Text Mining (2019)

Information extraction

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING
UNIVERSITY

This work is licensed under a
Creative Commons Attribution 4.0 International License.

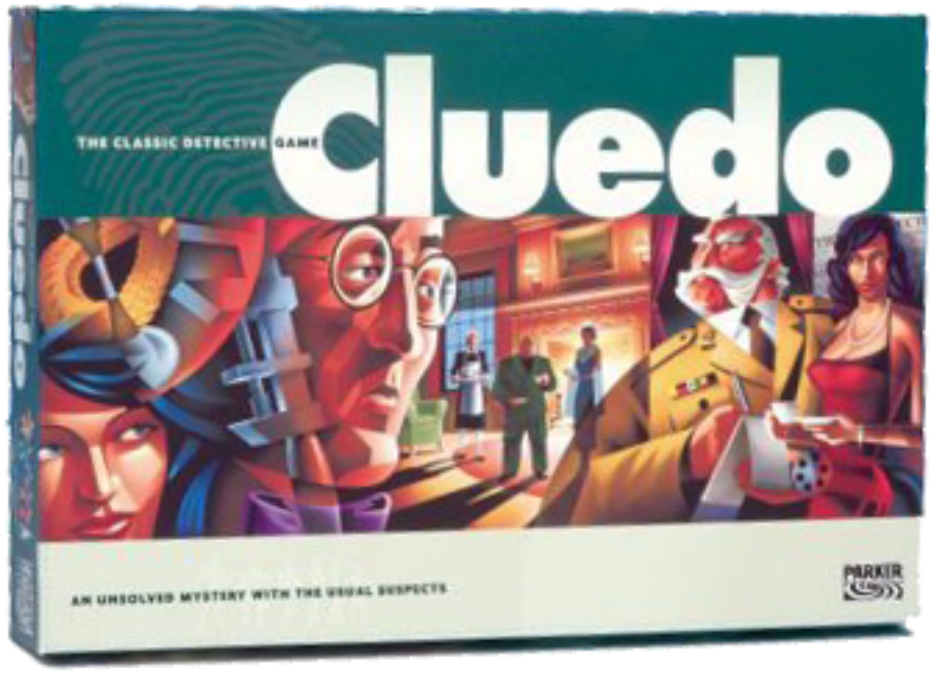
Information extraction

- **Information extraction (IE)** is the task of extracting structured information from running text.

- More specifically, the term ‘structured information’ refers to **named entities** and **semantic relations** between those entities.

  persons, organisations – *X* is-leader-of *Y*

Who did what to whom, where, and when?


THE CLASSIC DETECTIVE GAME
Cluedo
AN UNSOLVED MYSTERY WITH THE USUAL SUSPECTS
PARKER

named entities

semantic relations

Information extraction

As of 15 Mar 2002, Hawaii state health officials reported one additional recent case of dengue fever and 6 cases that occurred last year but were not confirmed by laboratory testing until 2002.

Attribute	Value
docno	ProMed.20020322.11
doc_date	2002.03.22
disease_name	dengue fever
norm_stime	2002.03.15
norm_etime	2002.03.15
victim_types	—
location	Hawaii

Source: Grishman et al. (2002)

Why information extraction?

