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**].

COMP90042 LECTURE 13

INFORMATION EXTRACTION

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”)

INTRODUCTION

- “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 lowercost 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
- Tagset is application-dependent: some domains deal with specific entities such as 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 HMM for this?
  - 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.

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

- In practice, IO tagging usually enough since entities tend to be separated by punctuation and function words.
  - Final choice may depend on domain though.
- Given a tagging scheme and an annotated corpus, one can train any sequence labelling model
- In theory, HMMs can be used but feature-based models such as MEMMs and CRFs are preferred
  - Character-level features (is the first letter uppercase?)
  - Extra resources: gazeteers, databases
  - 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’X[x]*
- 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

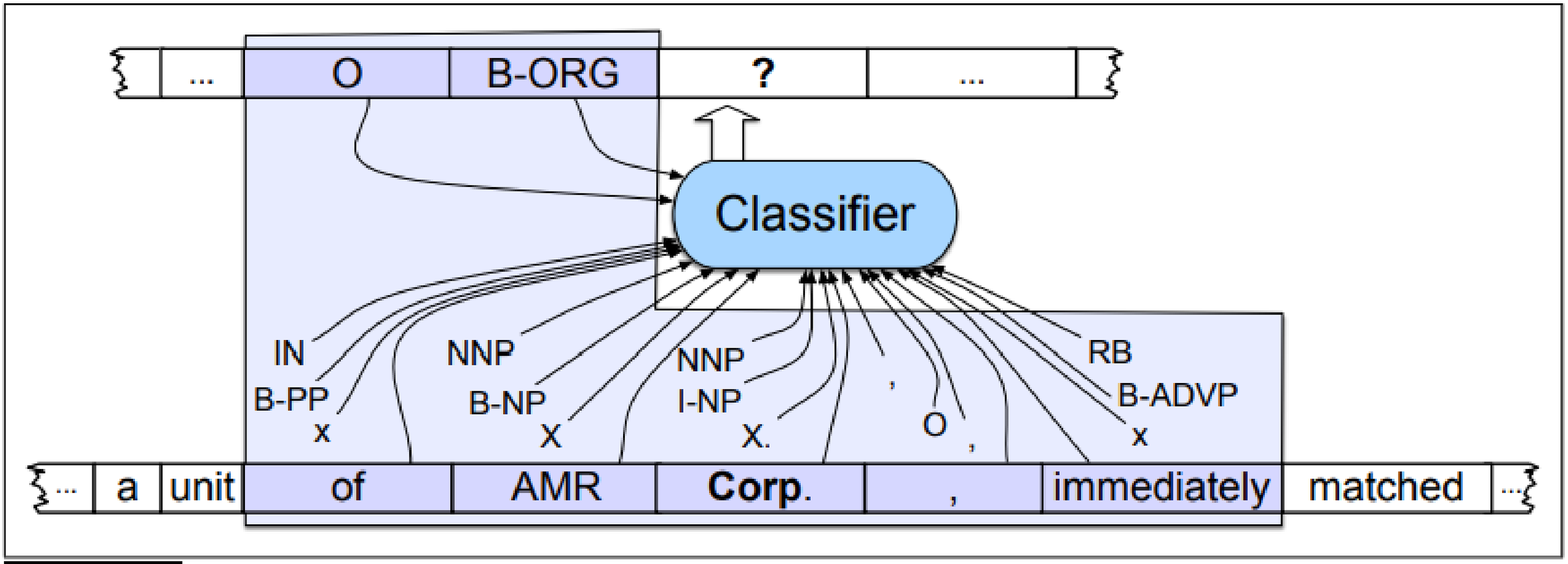

Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

NER SYSTEMS IN PRACTICE

- Purely sequence labelling approaches are the norm in academia
- In real world applications, pipeline approaches are more common. Key idea: some entities are easily caught by rules and lists.
  - 1) Apply high-precision rules
  - 2) Search for substring matches or previously detected entities (“Tim Wagner” -> “Tim” / “Wagner”)
  - 3) Consult domain-specific entity lists (ex: gazeteers)
  - 4) Apply sequence labelling using information from previous steps.

