Information Extraction

COMP90042

Natural Language Processing

Lecture 18

Semester 1 2022 Week 9

Jey Han Lau

THE UNIVERSITY OF MELBOURNE

Information Extraction

- Given this:
  - *"Brasilia, the Brazilian capital, was founded in 1960."*
- Obtain this:
  - capital(Brazil, Brasilia)
  - founded(Brasilia, 1960)
- Main goal: turn **text** into **structured data**

Applications

- **Stock analysis**
  - Gather information from news and social media
  - Summarise texts into a structured format
  - Decide whether to buy/sell at current stock price
- **Medical research**
  - Obtain information from articles about diseases and treatments
  - Decide which treatment to apply for new patient

How?

- Given this:
  - *"Brasilia, the Brazilian capital, was founded in 1960."*
- Obtain this:
  - capital(Brazil, Brasilia)
  - founded(Brasilia, 1960)
- Two steps:
  - **Named Entity Recognition (NER)**: find out entities such as "Brasilia" and "1960"
  - **Relation Extraction**: use context to find the relation between "Brasilia" and "1960" ("founded")

Machine learning in IE

- Named Entity Recognition (NER): **sequence** models such as RNNs, HMMs or CRFs.
- Relation Extraction: mostly **classifiers**, either binary or multi-class.
- This lecture: how to frame these two tasks in order to apply sequence labellers and classifiers.

Outline

- Named Entity Recognition
- Relation Extraction
- Other IE Tasks

Named Entity Recognition

Named Entity Recognition

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.

JM3, Ch 17

Named Entity Recognition

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 [GPE **Chicago**] to [GPE **Dallas**] and [GPE **Denver**] to [GPE **San Francisco**]

Typical Entity Tags

- **PER**: people, characters
- **ORG**: companies, sports teams
- **LOC**: regions, mountains, seas
- **GPE**: countries, states, provinces
(in some tagset this is labelled as **LOC**)
- **FAC**: bridges, buildings, airports
- **VEH**: planes, trains, cars
- Tag-set is application-dependent: some domains deal with specific entities e.g. proteins and genes

NER as Sequence Labelling

- NE tags can be ambiguous:
  - "Washington" can be a person, location or political entity
- Similar problem when doing POS tagging
  - Incorporate context
- Can we use a sequence tagger for this (e.g. HMM)?
  - No, as entities can span multiple tokens
  - Solution: modify the tag set

IO tagging

- **[**ORG **American Airlines],** a unit of **[**ORG **AMR Corp.]**, immediately matched the move, spokesman **[**PER **Tim Wagner]** said.
- 'I-ORG' represents a token that is **inside** an entity (ORG in this case).
- All tokens which are not entities get the 'O' token (for **outside**).
- Cannot differentiate between:
  - a single entity with multiple tokens
  - multiple entities with single tokens

Words	IO Label
American	I-ORG
Airlines	I-ORG
,	O
a	O
unit	O
of	O
AMR	I-ORG
Corp.	I-ORG
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	I-PER
Wagner	I-PER
said	O
.	O

IOB tagging

- [ORG **American Airlines]**, a unit of **[**ORG **AMR Corp.]**, immediately matched the move, spokesman **[**PER **Tim Wagner]** said.
- B-ORG represents the *beginning* of an ORG entity.
- If the entity has more than one token, subsequent tags are represented as I-ORG.

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

Annotate the following sentence with NER tags (IOB)

Steve Jobs founded Apple Inc. in 1976

Tagset: PER, ORG, LOC, TIME

PollEv.com/jeyhanlau569

NER as Sequence Labelling

- Given such tagging scheme, we can train any sequence labelling model
- In theory, HMMs can be used but **discriminative** models such as CRFs are preferred

NER: Features

- Example: *L'Occitane*
- Prefix/suffix:
  - L / L' / L'O / L'Oc / …
  - e / ne / ane / tane / …
- Word shape:
  - X'Xxxxxxxx / X'Xx
  - XXXX-XX-XX (date!)
- POS tags / syntactic chunks: many entities are nouns or noun phrases.
- Presence in a **gazeteer**: lists of entities, such as place names, people's names and surnames, etc.

