

Natural Language Processing

Text Mining: Information Extraction

Text mining is the process of deriving high-quality information from text.

Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics.

Subtasks:

- Information retrieval;
- Named Entity Recognition;
- Recognition of Pattern Identified Entities;
- Coreference;
- Relationship, fact, and event Extraction;
- Sentiment analysis;
- Quantitative text analysis

Text Mining Applications

- Security applications;
- Biomedical applications;
- Software applications;
- Online media applications;
- Business and marketing applications;
- Sentiment analysis;
- Academic applications;
- Digital humanities and computational sociology

Information Extraction

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.

Tasks:

- **Named Entity Extraction**
 - Named Entity Recognition;
 - Coreference resolution;
 - Relationship Extraction;
- **Semi-structured Information Extraction**
 - Table extraction;
 - Comments extraction;
- **Language and vocabulary analysis**
 - Terminology extraction
- **Audio extraction**



Named Entity Recognition (NER)

- A very important sub-task: **find** and **classify** names in text, for example:
 - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Information Extraction

- **Information extraction** (IE) turns the unstructured information embedded in texts into structured data, for example for populating a relational database to enable further processing.
- The task of **named entity recognition** (NER) is to find each mention of a named entity in the text and label its type.
- The task of **relation extraction** is to find and classify semantic relations among the text entities, often binary relations like spouse-of, child-of, employment, part-whole, membership, and geospatial relations.
- The task of **event extraction** is to find events in which these entities participate.
- To figure out when the events in a text happened we'll do recognition of **temporal expressions** like days of the week (Friday and Thursday), months, holidays, temporal expression etc.
- The task of **template filling** is to find situations in documents and fill the template slots with appropriate material.

Information Extraction

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.



FARE-RAISE ATTEMPT:

LEAD AIRLINE:	UNITED AIRLINES
AMOUNT:	\$6
EFFECTIVE DATE:	2006-10-26
FOLLOWER:	AMERICAN AIRLINES

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].

Named Entity Recognition

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	Turing is a giant of computer science.
Organization	ORG	companies, sports teams	The IPCC warned about the cyclone.
Location	LOC	regions, mountains, seas	The Mt. Sanitas loop is in Sunshine Canyon .
Geo-Political Entity	GPE	countries, states, provinces	Palo Alto is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the Tappan Zee Bridge .
Vehicles	VEH	planes, trains, automobiles	It was a classic Ford Falcon .

Figure 21.1 A list of generic named entity types with the kinds of entities they refer to.

Named Entity Recognition: Problems

Name	Possible Categories
<i>Washington</i>	Person, Location, Political Entity, Organization, Vehicle
<i>Downing St.</i>	Location, Organization
<i>IRA</i>	Person, Organization, Monetary Instrument
<i>Louis Vuitton</i>	Person, Organization, Commercial Product

Figure 21.2 Common categorical ambiguities associated with various proper names.

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.
The [VEH Washington] had proved to be a leaky ship, every passage I made...

Figure 21.3 Examples of type ambiguities in the use of the name *Washington*.

IOB and IO encodings

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

In IOB tagging we introduce a tag for the beginning (B) and inside (I) of each entity type, and one for tokens outside (O) any entity.

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

Figure 21.4

Named entity tagging as a sequence model, showing IOB and IO encodings.

Word Shape

Word shape features are used to represent the abstract letter pattern of the word by mapping lower-case letters to 'x', upper-case to 'X', numbers to 'd', and retaining punctuation.

Example:

I.M.F. → X.X.X.

DC10-30 → XXdd-dd

Short word shape: consecutive character types are removed, so DC10-30 would be mapped to Xd-d but I.M.F would still map to X.X.X.

Features for NER

identity of w_i
identity of neighboring words
part of speech of w_i
part of speech of neighboring words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazetteer**
 w_i contains a particular prefix (from all prefixes of length ≤ 4)
 w_i contains a particular suffix (from all suffixes of length ≤ 4)
 w_i is all upper case
word shape of w_i
word shape of neighboring words
short word shape of w_i
short word shape of neighboring words
presence of hyphen

Figure 21.5 Features commonly used in training named entity recognition systems.

For example the named entity token *L'Occitane* would generate the following non-zero valued feature values:

prefix(w_i) = L
prefix(w_i) = L '
prefix(w_i) = L 'O
prefix(w_i) = L 'Oc
suffix(w_i) = tane
suffix(w_i) = ane
suffix(w_i) = ne
suffix(w_i) = e
word-shape(w_i) = X'Xxxxxxxx
short-word-shape(w_i) = X'Xx

NER

A **gazetteer** is a list of place names, and they can offer millions of entries for all manner of locations along with detailed geographical, geologic, and political information.

