Information Extraction

COMP90042 Lecture 13

Introduction

- 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

Introduction

- 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 seq. classifiers, 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

"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 **[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.

IO tagging

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

Dealing with adjacent entities: IOB tagging

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

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

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’Xxxxxxxx / X’Xx
- 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.

NER - Features

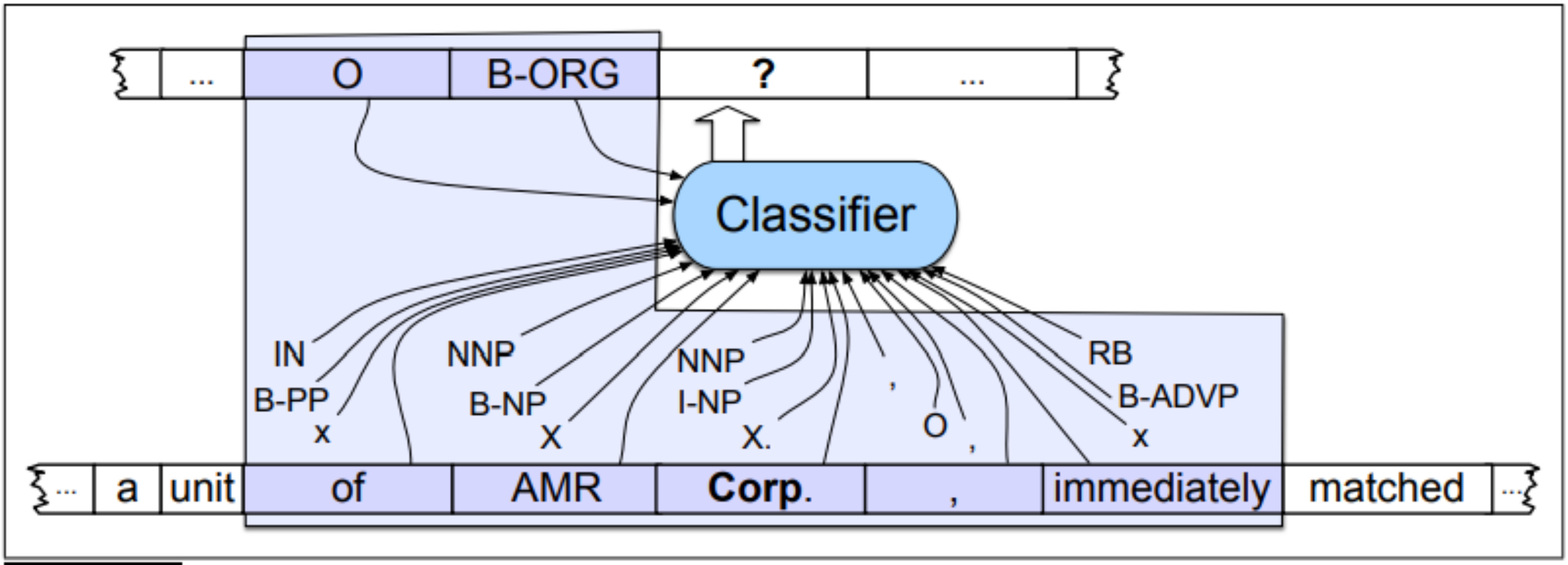
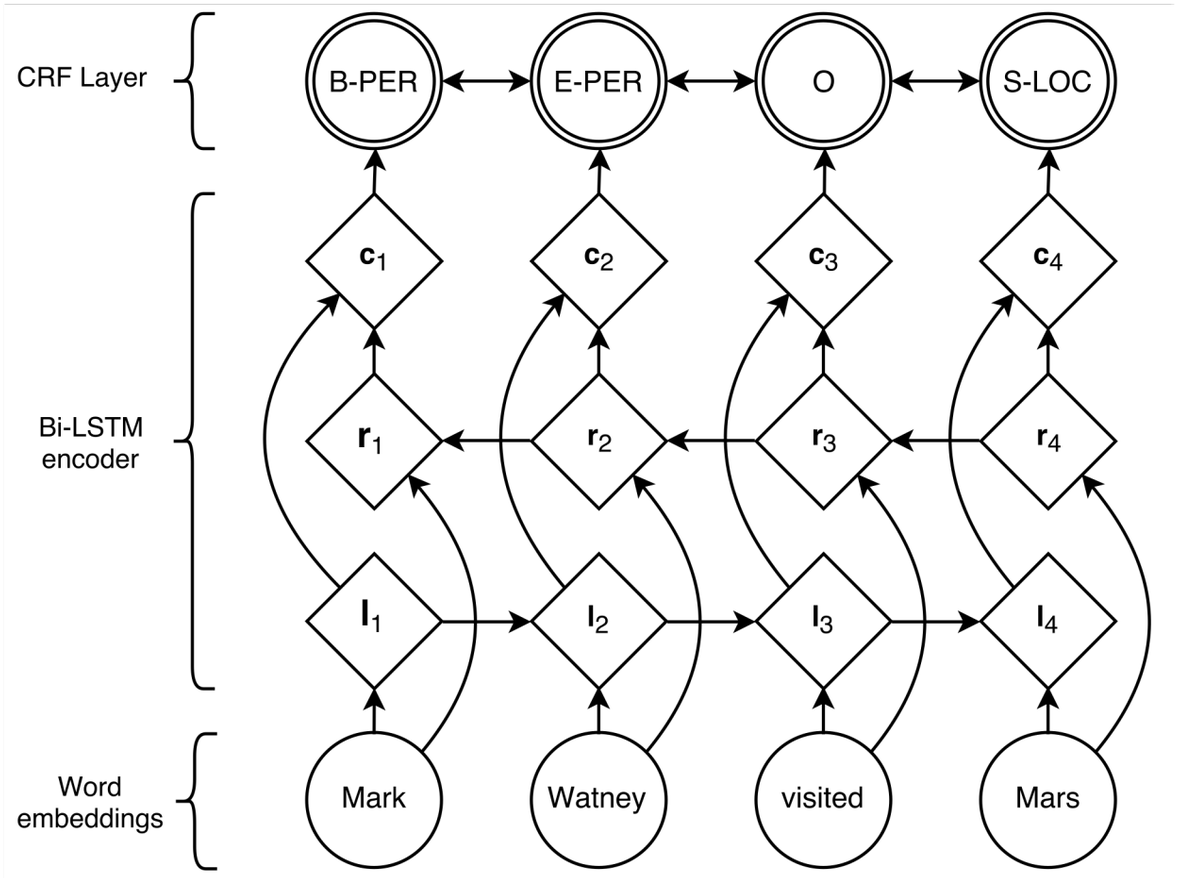
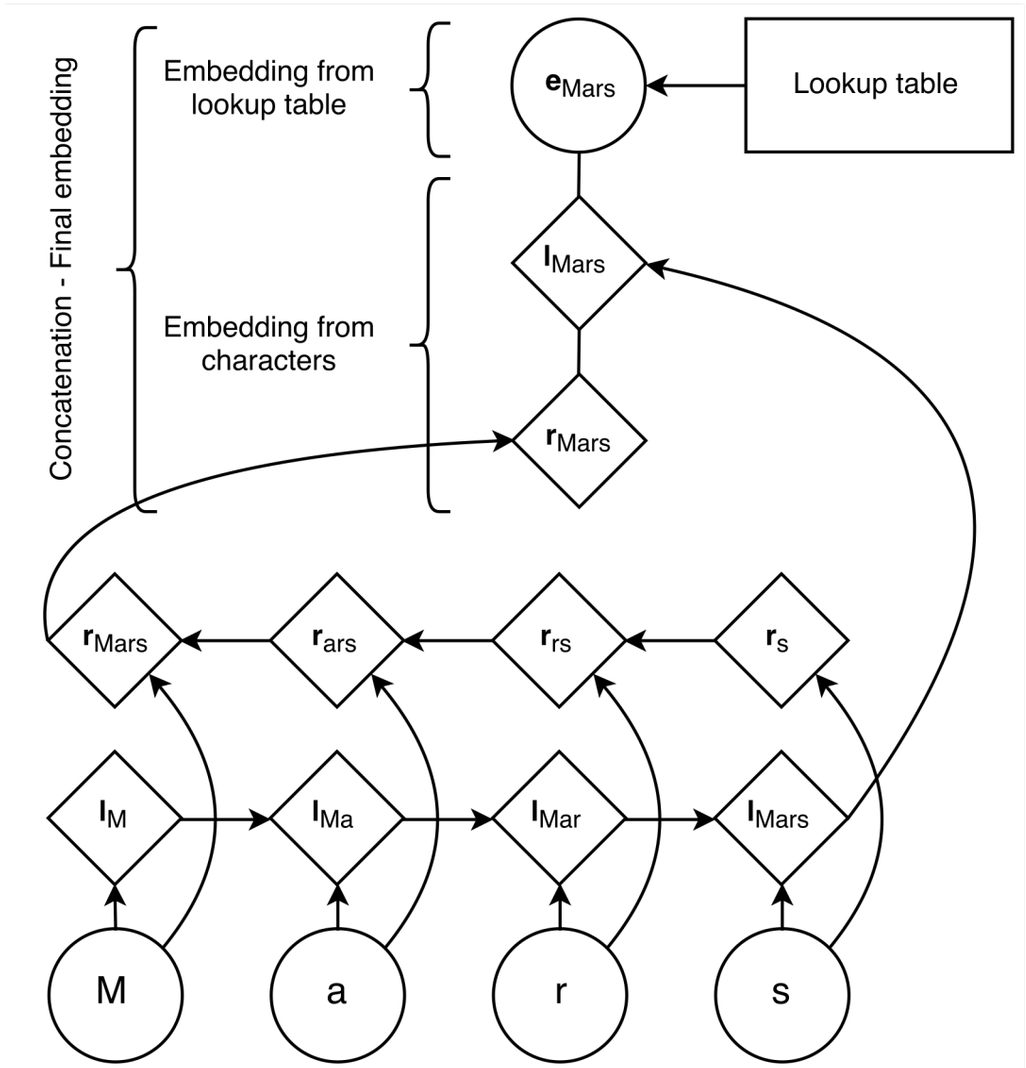


Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Deep models for NER

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





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 have access to a set of possible relations?
  - Answer depends on the application.

Relation extraction

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

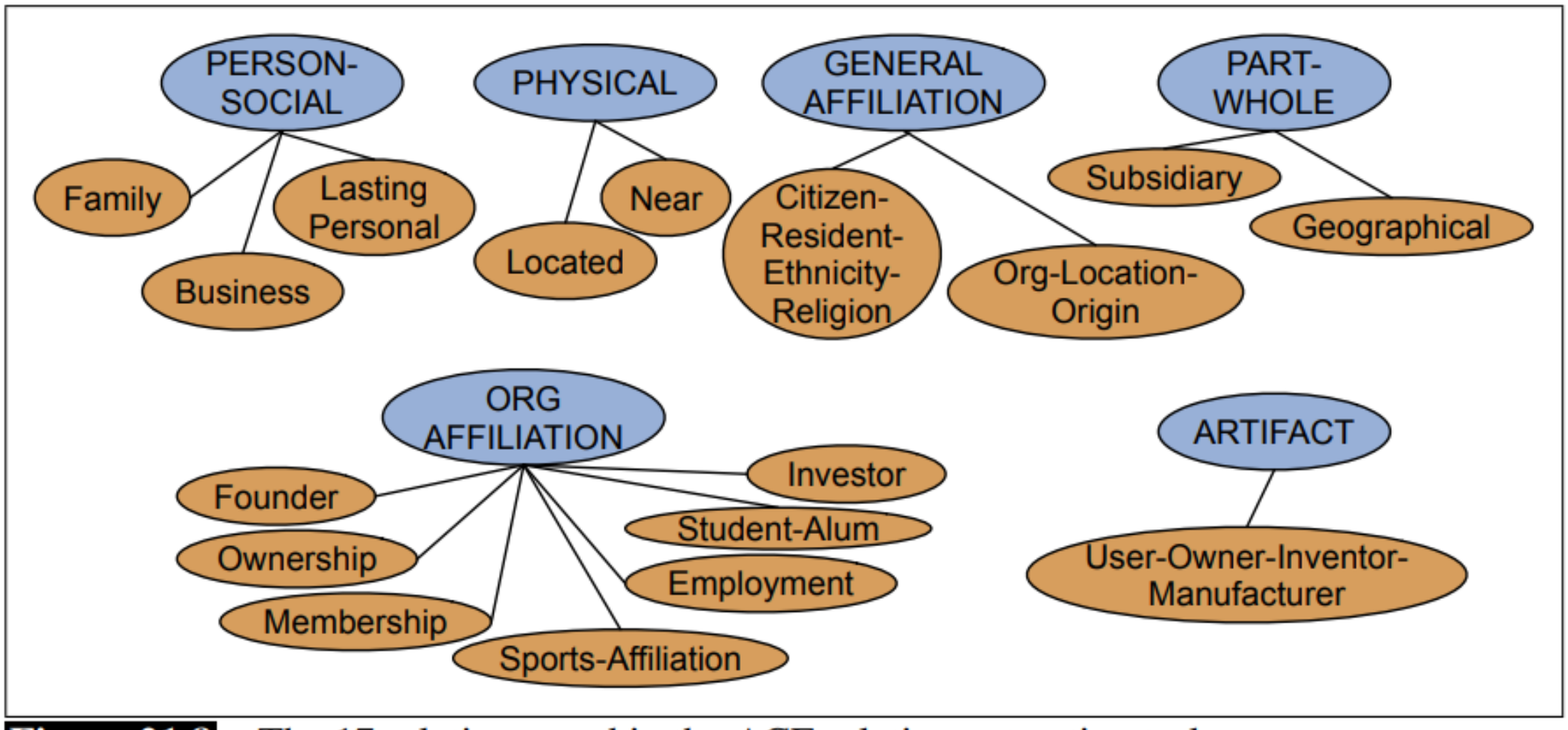


Figure 21.8 The 17 relations used in the ACE relation extraction task.

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

Supervised relation extraction

- Assume a corpus with annotated relations.
- Two steps. 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.
- Second, for pairs predicted as positive, use a multi-class classifier 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
  - (Tim Wagner, American Airlines) -> positive
  - (Tim Wagner, AMR Corp.) -> negative
- Second:
  - (American Airlines, AMR Corp.) -> subsidiary
  - (Tim Wagner, American Airlines) -> employment

Semi-supervised relation extraction

- Annotated corpora is very expensive to create.
- Assume we have a small set of **seed tuples**.
- Mine the web for text containing the tuples:
  - Given hub(Ryanair, Charleroi)
  - Get sentences containing all terms, e.g.,
    "Budget airline **Ryanair**, which uses **Charleroi** as a **hub**, scrapped all weekend flights out of the airport."
  - Use these patterns to new tuples, e.g., hub(Jetstar, Avalon) as these words occur in similar contexts; repeat
- Suffers from "semantic drift", where errors compound

Distant supervision

- Semi-supervised methods assume the existence of seed tuples.
- What about mining new tuples?
- 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
- 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.

Temporal expressions

"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"? Information usually present in metadata.
- **Normalisation**: mapping expressions to canonical forms.
- 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 relation extraction, including annotation and learning methods.
- **Event ordering**: detect how a set of events happened in a timeline.
  - Involves both event extraction and temporal extraction/normalisation.
  - Useful for rumour detection.

Template filling

- Some events can be represented as **templates**.
  - A “fare raise” event has an *airline*, an *amount* and a *date* when it occurred, among other possible **slots**.
- Goal is to fill these slots given a text. Models can take the template information into account to ease the learning and extraction process.
- Need to determine if a piece of text contain the information asked in the template (binary classification).

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
  - Template filling can help learning algorithms
- Machine learning methods involve classifiers and sequence labelling models.

Reading

- JM3 Ch. 17 – 17.2

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