- to find information expressed in natural language

  company acquisitions and mergers

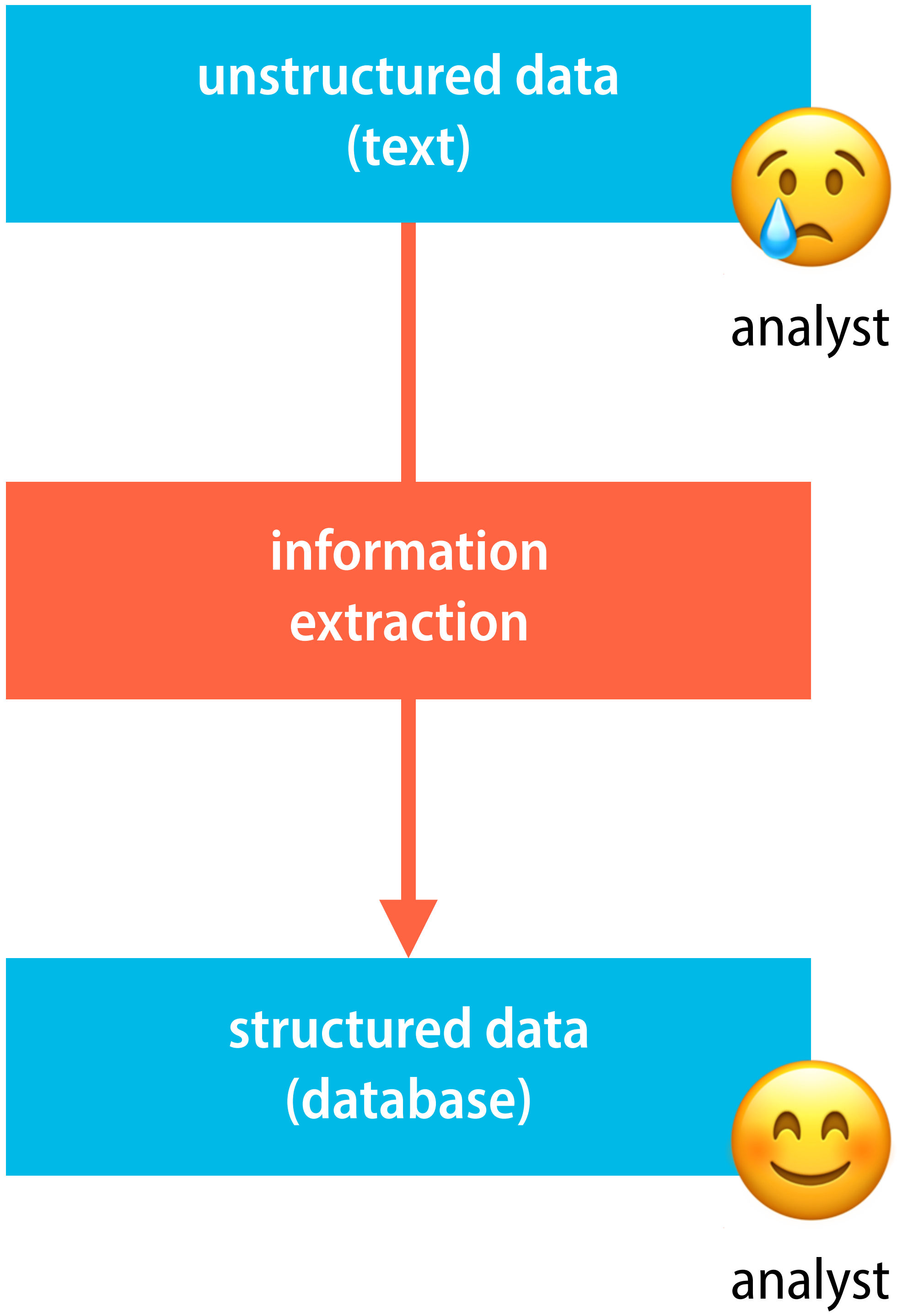
- to create or maintain knowledge bases

  Knowledge Graph, DBPedia

- to support question answering systems

  IBM's Watson

The Knowledge Gap



unstructured data
(text)

😢

analyst

information
extraction

structured data
(database)

😊

analyst

JEOPARDY!®

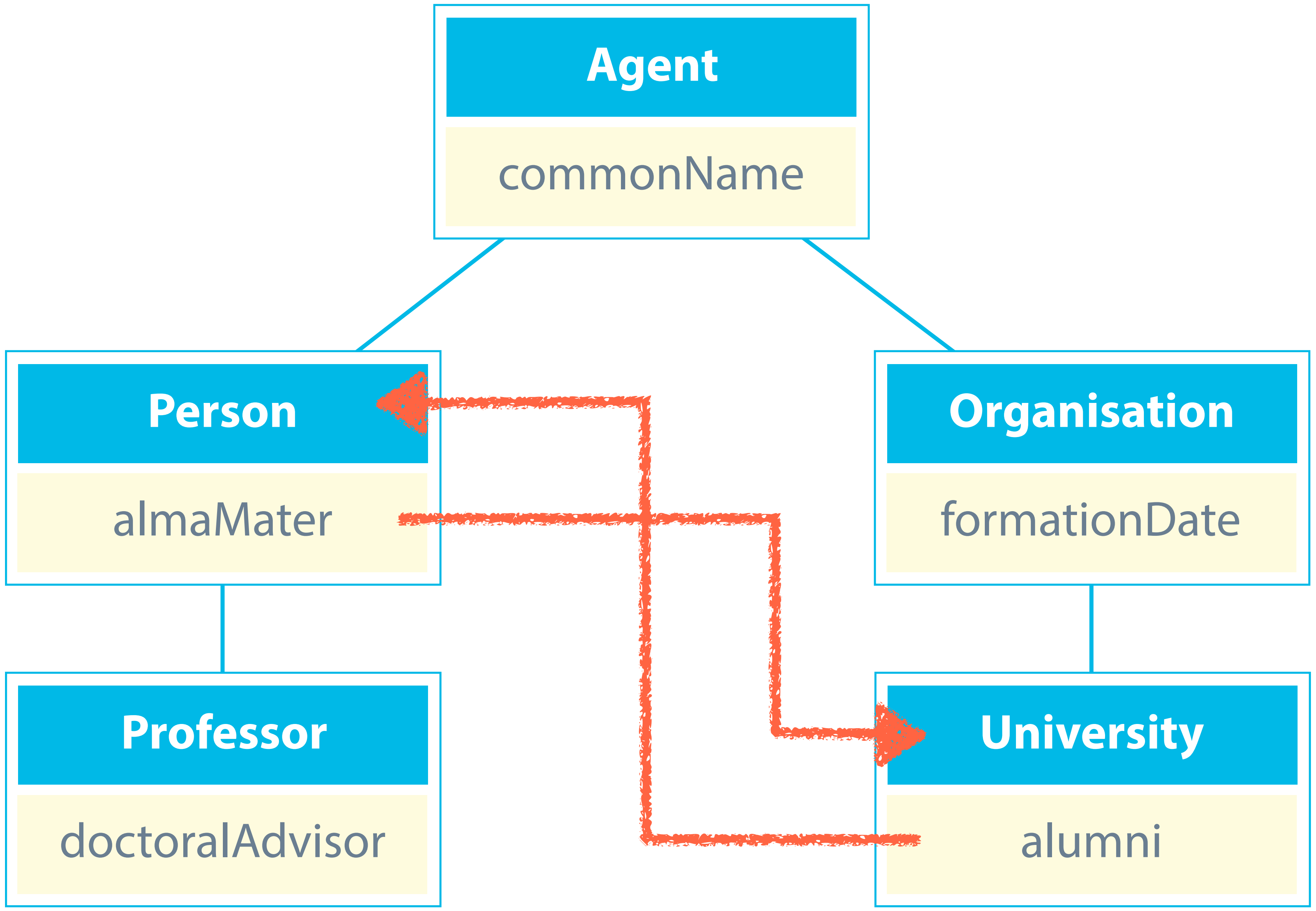
This Stanford University alumnus co-founded educational technology company Coursera.

Source: MacArthur Foundation

SPARQL query against DBPedia

```
SELECT DISTINCT ?x WHERE {
  ?x dbo:almaMater dbr:Stanford_University.
  dbr:Coursera dbo:foundedBy ?x.
}
```

Part of the DBPedia ontology



http://wiki.dbpedia.org/Ontology

This lecture

- Introduction to information extraction
- Named entity recognition
- Entity linking
- Relation extraction

Named entity recognition

Named entity recognition

Named entity recognition (NER) is the task of finding mentions of named entities in a text, and labelling them with their type.

person, company, organisation, geopolitical entity

Named entities, example

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].

Example from Jurafsky and Martin (2017)

Properties of named entities

- can be referred to with a proper name
- can be indexed and linked to
- participate in semantic relations
- are common answers in question answering systems
- can be associated with attitudes

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.

Type	Description
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ‘%’.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	first’, ‘second’, etc.
CARDINAL	Numerals that do not fall under another type.