Example: <http://www.geonames.org/>

NER as a Sequence Labeling

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 21.6 Word-by-word feature encoding for NER.

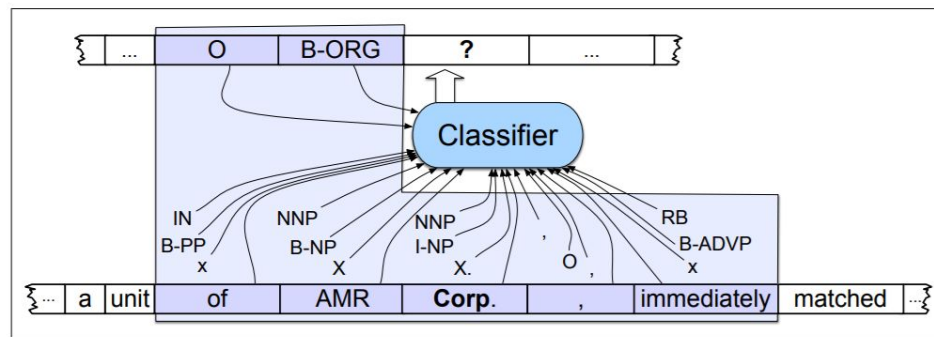


Figure 21.7 Named entity recognition as sequence labeling. The features available to the classifier during training and classification are those in the boxed area.

Evaluation of Named Entity Recognition

Precision

Recall

F-measure

Examples

iPavlov: <https://demo.ipavlov.ai/#en>

Flair: <https://github.com/zalandoresearch/flair>

Relation Extraction

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**].

Relation Extraction

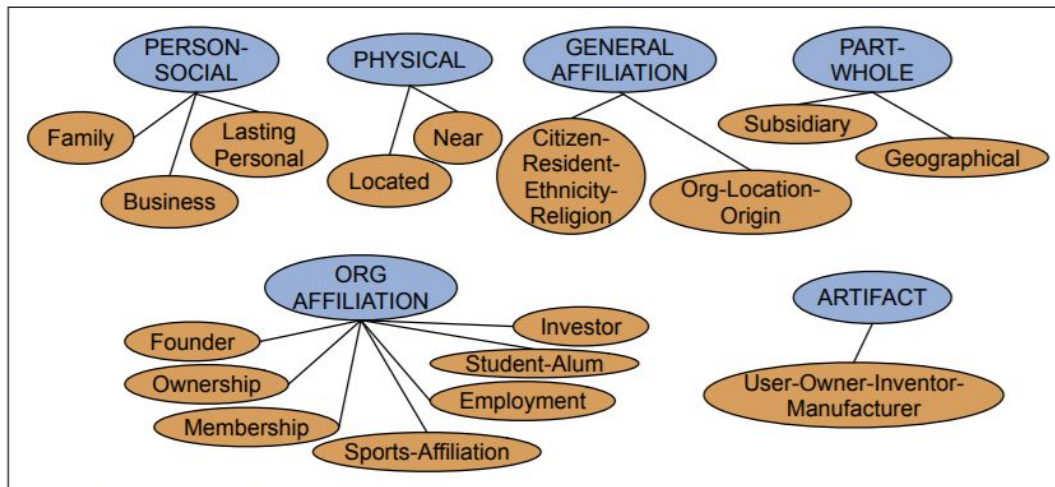


Figure 21.8 The 17 relations used in the ACE relation extraction task.

Relations	Types	Examples
Physical-Located	PER-GPE	He was in Tennessee
Part-Whole-Subsidiary	ORG-ORG	XYZ , the parent company of ABC
Person-Social-Family	PER-PER	Yoko 's husband John
Org-AFF-Founder	PER-ORG	Steve Jobs , co-founder of Apple ...

Figure 21.9 Semantic relations with examples and the named entity types they involve.


Relation Extraction

Domain	$\mathcal{D} = \{a, b, c, d, e, f, g, h, i\}$
United, UAL, American Airlines, AMR	a, b, c, d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f, g, h, i
Classes	
United, UAL, American, and AMR are organizations	$Org = \{a, b, c, d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver, and San Francisco are places	$Loc = \{f, g, h, i\}$
Relations	
United is a unit of UAL	$PartOf = \{\langle a, b \rangle, \langle c, d \rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c, e \rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a, f \rangle, \langle a, g \rangle, \langle a, h \rangle, \langle a, i \rangle\}$

Figure 21.10 A model-based view of the relations and entities in our sample text.

