Lecture 24:
Information Extraction

WordNet as a database of relations between *concepts*

Hyponym relations (is-a relation)

cats are mammals

Meryonym relations (part-of/has-a relations):

Part meronyms: bumpers are parts of cars,
cars have bumpers

Member meronyms: musicians belong to bands/orchestras,

Substance meronyms: dough contains flour

NB: some of these are inherited via hypernyms:
'musician' is a member meronym of 'musical organization',
which has hyponyms such as 'orchestra', 'band', 'choir', etc.

Domain knowledge expressed as relations

Wikipedia's **infoboxes** provide structured facts about **named entities**:

These can be turned into **structured relations** between these entities, e.g.

location-of(UIUC, Illinois)

or **RDF** (Resource Description Framework) **triples**

```
(entity, relation, entity):
(UIUC, location, Illinois)
```

Freebase and **DBPedia** (2 billion RDF triples) are both very large knowledge bases of such relations, extracted from Wikipedia.

University of Illinois at Urbana–Champaign

Former names	Illinois Industrial University (1867–1885) University of Illinois (1885–1982)
Motto	*Learning & Labor*
Type	Public land-grant research university
Established	1867; 153 years ago
Academic affiliations	University of Illinois system AAU BTAA APLU URA Sea-grant Space-grant
Endowment	$2.35 billion (2018)[1]
Chancellor	Robert J. Jones[2]
Provost	Andreas C. Cangellaris[3]
Academic staff	2,548
Administrative staff	7,901
Students	51,196 (Fall 2019)[4]
Undergraduates	33,850 (Fall 2019)[4]
Postgraduates	16,319 (Fall 2019)[4]
Location	Urbana and Champaign, Illinois, United States
Campus	Urban, 8,370 acres (2,578 ha)[5]

https://en.wikipedia.org/wiki/University_of_Illinois_at_Urbana–Champaign

Relation Extraction from text

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**].

Can we identify that…

…American Airlines is part of (a unit of) AMR,

…United Airlines is part of (a unit of) UAL Corp,

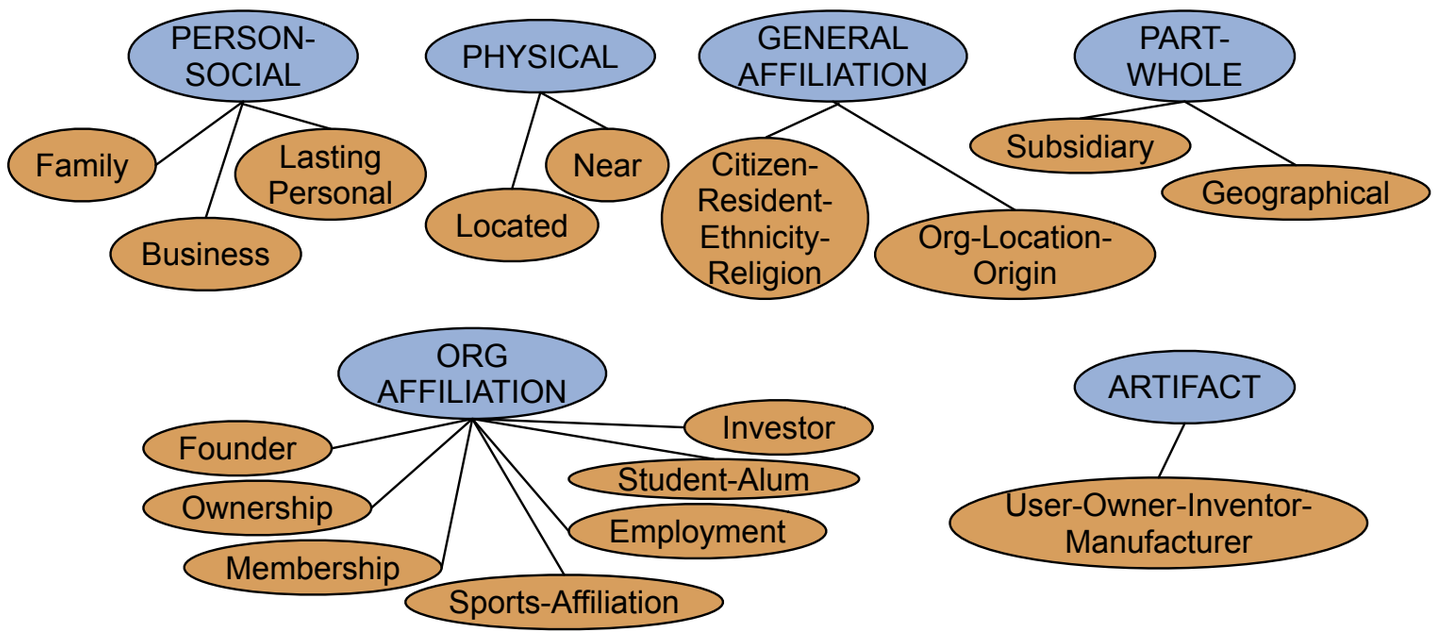
…Tim Wagner is employed by (a spokesman of) AMR