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

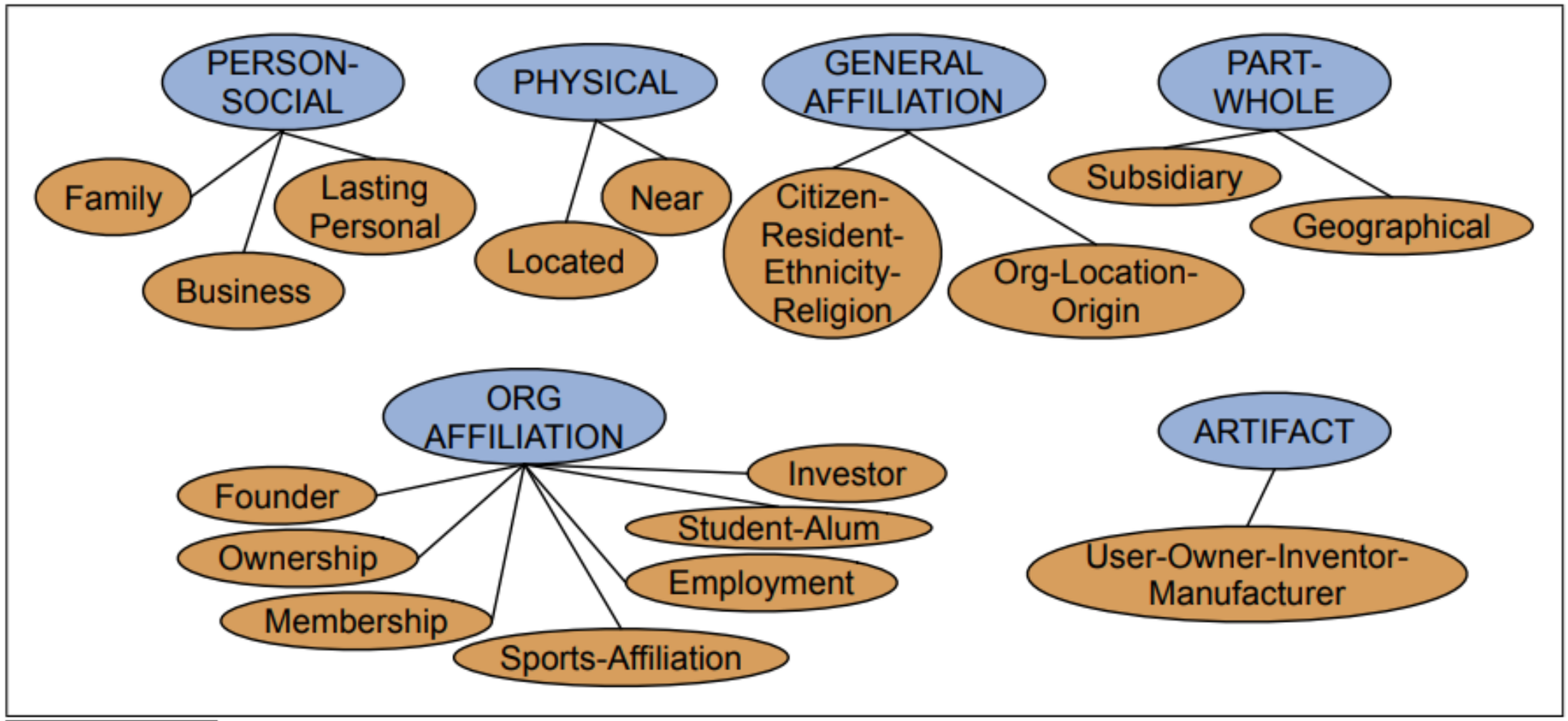

Figure 21.8 The 17 relations used in the ACE relation extraction task.

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

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 framed 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 set of **seed tuples**. These can be get from annotated corpora.
- Mine the web for text containing the tuples:
  - Given hub(Ryanair, Charleroi)
  - Get sentences such as: “Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.”
  - Add these to your training data.

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 very large training sets, enabling the use of richer features
- Still rely on a fixed set of relations.

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 (there is no such thing as unsupervised learning…)

TEMPORAL EXPRESSIONS

- "A fare increase initiated **last week** by UAL Corp's United Airlines was matched by competitors over **the weekend**, marking the second successful fare increase in **two weeks**."
- **Anchoring**: when is "last week"? Information usually present in metadata.
- **Normalisation**: mapping expressions to canonical forms.
- Mostly rule-based approaches. False positives are a problem:
  - …U2's classic *Sunday* Bloody *Sunday*

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.
  - Example application: 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
  - Database population usually a two-step process: 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. Many choices are possible, including deep learning approaches
- Domain is key for good performance

ADDITIONAL READING

- JM3 Ch. 21