Some types of named entities in DBPedia

- **Persons:** Actor, Curler, FictionalCharacter
- **Organisations:** Band, Company, SportsTeam
- **Places:** Building, Mountain, Country
- **Dates and times:** Date, Year, HistoricalPeriod
- **Medical terms:** Muscle, Enzyme, Disease

Gazetteers

- In a narrow sense, a **gazetteer** is a 'geographical index or dictionary' (Oxford English Dictionary).
- In the broader sense in which this term is used in named entity recognition, it often refers to a list of names.

Gazetteers, example

Ale Alingsås Alvesta Aneby Arboga Arjeplogs Arvidsjaurs Arvika Askersunds Avesta Bengtsfors Bergs Bjurholms Bjuvs Bodens Bollebygds Bollnäs Borgholms Borlänge Borås Botkyrka Boxholms Bromölla Bräcke Burlövs Båstads Dals-Eds Danderyds Degerfors Dorotea Eda Ekerö Eksjö Emmaboda Enköpings Eskilstuna Eslövs Essunga Fagersta Falkenbergs Falköpings Falu Filipstads Finspångs Flens Forshaga Färgelanda Gagnefs Gislaveds Gnesta Gnosjö Gotlands Grums Grästorps Gullspångs Gällivare Gävle Göteborgs Götene Habo Hagfors Hallsbergs Hallstahammars Halmstads Hammarö Haninge Haparanda Heby Hedemora Helsingborgs Herrljunga Hjo Hofors Huddinge Hudiksvalls Hultsfreds Hylte Håbo Hällefors Härjedalens Härnösands Härryda Hässleholms Höganäs Högsby Hörby Höörs Jokkmokks Järfälla Jönköpings Kalix Kalmar Karlsborgs Karlshamns Karlskoga Karlskrona Karlstads Katrineholms Kils Kinda Kiruna Klippans Knivsta Kramfors Kristianstads Kristinehamns Krokoms Kumla Kungsbacka Kungsörs Kungälvs Kävlinge Köpings Laholms Landskrona Laxå Lekebergs Leksands Lerums Lessebo Lidingö Lidköpings Lilla Edets Lindesbergs Linköpings Ljungby Ljusdals Ljusnarsbergs Lomma Ludvika Luleå Lunds Lycksele Lysekils Malmö Malung-Sälens Malå Mariestads Marks Markaryds Melleruds Mjölby Mora Motala Mullsjö Munkedals Munkfors Mölndals Mönsterås Mörbylånga Nacka Nora Norbergs Nordanstigs Nordmalings Norrköpings Norrtälje Norsjö Nybro Nykvarns Nyköpings Nynäshamns Nässjö Ockelbo Olofströms Orsa Orusts Osby Oskarshamns Ovanåkers Oxelösunds Pajala Partille Perstorps Piteå Ragunda Robertsfors Ronneby Rättviks Sala Salems Sandvikens Sigtuna Simrishamns Sjöbo Skara Skellefteå Skinnskattebergs Skurups Skövde Smedjebackens Sollefteå Sollentuna Solna Sorsele Sotenäs Staffanstorps Stenungsunds Stockholms Storfors Storumans Strängnäs Strömstads Strömsunds Sundbybergs Sundsvalls Sunne Surahammars Svalövs Svedala Svenljunga Säffle Säters Sävsjö Söderhamns Söderköpings Södertälje Sölvesborgs Tanums Tibro Tidaholms Tierps Timrå Tingsryds Tjörns Tomelilla Torsby Torsås Tranemo Tranås Trelleborgs Trollhättans Trosa Tyresö Täby Töreboda Uddevalla Ulricehamns Umeå Upplands Väsby Upplands-Bro Uppsala Uppvidinge Vadstena Vaggeryds Valdemarsviks Vallentuna Vansbro Vara Varbergs Vaxholms Vellinge Vetlanda Vilhelmina Vimmerby Vindelns Vingåkers Vårgårda Vänersborgs Vännäs Värmdö Värnamo Västerviks Västerås Växjö Ydre Ystads Åmåls Ånge Åre Årjängs Åsele Åstorps Åtvidabergs Älmhults Älvdalens Älvkarleby Älvsbyns Ängelholms Öckerö Ödeshögs Örebro Örkelljunga Örnsköldsviks Östersunds Österåkers Östhammars Östra Göinge Överkalix Övertorneå

Inflected names in Polish

Case	Form
Nominative	Muammar Kaddafi
Genitive	Muammara Kaddafiego
Dative	Muammarowi Kaddafiemu
Accusative	Muammara Kaddafiego
Instrumental	Muammarem Kaddafim
Locative	Muammarze Kaddafim
Vocative	Muammarze Kaddafi

Piskorsky and Yangarber (2013)

Type ambiguities

- [PER Washington] was born into slavery.
- [ORG Washington] went up 2 games to 1 in the four-game series.
- Blair arrived in [LOC Washington] for his last state visit.
- In June, [GPE Washington] passed a primary seatbelt law.
- The [VEH Washington] had proved to be a leaky ship,…

Examples adapted from Jurafsky and Martin (2017)

Evaluation of named entity recognition

- One way to view named entities is as spans, specified by their start position and their end position in the sentence.

  more generally, also labelled with entity type

- Based on this view, both gold-standard entities and predicted entities can be represented as sets of tuples.
- We can then use standard measures such as precision, recall, and F1 to evaluate named entity recognizers.

