

Information Extraction

COMP90042

Natural Language Processing

Lecture 18



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** such as databases, etc.
- Help **decision makers** in applications.

Examples

- Stock analysis
 - ▶ Gather information from news and social media → summarise into a structured format → decide whether to buy/sell at current stock price
- Medical and biological research
 - ▶ Obtain information from articles about diseases and treatments
 - decide which treatment to apply for a new patient
- Rumour detection
 - ▶ Detect events in social media
 - decide where, when and how to act

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”)

Extract
Entities

Machine learning in IE

Sequence labeling

- 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 classifiers and sequence labellers.
- Choice of machine learning methods is up to the user (yes, deep learning methods can be applied).

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.

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 [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco]

Typical Entity Tags

- **PER**: people, characters
- **ORG**: companies, sports teams
- **LOC**: regions, mountains, seas
- **GPE**: countries, states, provinces (sometimes conflated with **LOC**)
- **FAC**: bridges, buildings, airports
- **VEH**: planes, trains, cars
- Tag-set is application-dependent: some domains deal with specific entities e.g. proteins, genes or works of art.

NER as Sequence Labelling

- NE tags can be ambiguous:
 - ▶ “Washington” can be either a person, a location or a political entity.
- We faced a similar problem when doing POS tagging.
 - ▶ Solution: incorporate context by treating NER as sequence labelling.
- Can we use an out-of-the-box sequence tagger for this (e.g., HMM)?
 - ▶ Not really: entities can span multiple tokens.
 - ▶ Solution: adapt the tag set.

Sequence labelling.

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*).
- Can not differentiate between a single entity with multiple tokens or multiple entities with single tokens.

one entity
two words!

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

two
orgs

can not differentiate

Solved by

IOB tagging

B-ORG

- **[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

NER as Sequence Labelling

- Given a tagging scheme and an annotated corpus, one can train any sequence labelling model
- In theory, HMMs can be used but *discriminative models* such as MEMMs and CRFs are preferred
 - ▶ Character-level features (is the first letter uppercase?)
 - ▶ Extra resources, e.g., lists of names
 - ▶ POS tags

add

add
features
more
easily.

NER: Features

- Character and word shape features (ex: “L’Occitane”)
 - Prefix/suffix:
 - ▶ L / L’ / L’O / L’Oc / ...
 - ▶ e / ne / ane / tane / ...
 - Word shape:
 - ▶ X’Xxxxxxxxx / 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.
- a list of entities
common people
surnames. ...*