Relation Extraction from text

Identify **relations between named entities**, typically from a small set of predefined relations.

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

The 17 relations (orange) used in ACE:



A logical interpretation

We can construct a model for these relations:

- The **domain** (universe) is a **set of named entities,** partitioned into different types or classes of entities
- Each **relation** is a **set of tuples** of entities (restricted to relation-specific tuples of types)

Domain	$\mathscr{D} = \{a, b, c, d, e, f, g, h, i\}$
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	$Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$

Rule-based relation extraction

Handwritten rules to identify **lexico-syntactic patterns** (Hearst, 1992) can be used for high-precision (and low-recall) relation extraction:

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

The **pattern** "X, such as Y (and/or Z)" implies that X is a hypernym of Y and Z.

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

Figure 18.12 Hand-built lexico-syntactic patterns for finding hypernyms, using {} to mark optionality (Hearst 1992a, Hearst 1998).

Relation Extraction via supervised learning

Learn a classifier that identifies whether there is a relation between a pair of entities that appear in the same sentence (or nearby within a document).

Classifier output: *n*+1 classes for *n* rels (incl. NONE)

Useful features:

— the words appearing in and next to the entities

— the words between the entities

— the NER types of both entities

— the distance between both entities (#words, #NERs,…)

— the syntactic path between the entities

Semi-supervised Relation Extraction

Use **high-precision seed patterns** (e.g. "X's Y") relations to identify **high-confidence seed tuples**.

Ryanair's hub Charleroi `-> (Ryanair, has-hub-in, Charleroi)`

Bootstrap a classifier with increasing coverage:

— Find sentences containing entity pairs from seeds.

"Ryanair, which uses Charleroi as hub"
"Ryanair's Belgian hub at Charleroi"

— These will contain new patterns

(as well as some noise: "Sydney has a ferry hub at Circular Quay")

— Noise needs to be controlled so as not to propagate

(Confidence values, combined across patterns via noisy-or)

Distant Supervision for Relation Extraction

— Use **a very large database of known relations** (Freebase, DBPedia) to obtain a very large number of seed tuples.

```
(John F. Kennedy, died-in, Dallas)
(Princess Diana, died-in, Paris)
(Elvis Presley, died-in, Memphis)
```

— Search **large amounts of text** for sentences containing pairs of entities in a known relation

(plus entities in this list not in any known relation, to get no-relation examples)

— Process these sentences with NER, syntactic parsing, etc.

— **Learn a classifier** on these sentences to predict relations between entities that are not in the database

What is the intuition why this might work?

This returns a lot of noise: Elvis performed/lived/is buried in/sang about/… Memphis

But if trained on enough data, high-confidence predictions of this classifier are likely to be correct (since many true positive examples will be similar to each other)

Unsupervised Relation Extraction ("Open Information Extraction/IE")

Goal: Extract any relation (from large amounts of text, e.g. web) without being restricted to a predefined set of relations

Relations: Raw strings of words (often beginning with verbs, and possibly subject to some predefined syntactic constraints)

Example: The ReVerb algorithm:

— Run a POS tagger and entity chunker over each sentence

— Identify any potential relations (any string between entities that starts with a verb and obeys predefined constraints)

— Normalize relations (remove inflection, auxiliary verbs, adjectives, adverbs)

— Add relations that occur with at least N different arguments to database

— Train a classifier on small number (1000) hand-labeled sentences to obtain confidence scores for relations in the database.