Evaluation of named entity recognition

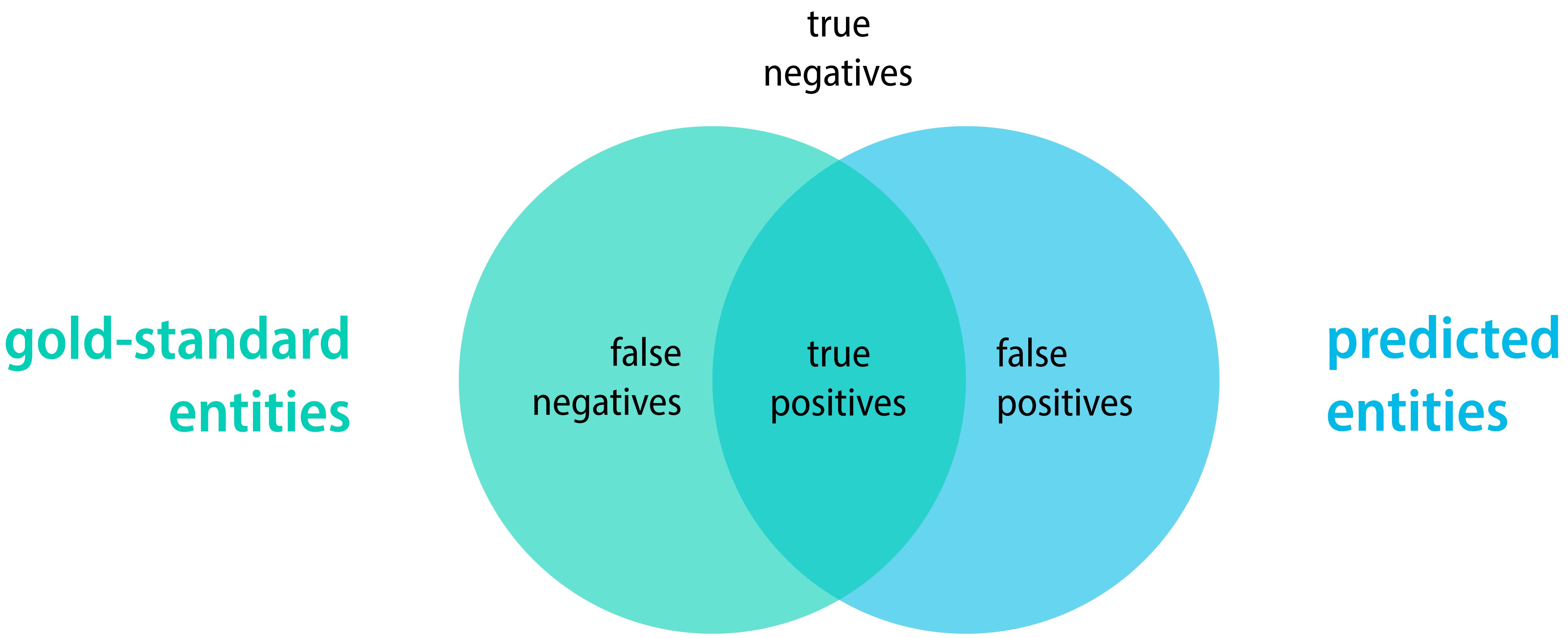
Position	Start	Entity
1	Foreign	ORG
2	ministry	ORG
3	spokesman	
4	Shen	PERSON
5	Guofang	PERSON
6	told	
7	Reuters	ORG

Corresponding entity spans:

(1, 2, ORG)
(4, 5, PERSON)
(7, 7, ORG)

Example from the CoNLL 2003 NER data set

Precision and recall for named entity recognition



$$P = \frac{|\text{gold} \cap \text{predicted}|}{|\text{predicted}|} \qquad R = \frac{|\text{gold} \cap \text{predicted}|}{|\text{gold}|}$$

Issues in the evaluation of NER systems

		gold standard	system
1	First	ORG	
2	Bank	ORG	ORG
3	of	ORG	ORG
4	Chicago	ORG	ORG
5	announced		
6	earnings		

1–4 ORG

2–4 ORG

Example from Jurafsky and Manning

Named entity recognition as sequence labelling

- State-of-the algorithms treat named entity recognition as a word-by-word sequence labelling task.

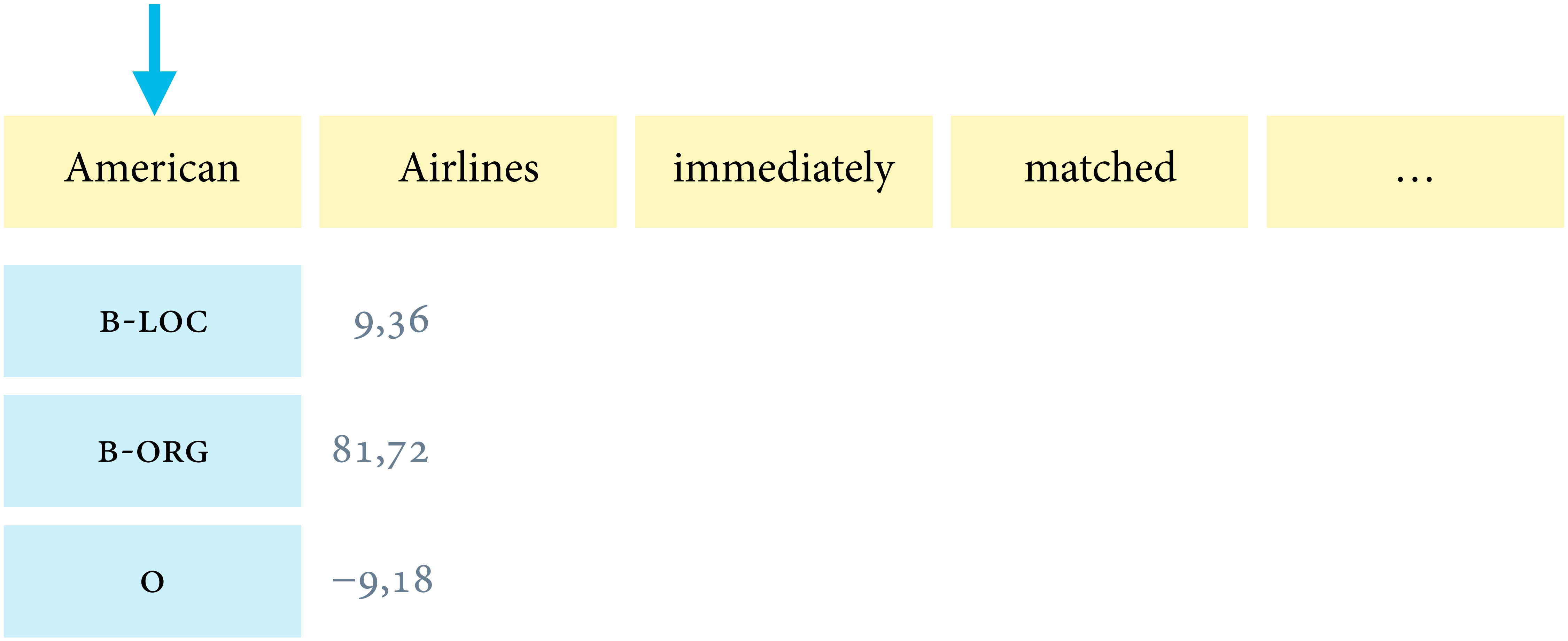
- The basic idea is to label words with tags that can encode the boundaries and the types of named entity mentions.

