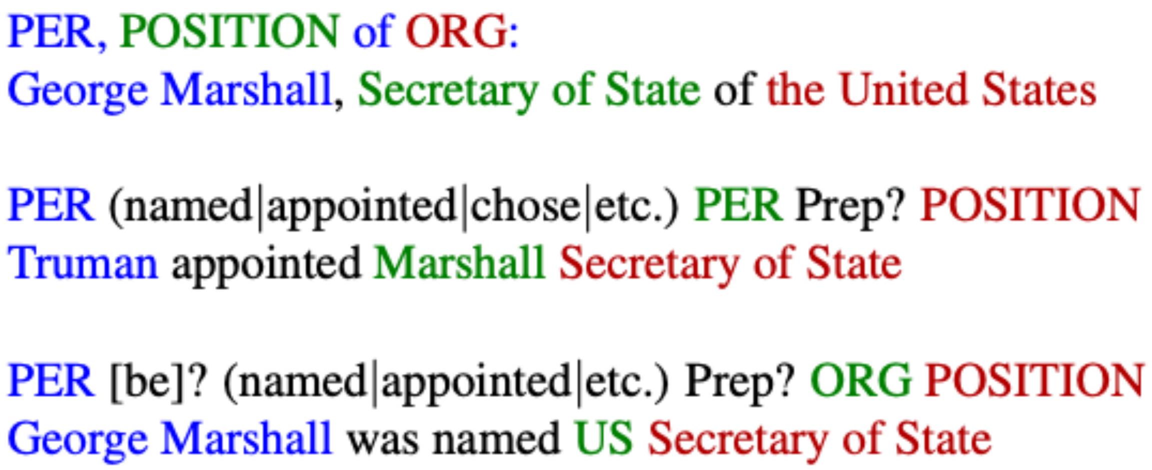
NP {, NP}* {,} (and|or) other NP$_H$	temples, treasuries, and other important civic buildings
NP$_H$ such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP$_H$ as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP$_H$ {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP$_H$ {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

https://github.com/abyssnlp/Hearst-Hypernym-Extractor

Handwritten patterns

Modern pattern-based approach
Hearst (1992a) + named entity constraints

Who holds what office in which organization?



PER, POSITION of ORG:
George Marshall, Secretary of State of the United States

PER (named|appointed|chose|etc.) PER Prep? POSITION
Truman appointed Marshall Secretary of State

PER [be]? (named|appointed|etc.) Prep? ORG POSITION
George Marshall was named US Secretary of State

Handwritten patterns

- High-precision
- Can be tailored to specific domains

- Often, low-recall
- Lot of work to create them for all possible patterns.

Supervised Learning

- Choose fixed set of relations and entities
- Annotate training corpus with relations and entities
- Use annotated texts to train classifiers to annotate unseen test set

Straightforward approach:

1. Find pairs of named entities (usually in the same sentence)
2. Apply relation classification on each pair. Classifier can use any supervised technique (logistic regression, RNN, Transformer, etc.)

Intermediate filtering classifier (optional):

- Makes binary decision on whether a given pair of named entities are related (by any relation)
- Trained on positive examples from all relations in annotated corpus
- Trained on negative examples from within-sentence entity pairs not annotated with a relation

Supervised Learning

Feature-based supervised relation classifiers

- Use hand-crafted features to represent relation between two entities (M1 and M2)
- Train classifier using annotated data to predict relation between two entities
- Can use any supervised learning algorithm (e.g. logistic regression, random forests)

Features:

- Word features:
  - Headwords of M1 and M2 and their concatenation
  - Bag-of-words and bigrams **in** M1 and M2
  - Words or bigrams in particular positions
  - Bag-of-words or bigrams **between** M1 and M2
- Named entity features:
  - Named-entity types and their concatenation
  - Entity level of M1 and M2 (from the set NAME, NOMINAL, PRONOUN)
  - Number of entities between M1 and M2
- Syntactic structure:
  - Constituent paths between M1 and M2
  - Dependency-tree paths

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said

Airlines Wagner Airlines-Wagner

American, Airlines, Tim, Wagner, American Airlines, Tim Wagner

M2: -1 spokesman M2: +1 said

a, AMR, of, immediately, matched, move, spokesman, the, unit

M1: ORG, M2: PER, M1M2: ORG-PER

M1: NAME [it or he would be PRONOUN] M2: NAME [the company would be NOMINAL]