Word	POS	Chunk	Short shape	Label
American	NNP	B-NP	Xx	B-ORG
Airlines	NNPS	I-NP	Xx	I-ORG
,	,	O	,	O
a	DT	B-NP	x	O
unit	NN	I-NP	x	O
of	IN	B-PP	x	O
AMR	NNP	B-NP	X	B-ORG
Corp.	NNP	I-NP	Xx.	I-ORG
,	,	O	,	O
immediately	RB	B-ADVP	x	O
matched	VBD	B-VP	x	O
the	DT	B-NP	x	O
move	NN	I-NP	x	O
,	,	O	,	O
spokesman	NN	B-NP	x	O
Tim	NNP	I-NP	Xx	B-PER
Wagner	NNP	I-NP	Xx	I-PER
said	VBD	B-VP	x	O
.	,	O	.	O

Figure 18.6 Word-by-word feature encoding for NER.

NER: Classifier

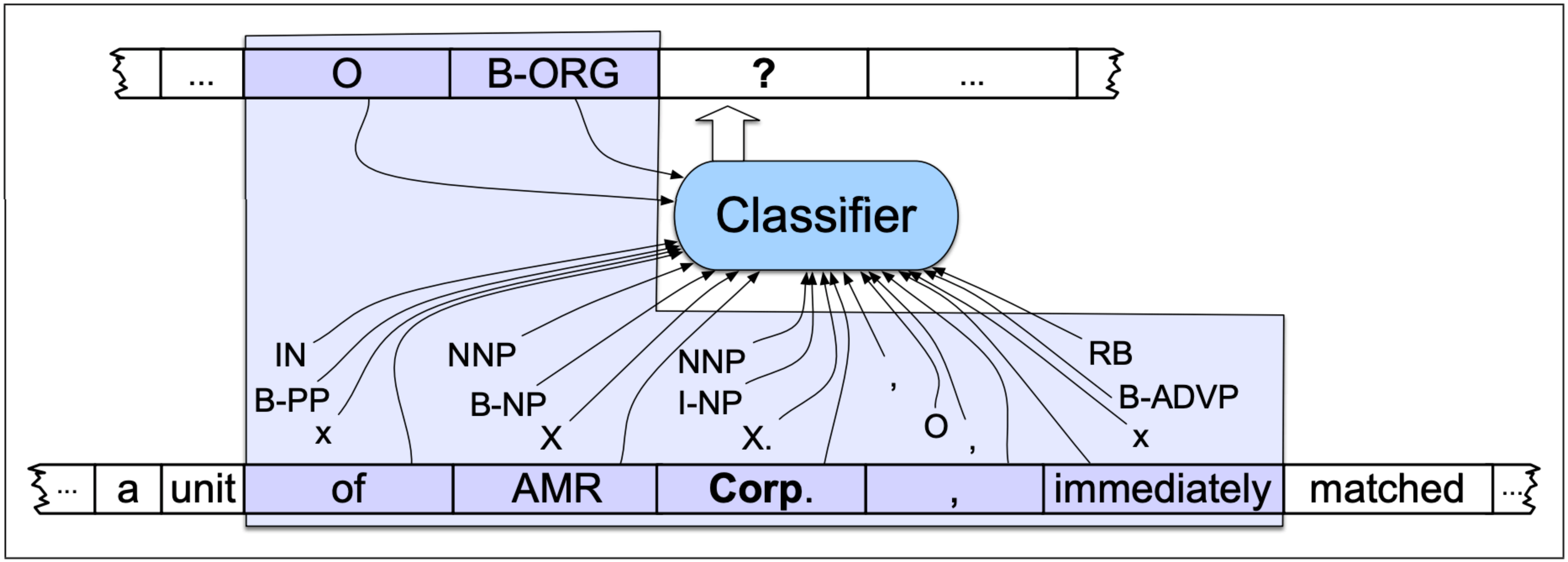
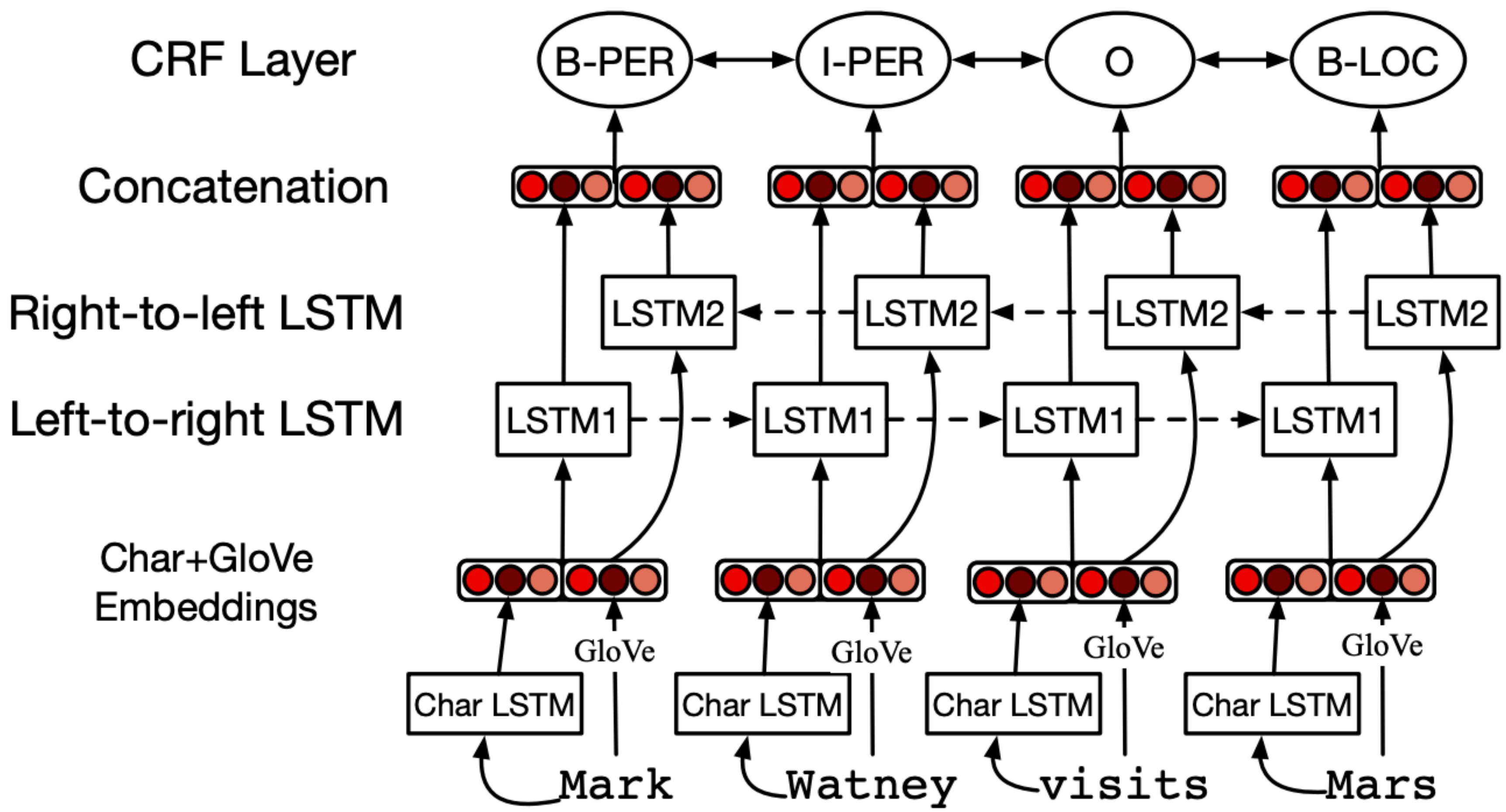