- A common encoding is the **IOB scheme**, where there is a tag for the beginning (B) and inside (I) of each entity type, as well as an additional tag for tokens outside (O) any entity.

Named entity recognition as sequence labelling

Token	IOB tag
American	B-ORG
Airlines	I-ORG
immediately	O
matched	O
the	O
move	O
Wagner	B-PER
said	O
.	O

Named entity tagging with a sequence classifier



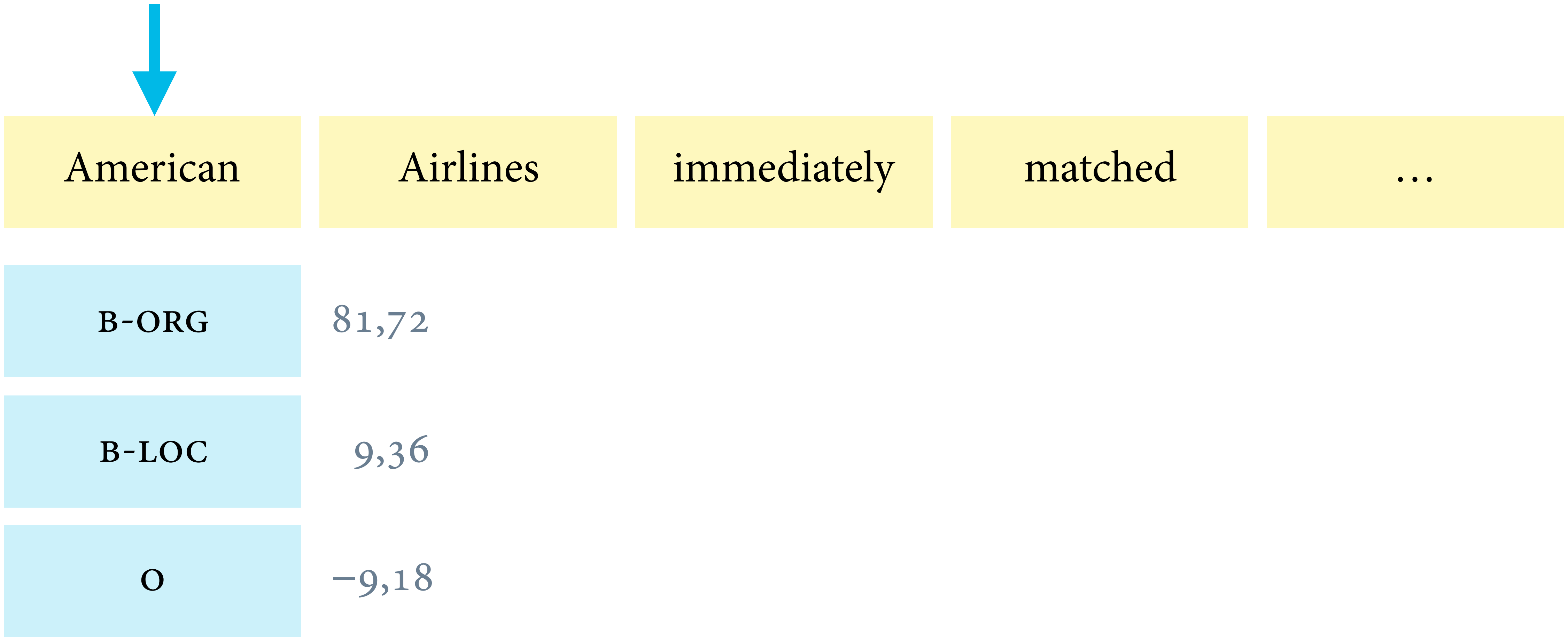
American Airlines immediately matched …

B-LOC 9,36

B-ORG 81,72

O −9,18

Named entity tagging with a sequence classifier



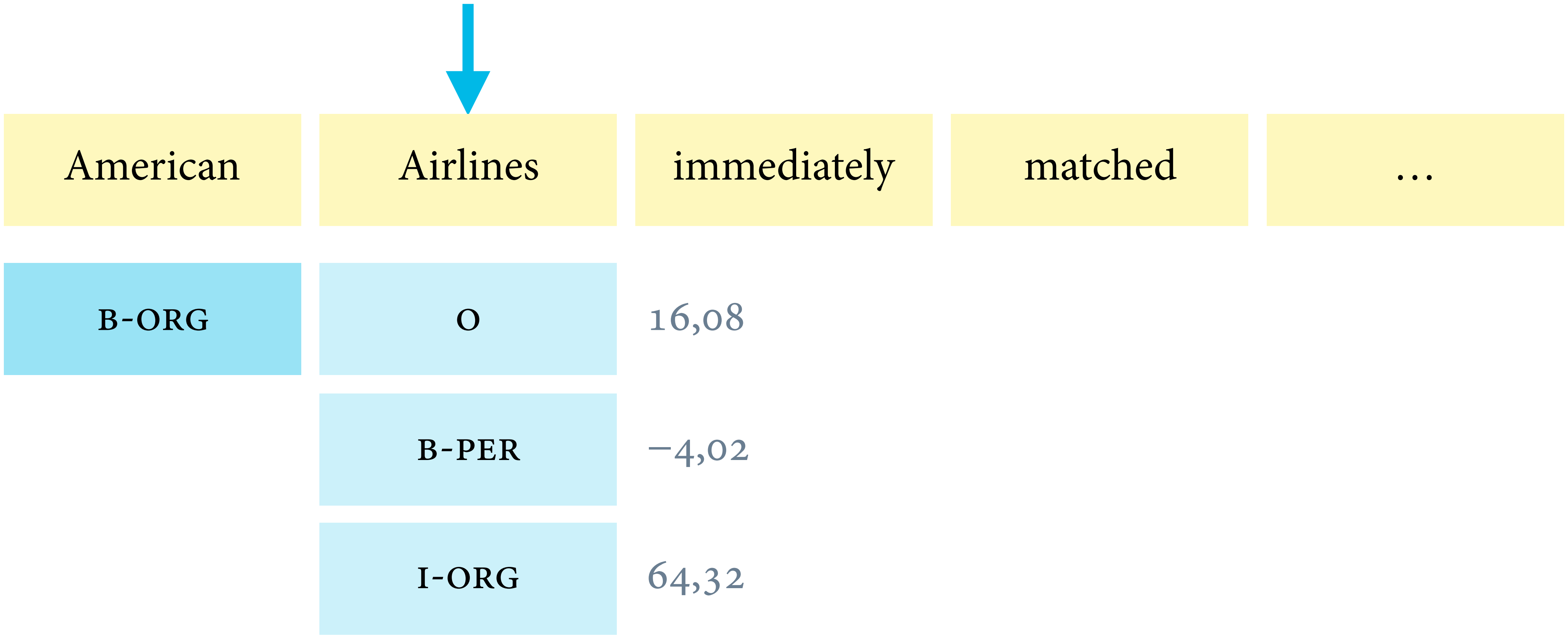
American | Airlines | immediately | matched | …