feature

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

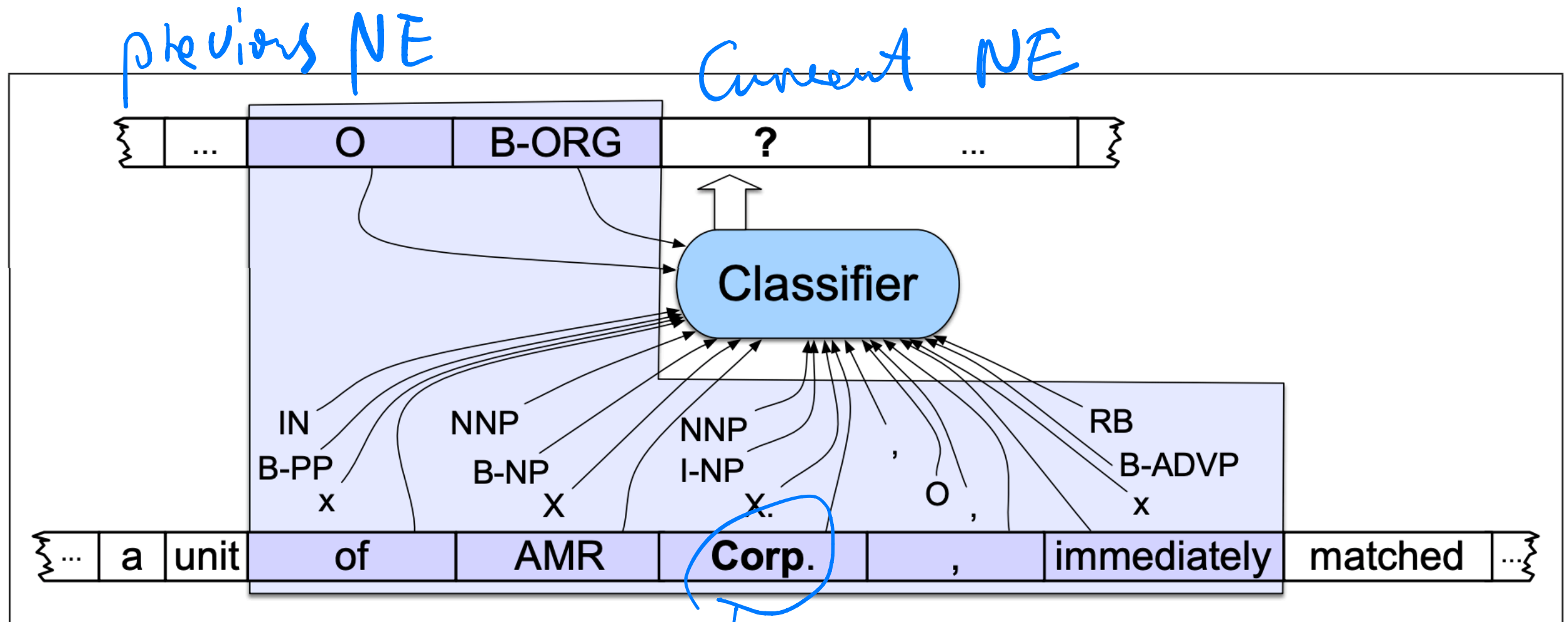
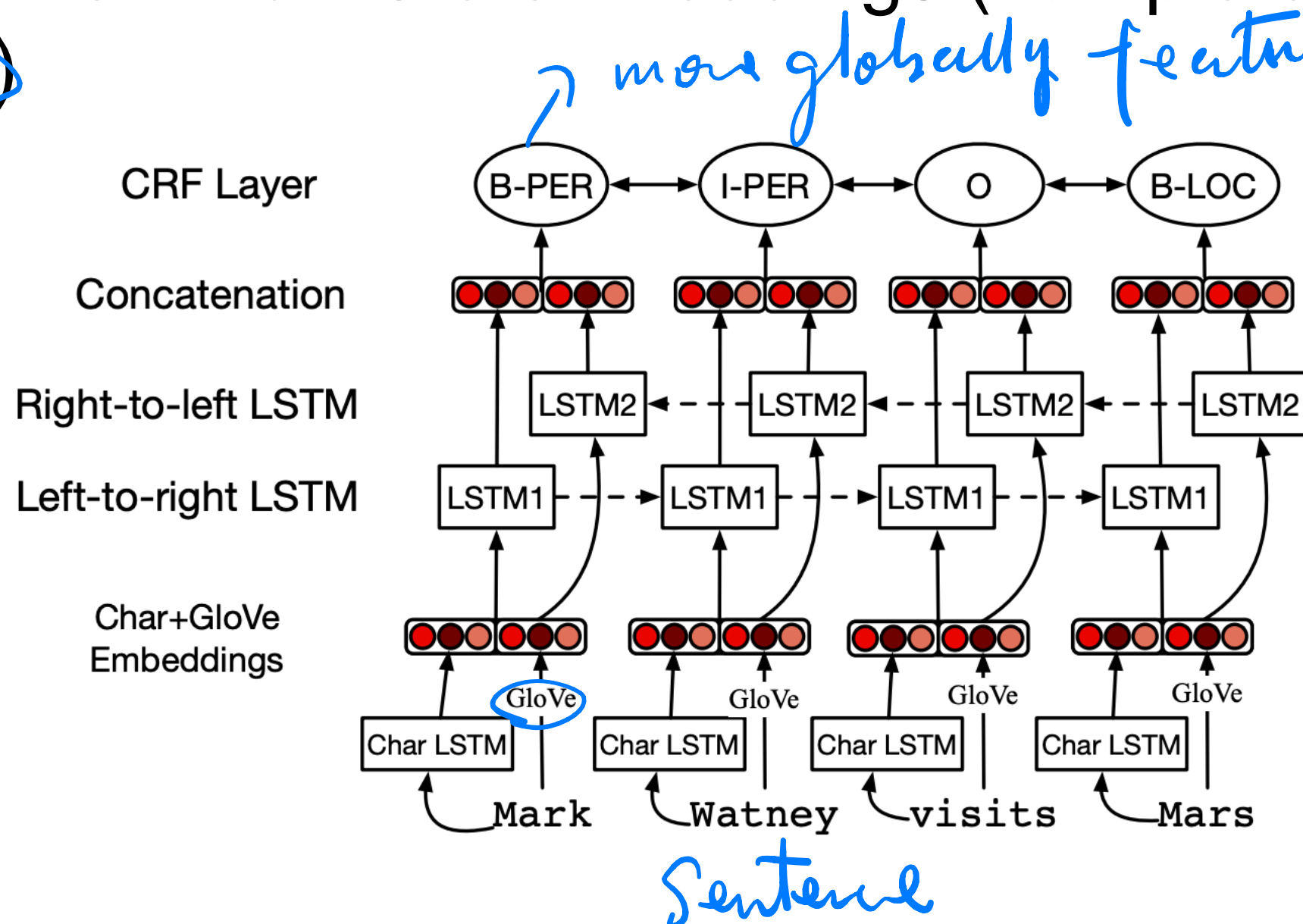