1 FOR AMR

NP↑NP↑S↑S↓NP

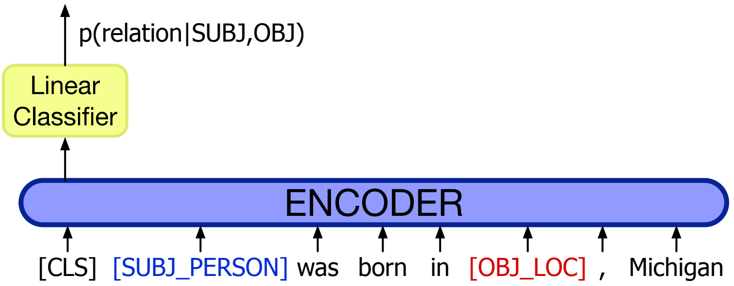
Airlines $\leftarrow_{subj}$ *matched* $\leftarrow_{comp}$ *said* $\rightarrow_{subj}$ *Wagner*

Supervised Learning

Neural supervised relation classifiers

Treat relation extraction as supervised classification

- **Input:** a sentence and two spans (subject and object)
- **Output:** one of 43 labels (42 TAC relations or no relation)
- Method:
  - Use pretrained encoder (e.g. BERT)
  - Add linear layer on top of sentence representation
  - Finetune linear layer as 1-of-N classifier
- Input to encoder partially de-lexified (replace subject and object entities with NER tags)
- Can use versions of BERT designed for single long sequence of sentences (RoBERTa, SPANbert)



Zhang et al. 2017, Joshi et al. 2020

Supervised Learning

Pros:

- High accuracy (if enough labeled data and similarity to test set)

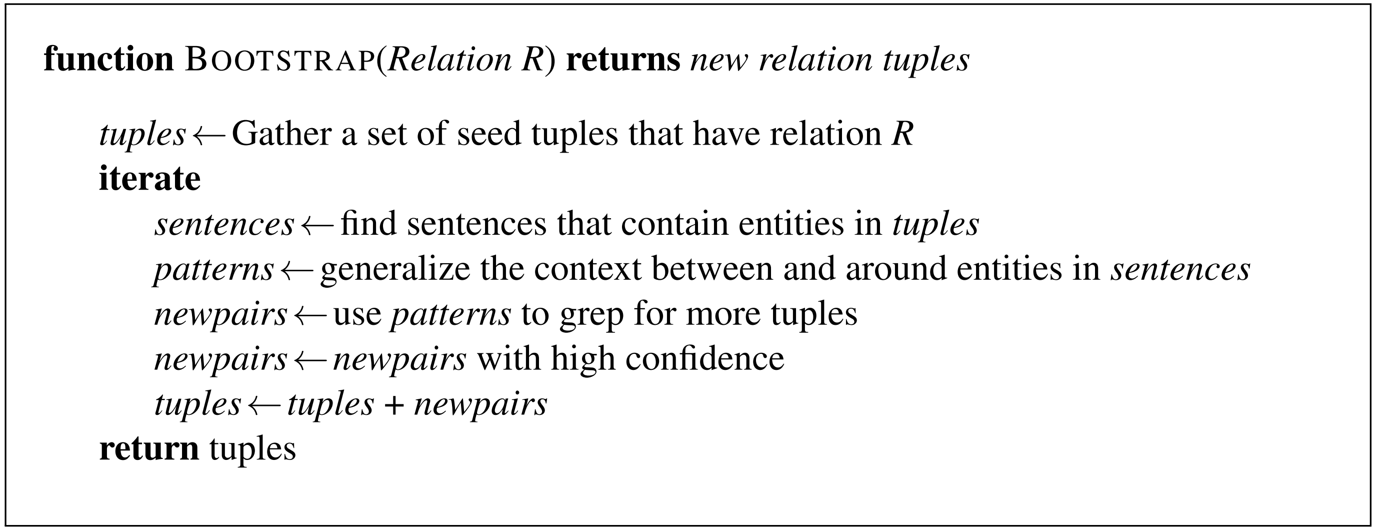
Cons:

- Expensive (labeling large training set)
- Brittle (poor generalization to different genres)

Semi-supervised (via Bootstrapping)

Does not require lots of labeled data, instead; seed patterns, or perhaps a few seed tuples

Bootstrapping takes the entities in the seed pair, and then finds sentences (on the web, or whatever dataset we are using) that contain both entities.



```
function BOOTSTRAP(Relation R) returns new relation tuples

    tuples ← Gather a set of seed tuples that have relation R
    iterate
        sentences ← find sentences that contain entities in tuples
        patterns ← generalize the context between and around entities in sentences
        newpairs ← use patterns to grep for more tuples
        newpairs ← newpairs with high confidence
        tuples ← tuples + newpairs
    return tuples
```

Semi-supervised (via Bootstrapping)

Suppose, for example, that we need to create a list of airline/hub pairs, and we know only that *Ryanair has a hub at Charleroi*.