B-ORG 81,72

B-LOC 9,36

O −9,18

Named entity tagging with a sequence classifier



American

Airlines

immediately

matched

…

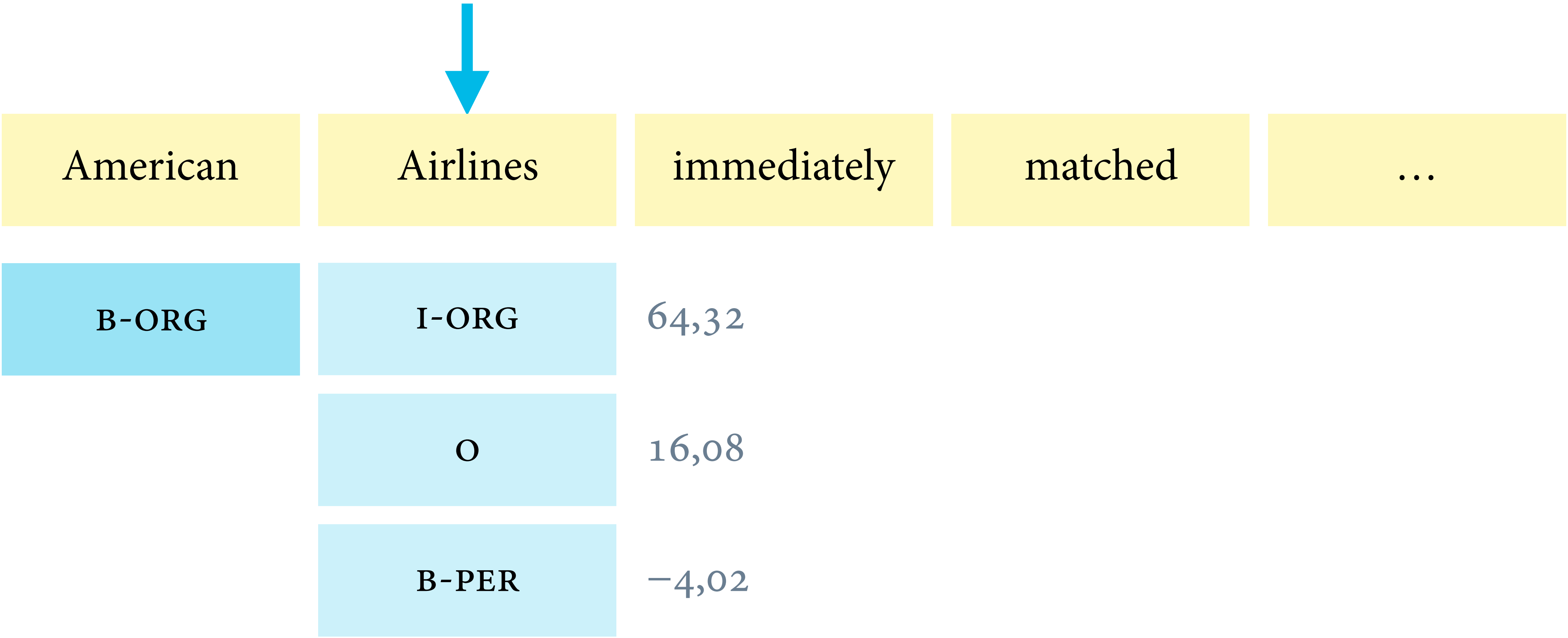
B-ORG

O 16,08

B-PER −4,02

I-ORG 64,32

Named entity tagging with a sequence classifier



American | Airlines | immediately | matched | …

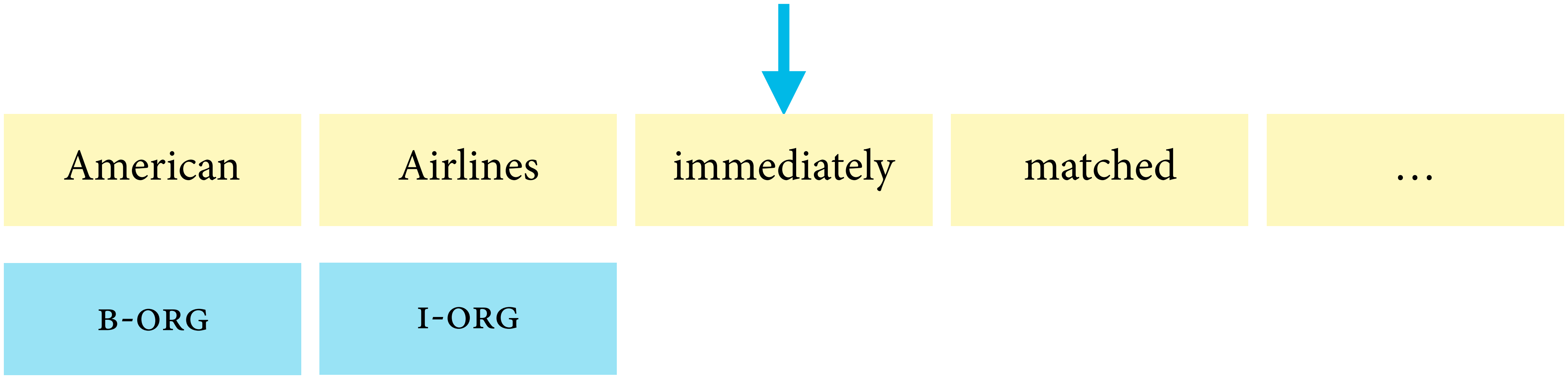
B-ORG

I-ORG 64,32

O 16,08

B-PER −4,02

Named entity tagging with a sequence classifier



American	Airlines	immediately	matched	…
B-ORG	I-ORG			

Features commonly used in NER systems

- identity of current word and neighbouring words
- part-of-speech of current word and neighbouring words
- presence of current word in a gazetteer
- current word has a particular prefix or suffix
- word shape of current word (USA, IMF → XXX)
- syntactic contexts (dependency trees)

Adapted from Jurafsky and Martin (2017)

Standard models for named entity recognition

- **Maximum Entropy Markov Model (MEMM)**

  The classifier makes a single decision at a time. A globally optimal sequence of decisions is found using Viterbi search.

  or beam search

- **Conditional Random Fields (CRF)**

  whole-sequence model