Figure 18.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Deep Learning for NER

- State of the art approach uses LSTMs with character and word embeddings (Lample et al. 2016)



Relation Extraction

Relation Extraction

- [ORG **American Airlines],** a unit of **[**ORG **AMR Corp.]**, immediately matched the move, spokesman **[**PER **Tim Wagner]** said.
- Traditionally framed as triple extraction:
  - unit(American Airlines, AMR Corp.)
  - spokesman(Tim Wagner, American Airlines)
- Key question: do we know all the possible relations?

Relation Extraction

- unit(American Airlines, AMR Corp.) → subsidiary
- spokesman(Tim Wagner, American Airlines) → employment

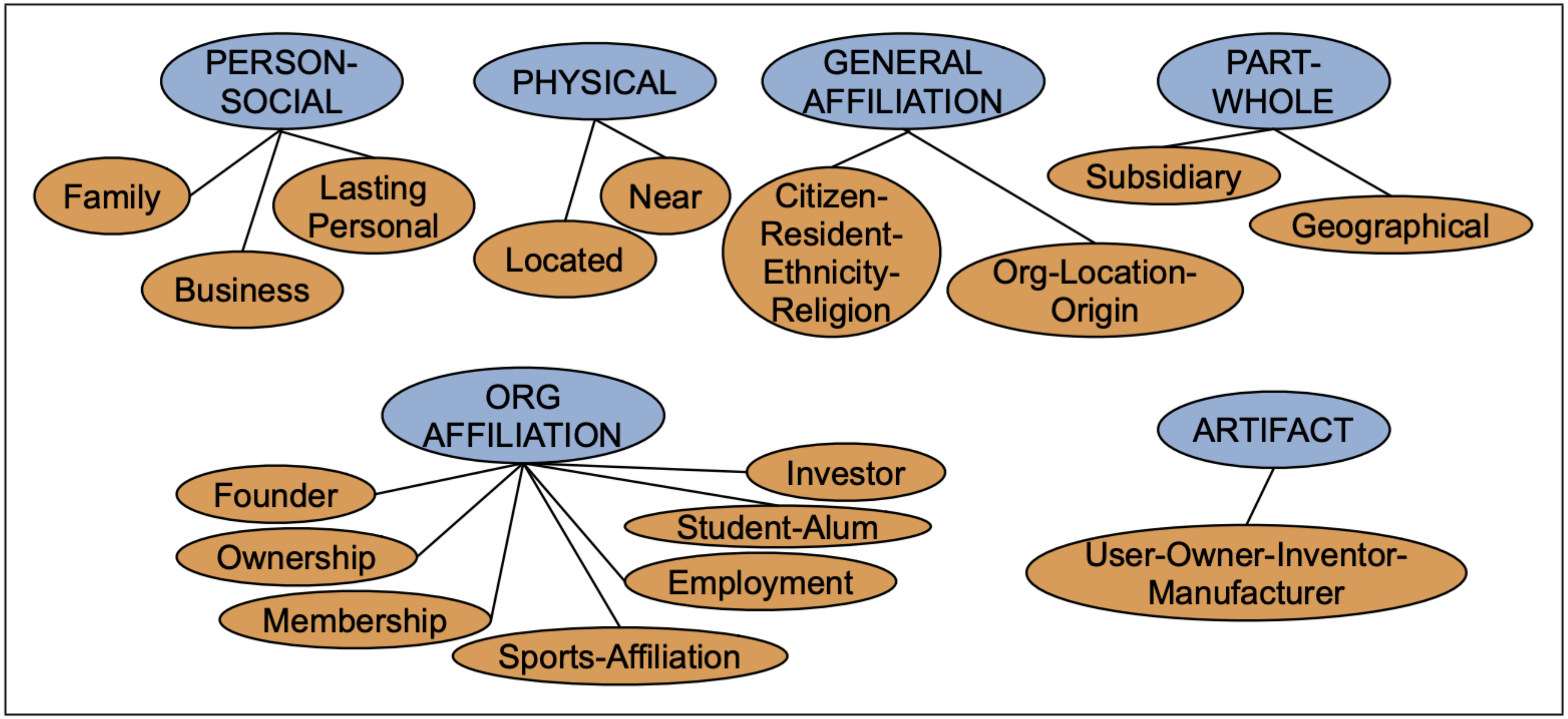


Figure 18.9 The 17 relations used in the ACE relation extraction task.

Methods

- If we have access to a fixed relation database:
  - Rule-based
  - Supervised
  - Semi-supervised
  - Distant supervision
- If no restrictions on relations:
  - Unsupervised
  - Sometimes referred as “OpenIE”

Rule-Based Relation Extraction

- “Agar is a substance prepared from a mixture of red algae such as Gelidium, for laboratory or industrial use.”
- **[NP red algae]** such as **[NP Gelidium]**
- NP_0 such as NP_1 → hyponym(NP_1, NP_0)

  下位词

  属于
- hyponym(Gelidium, red algae)
- Lexico-syntactic patterns: high precision, low recall, manual effort required

More Rules

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).

Supervised Relation Extraction

- Assume a corpus with annotated relations
- Two steps. ① First, find if an entity pair is related or not (binary classification)
  - For each sentence, gather all possible entity pairs
  - Annotated pairs are considered positive examples
  - Non-annotated pairs are taken as negative examples
- ② Second, for pairs predicted as positive, use a multi-class classifier (e.g. SVM) to obtain the relation

Supervised Relation Extraction

- [ORG **American Airlines]**, a unit of **[**ORG **AMR Corp.]**, immediately matched the move, spokesman **[**PER **Tim Wagner]** said.
- First:
  - (American Airlines, AMR Corp.) → positive
  - (American Airlines, Tim Wagner) → positive
  - (AMR Corp., Tim Wagner) → negative
- Second:
  - (American Airlines, AMR Corp.) → subsidiary
  - (American Airlines, Tim Wagner) → employment

Features

- [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said.
- (American Airlines, Tim Wagner) → employment

M1 headword	*airlines* (as a word token or an embedding)
M2 headword	*Wagner*
Word(s) before M1	NONE
Word(s) after M2	*said*
Bag of words between	{*a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
M1 type	ORG
M2 type	PERS
Concatenated types	ORG-PERS
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Semi-supervised Relation Extraction

- Annotated corpora is very expensive to create
- Use seed tuples to bootstrap a classifier

Semi-supervised Relation Extraction

1. Given seed tuple: hub(Ryanair, Charleroi)
2. Find sentences containing terms in seed tuples
    - *Budget airline **Ryanair**, which uses **Charleroi** as a hub, scrapped all weekend flights out of the airport.*
3. Extract general patterns
    - [ORG], which uses [LOC] as a hub
4. Find new tuples with these patterns
    - hub(Jetstar, Avalon)
5. Add these new tuples to existing tuples and repeat step 2

What are some issues of such semi-supervised relation extraction method?