Figure 18.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

target word.

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: *3 arguments*
 - ▶ unit(American Airlines, AMR Corp.)
 - ▶ spokesman(Tim Wagner, American Airlines)
- Key question: do we have access to a set of possible relations?
 - ▶ Answer depends on the application

if you can have limited relations there can be enumerated.

Relation Extraction

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

Mapping.

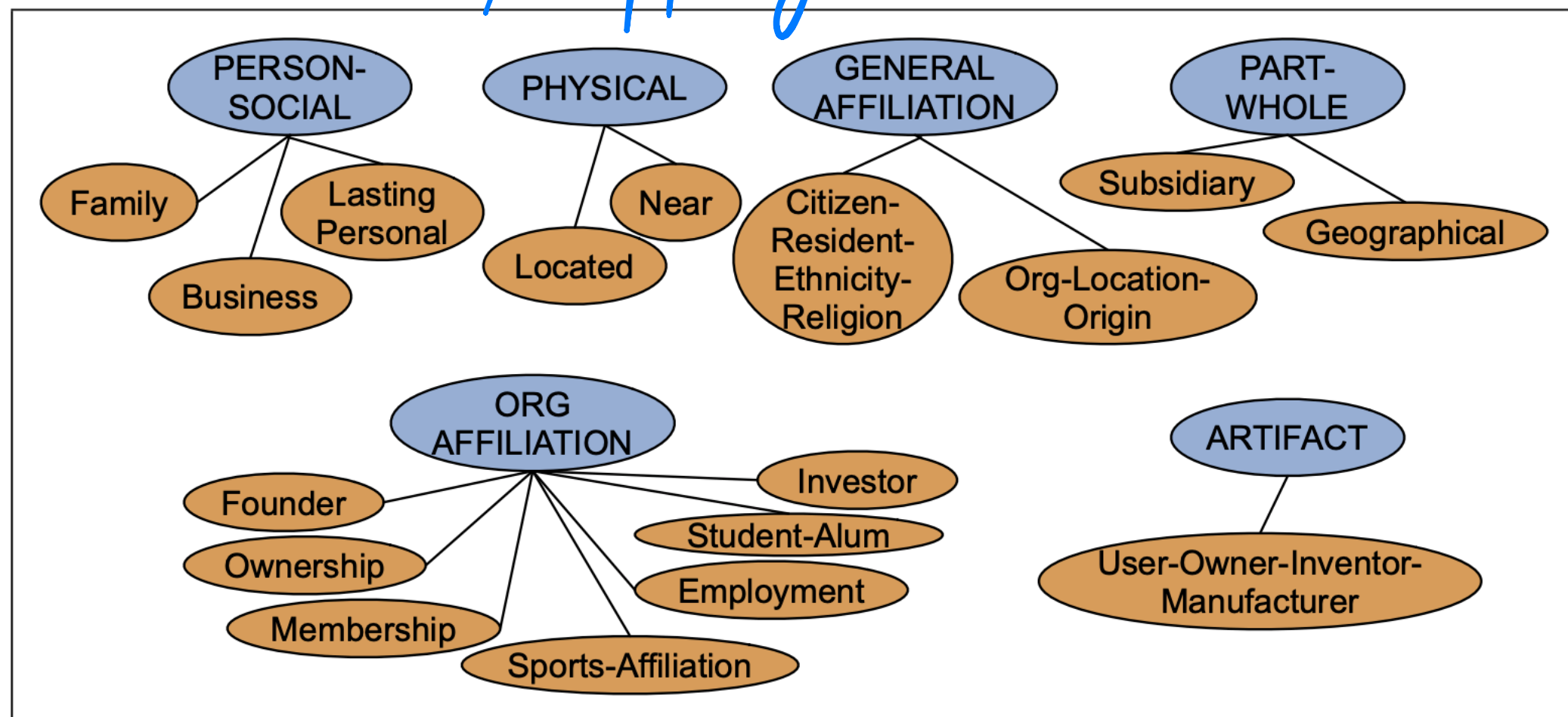


Figure 18.9 The 17 relations used in the ACE relation extraction task.

Methods

- If we have access to a fixed relation database:
 - ▶ Rule-based
 - ▶ Supervised
 - ▶ Semi-supervised
 - ▶ Distant supervision
- If no restrictions on relations: *Open 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 $NP_1 \rightarrow$ hyponym(NP_1, NP_0)
- $\text{hyponym}(\text{Gelidium}, \text{red algae})$
- Lexico-syntactic patterns: high precision, low recall, manual effort required

↓ Rules cannot capture all ways for relations.

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. First, find if an entity pair is related or not (binary classification) *→ has relation or Not.*
 - ▶ For each sentence, gather all possible entity pairs
 - ▶ Annotated pairs are considered positive examples
 - ▶ Non-annotated pairs are taken as negative examples
- 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

label

predict.

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


chunk parser .

Semi-supervised Relation Extraction

- Annotated corpora is very expensive to create.
- Use seed tuples to bootstrap a classifier
 1. Given a set of seed tuples
 2. Find sentences containing these seed tuples
 3. Extract general patterns from these sentences
 4. Use these patterns to find new tuples
 5. Repeat from step 2

Semi-supervised Relation Extraction

air company *city in Belgium*

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

Issue: 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*existing knowledge bases*
- Generate massive training sets, enabling the use of richer features, and no risk of semantic drift
- Still rely on a fixed set of relations

ReVERB: Unsupervised Relation Extraction

- If there is no relation database or the goal is to find new relations, unsupervised approaches must be used.
- Relations become substrings, usually containing 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 the substring 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 a small sample
 - ▶ Can only obtain (approximate) precision, not recall.

random sample
for human evaluation.

not available

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

pinpoint the real time

- **Normalisation**: mapping expressions to canonical forms.

- ▶ July 2, 2007 → 2007-07-02

Temporal

- 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.

on top of event Extraction

I move.

A Final Word

- Information Extraction is a vast field with many different tasks and applications
 - ▶ Named Entity Recognition + Relation Extraction
 - ▶ Events can be tracked by combining event and temporal expression extraction
- Machine learning methods involve classifiers and sequence labelling models.

Reading

- JM3 Ch. 18 – 18.2
- References:
 - ▶ Lample et al, Neural Architectures for Named Entity Recognition, NAACL 2016
<https://github.com/glample/tagger>