- **Recurrent Neural Networks (LSTMs)**

This lecture

- Introduction to information extraction
- Named entity recognition
- Entity linking
- Relation extraction

Entity linking

Knowledge bases

- A **knowledge base (KB)** stores structured and unstructured information in a machine-readable way.

- In contrast to standard relational databases, knowledge bases are often based on an explicit object model.

  type hierarchy, ontology

WordNet

- Three separate databases: nouns, verbs, adjectives and adverbs.

  WordNet 3.0: 117,798 nouns, 11,529 verbs, 26,960 adjectives and adverbs

- Each lemma is annotated with one or more senses, represented as **synsets**, sets of cognitive synonyms.

  https://wordnet.princeton.edu

Relations between senses of different words

- **Synonymy – Antonymy**

  the situation where two senses of two different words (lemmas) are identical or nearly identical – opposite of each other

  couch/sofa, car/automobile – cold/hot, leader/follower

- **Hyponymy – Hypernymy**

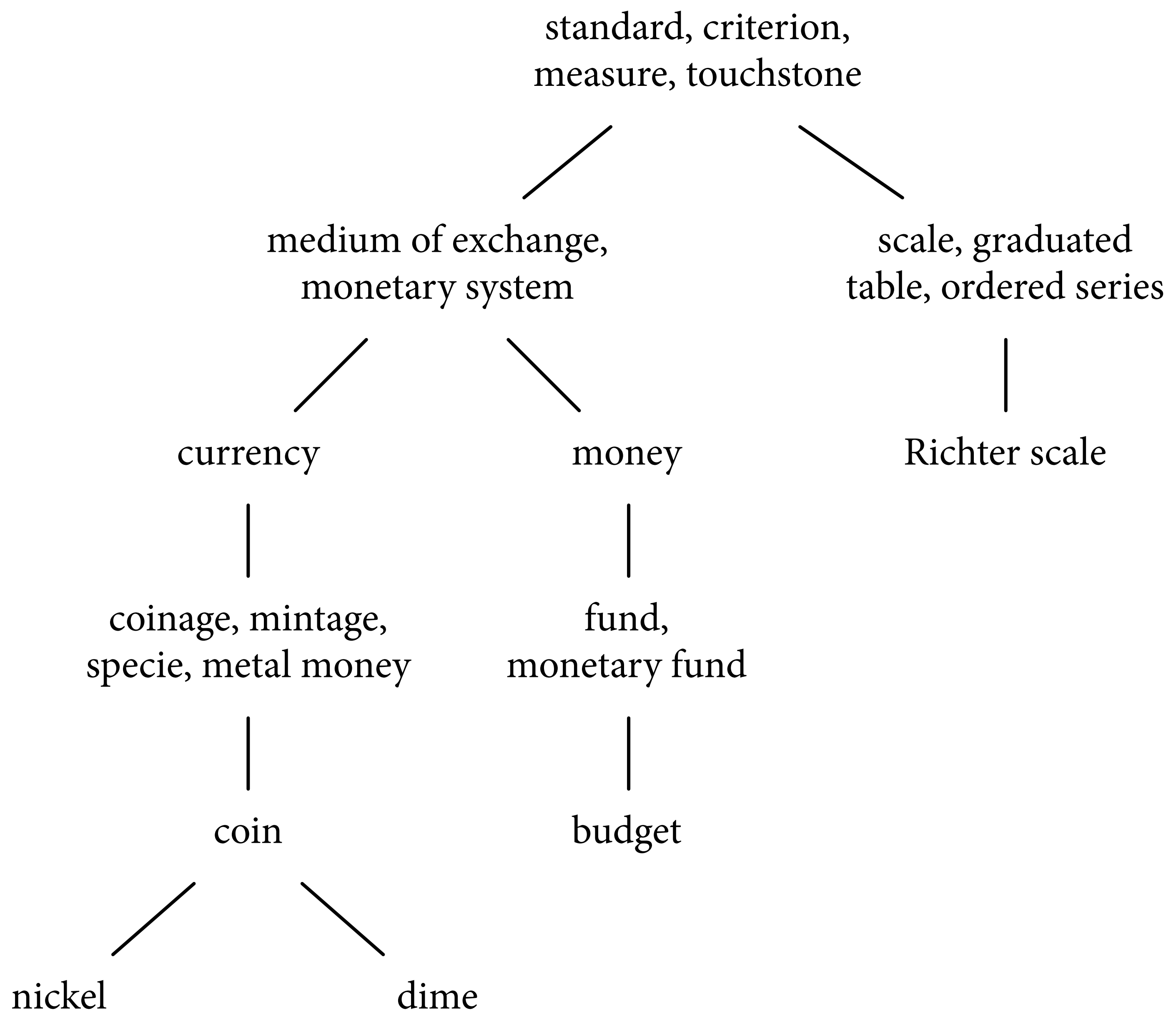
  the situation where in a pair of two senses of different words, one is more specific – less specific than the other