- Difficult to create seed tuples
- Extracted tuples deviate from original relation over time
- Difficult to evaluate
- Tend not to find many novel tuples given seed tuples
- Extracted general patterns tend to be very noisy

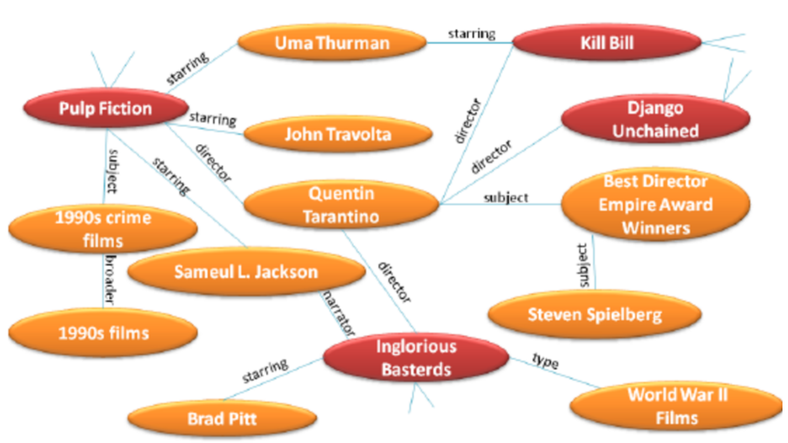
PollEv.com/jeyhanlau569

Semantic Drift

- Pattern: [NP] has a {NP}* hub at [LOC]
- *Sydney has a ferry hub at Circular Quay*
  - hub(Sydney, Circular Quay)
- More erroneous patterns extracted from this tuple…
- Should only accept patterns with high confidences

Distant Supervision

- Semi-supervised methods assume the existence of seed tuples to mine new tuples
- Can we mine new tuples **directly**?
- Distant supervision obtain new tuples from a range of sources:
  - DBpedia
  - Freebase



- Generate massive training sets, enabling the use of richer features, and no risk of semantic drift

Unsupervised Relation Extraction ("OpenIE")

- No fixed or closed set of relations
- Relations are sub-sentences; usually has a verb
- "United has a hub in Chicago, which is the headquarters of United Continental Holdings."
  - "has a hub in"(United, Chicago)
  - "is the headquarters of"(Chicago, United Continental Holdings)
- Main problem: mapping relations into canonical forms

Evaluation

- NER: F1-measure at the **entity** level.
- Relation Extraction with known relation set: F1-measure
- Relation Extraction with unknown relations: much harder to evaluate
  - Usually need some human evaluation
  - Massive datasets used in these settings are impractical to evaluate manually (use samples)
  - Can only obtain (approximate) precision, not recall.

Other IE Tasks

Temporal Expression Extraction

"**[**TIME **July 2, 2007]**: A fare increase initiated **[**TIME **last week]** by UAL Corp's United Airlines was matched by competitors over **[**TIME **the weekend]**, marking the second successful fare increase in **[**TIME **two weeks]**."

- **Anchoring**: when is "last week"?
  - "last week" → 2007–W26
- **Normalisation**: mapping expressions to canonical forms.
  - July 2, 2007 → 2007-07-02
- Mostly rule-based approaches

Event Extraction

- "American Airlines, a unit of AMR Corp., immediately **[**EVENT **matched] [**EVENT **the move]**, spokesman Tim Wagner **[**EVENT **said]**."
- Very similar to NER, including annotation and learning methods.
- **Event ordering**: detect how a set of events happened in a timeline.
  - Involves both event extraction and temporal expression extraction.

A Final Word

- Information Extraction is a vast field with many different tasks and applications
  - Named Entity Recognition
  - Relation Extraction
  - Event Extraction
- Machine learning methods involve classifiers and sequence labelling models.

Reading

- JM3 Ch. 8.3, 17-17.2
- References:
  - Lample et al, Neural Architectures for Named Entity Recognition, NAACL 2016 https://github.com/glample/tagger