Budget airline ***Ryanair****, which uses* ***Charleroi*** *as a* ***hub****, scrapped all weekend flights out of the airport.*	/ [ORG], which uses [LOC] as a hub /
All flights in and out of ***Ryanair's hub*** *at* ***Charleroi*** *airport were grounded on Friday…*	/ [ORG]'s hub at [LOC] /
A spokesman at ***Charleroi****, a main* ***hub*** *for* ***Ryanair****, estimated that 8000 passengers had already been affected.*	/ [LOC], a main hub for [ORG] /

Semi-supervised (via Bootstrapping)

Pros:

- Does not require labeled data

Cons:

- Semantic drift.

An erroneous pattern leads to the introduction of erroneous tuples, which, in turn, lead to the creation of problematic patterns and the meaning of the extracted relations 'drifts'.

Sydney has a ferry hub at Circular Quay.

If accepted as a positive example, this expression could lead to the incorrect introduction of the tuple ⟨Sydney,CircularQuay⟩. Patterns based on this tuple could propagate further errors into the database.

Semi-supervised (distant supervision)

Combines the advantages of **bootstrapping** with **supervised learning**.

Instead of just a handful of seeds, distant supervision uses a large database to acquire a huge number of seed examples, creates lots of noisy pattern features from all these examples and then combines them in a supervised classifier.

Example: learn the **place-of-birth relationship** between **people** and their **birth cities.**

- DBPedia has thousands of examples of many relations; including over 100,000 examples of place-of-birth

<Edwin **Hubble**, **Marshfield**> <Albert **Einstein**, **Ulm**>

- Run named entity taggers on large amounts of text (Mintz et al. (2009) used 800,000 articles from Wikipedia) and extract all sentences that have two named entities that match the tuple

…**Hubble** was born in **Marshfield**…
…**Einstein**, born (1879), **Ulm**…
…**Hubble’s** birthplace in **Marshfield**…

- Training instances can now be extracted from this data, one training instance for each identical tuple <relation, entity1, entity2>.

…**Hubble** was born in **Marshfield**…	born-in
…**Einstein**, born (1879), **Ulm**…	born-year
…**Hubble’s** birthplace in **Marshfield**…	born-in

Semi-supervised (distant supervision)

Pros:

- Unlike the iterative expansion, there's **no semantic drift** due to the use a large number of features simultaneously
- Like unsupervised classification, **it doesn't use a labeled training corpus of texts**, so it isn't sensitive to genre issues in the training corpus, and relies on very large amounts of unlabeled data.
- It can create training tuples to be used with neural classifiers, where features are not required

Cons:

- it tends to produce **low-precision results** (current research focuses on ways to improve precision)
- it can only help in extracting relations for which a large enough database already exists. To extract new relations without datasets, or relations for new domains, purely unsupervised methods must be used.

Unsupervised

- **Goal:** extract relations from the web
- No labeled training data
- No list of relations
- Also known as Open information extraction (Open IE)

ReVerb system (Fader et al., 2011):

1. Run a part-of-speech tagger and entity chunker over *S*
2. For **each verb in *S***, find the **longest sequence of words *W* that start with a verb** and satisfy syntactic and lexical constraints, merging adjacent matches.
3. For **each phrase *W***, find the **nearest noun phrase x to the left** which is not a relative pronoun, wh-word or existential "there". Find the **nearest noun phrase y to the right**.
4. Assign confidence **c** to the relation r = (x, w, y) using a confidence classifier and return it.

United has a hub in Chicago, which is the headquarters of United Continental Holdings.

United **has** a hub in Chicago, which **is** the headquarters of United Continental Holdings.

r1 = United, has a hub in, Chicago
r2 = Chicago, is the headquarters of, United Continental Holdings

Unsupervised

Pros:

- Ability to handle large number of relations without specification

Cons:

- need to map large sets of strings to canonical form

IE Evaluation

- Supervised models: test set with human-annotated, gold-standard relations; precision, recall, F-measure
- Semi-supervised, unsupervised: difficult to evaluate, not possible to pre-annotate gold set of correct instances of relations
    - Approximate precision by drawing random sample of extracted relations; human checks accuracy
        - Evaluation focuses on tuples (sets of extracted relations) rather than relation mentions
        - Estimated precision: # correctly extracted relation tuples in the sample / total # extracted relation tuples in sample
    - Precision at different levels of recall
        - System ranks relations by probability or confidence
        - Precision computed for top 1000, 10,000, 100,000, etc. new relations
        - Random sample at each level
        - Limitation: no way to directly evaluate recall