  car/vehicle, mango/fruit – furniture/chair, mammal/dog

Relations between noun senses in WordNet

Concept A	Semantic Relation	Concept B
breakfast^1	hyponym of	meal^1
meal^1	hypernym of	lunch^1
Bach^1	instance hyponym of	composer^1
author^1	instance hypernym of	Austen^1
leader^1	antonym of	follower^1

A small part of WordNet


standard, criterion, measure, touchstone
medium of exchange, monetary system
scale, graduated table, ordered series
currency
money
Richter scale
coinage, mintage, specie, metal money
fund, monetary fund
coin
budget
nickel
dime

DBPedia

- DBPedia is a crowd-sourced community effort to extract structured content from various Wikimedia projects.
- The English version of the DBPedia knowledge base contains information about more than 4 million entities.

http://dbpedia.org

YAGO

- **YAGO** was automatically constructed from several Wikipedias and other sources, such as GeoNames.
- Currently, YAGO has knowledge of more than 10 million entities and more than 120 million facts about these entities.
- Many of the facts and entities in YAGO are labelled with temporal and spatial information.

http://yago-knowledge.org

YAGO about Arthur Conan Doyle

Fact	Entity
<wasBornIn>	<Edinburgh>
<wasBornOnDate>	“1859-05-22”^^xsd:date
<diedOnDate>	“1930-07-07”^^xsd:date
<hasGivenName>	“Arthur”@eng
<rdf:type>	<wikicat_People_from_Edinburgh>

This lecture

- Introduction to information extraction
- Named entity recognition
- Entity linking
- Relation extraction

Relation extraction

Semantic relations, example

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.

is spokesman for

Unified Medical Language

135 entity types, 54 relation types

Injury	**disrupts**	Physiologic Function
Bodily Location	**location-of**	Biologic Function
Anatomical Structure	**part-of**	Organism
Pharmacologic Substance	**causes**	Pathologic Function
Pharmacologic Substance	**treats**	Pathologic Function

Jurafsky and Martin (2017)

Relation extraction using regular expressions

Semantic relations can be extracted using regular expressions.

Example: .*\bfödd.*\b

- [PER August Strindberg], född [DATE 22 januari 1849] …

  → \1 was-born-year \2

- [PER August Strindberg], som föddes [DATE 1849], …

  → \1 was-born-year \2

Text patterns for the X is-a Y relation

Pattern	Example
X and other *Y*	… temples, treasuries, **and other** civic buildings.
X or other *Y*	Bruises, wounds, broken bones **or other** injuries …
Y such as *X*	The bow lute, **such as** the Bambara ndang …
Such *Y* as *X*	… **such** authors **as** Herrick, Goldsmith, and Shakespeare.
Y including *X*	… common-law countries, **including** Canada.
Y, especially *X*	European countries, **especially** France and Spain, …

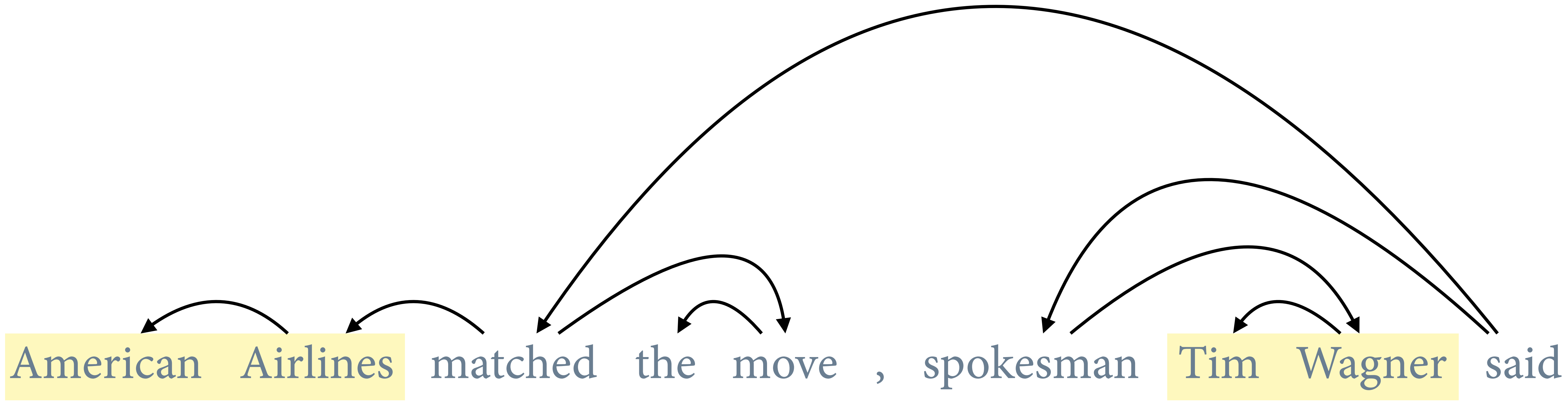
Jurafsky and Martin (2017)

Relation extraction based on dependency trees

- Run the sentence through a named entity recogniser and a dependency parser.
- For each pair of candidate entities, extract the shortest path between the two entities in the tree.
- Feed this path into a neural network and let it predict whether there is a relation between the two entities as well as its type.

  requires recurrent neural networks such as LSTM

Relation extraction based on dependency trees



extracted path between the two entities:
[ORG American Airlines] <matched said >spokesman [PER Tim Wagner]

This lecture

- Introduction to information extraction
- Named entity recognition
- Entity linking
- Relation extraction