Relation Extraction: Sources

- Infoboxes in Wikipedia;
- FreeBase;
- WordNet

Novosibirsk State University	
Новосибирский государственный университет (НГУ)	
	
Type	Public
Established	1959
Affiliation	Specialized Educational Scientific Center
Endowment	21 million rubles
Rector	Mikhail Fedoruk
Academic staff	2000
Students	7,131
Location	Novosibirsk, 630090, Russian Federation  54.846°N 83.094°E
Campus	Urban
Website	english.nsu.ru

Relation Extraction: Algorithms

- hand-written patterns;
- supervised machine learning;
- semi-supervised machine learning

Relation Extraction: Patterns

- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In COLING-92, Nantes, France
- Hearst, M. A. (1998). Automatic discovery of WordNet relations. In Fellbaum, C. (Ed.), WordNet: An Electronic Lexical Database. MIT Press.

Agar is a substance prepared from a mixture of **red algae**, such as **Gelidium**, for laboratory or industrial use.

NP {, NP}* {,} (and or) other NP _H	temples, treasures, 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 21.11 Hand-built lexico-syntactic patterns for finding hypernyms, using { } to mark optionality (Hearst, 1992a, 1998).

PER, **POSITION** of **ORG**:

George Marshall, **Secretary of State** of the **United States**

PER (named|appointed|chose|etc.) **PER** Prep? **POSITION**

Truman appointed **Marshall** **Secretary of State**

PER [be]? (named|appointed|etc.) Prep? **ORG** **POSITION**

George Marshall was named **US** **Secretary of State**

Relation Extraction: Patterns

- + high-precision
- + they can be tailored to specific domains
- they are often low-recall
- it's a lot of work to create them for all possible patterns

Relation Extraction: Supervised Learning

A fixed set of relations and entities is chosen, a training corpus is hand-annotated with the relations and entities, and the annotated texts are then used to train classifiers to annotate an unseen test set.

```
function FINDRELATIONS(words) returns relations  
  
  relations  $\leftarrow$  nil  
  entities  $\leftarrow$  FINDENTITIES(words)  
  forall entity pairs  $\langle e1, e2 \rangle$  in entities do  
    if RELATED?(e1, e2)  
      relations  $\leftarrow$  relations + CLASSIFYRELATION(e1, e2)
```

Figure 21.12 Finding and classifying the relations among entities in a text.

Relation Extraction: Supervised Learning

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Relation Extraction: Supervised Learning

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

Useful word features include

- The headwords of M1 and M2 and their concatenation
Airlines Wagner Airlines-Wagner
- Bag-of-words and bigrams in M1 and M2
American, Airlines, Tim, Wagner, American Airlines, Tim Wagner
- Words or bigrams in particular positions
M2: -1 *spokesman*
M2: +1 *said*
- Bag of words or bigrams between M1 and M2:
a, AMR, of, immediately, matched, move, spokesman, the, unit
- Stemmed versions of the same

Useful named entity features include

- Named-entity types and their concatenation
(M1: *ORG*, M2: *PER*, M1M2: *ORG-PER*)
- Entity Level of M1 and M2 (from the set NAME, NOMINAL, PRONOUN)
M1: *NAME* [*it* or *he* would be *PRONOUN*]
M2: *NAME* [*the company* would be *NOMINAL*]
- Number of entities between the arguments (in this case *1*, for AMR)

Syntactic features:

- Base syntactic chunk sequence from M1 to M2
NP NP PP VP NP NP
- Constituent paths between M1 and M2
NP ↑ NP ↑ S ↑ S ↓ NP
- Dependency-tree paths
Airlines ←_{subj} matched ←_{comp} said →_{subj} Wagner

Relation Extraction: Supervised Learning

- + can get high accuracies with enough hand-labeled training data, if the test set is similar enough to the training set
- labeling a large training set is extremely expensive
- supervised models don't generalize well to different genres

Relation Extraction: Semi-supervised Learning

Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.

All flights in and out of Ryanair's Belgian hub at Charleroi airport were grounded on Friday...

A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had already been affected.



/ [ORG], which uses [LOC] as a hub /
/ [ORG]'s hub at [LOC] /
/ [LOC] a main hub for [ORG] /

Extracting Times

- **Absolute temporal expressions** are those that can be mapped directly to calendar dates, times of day, or both.
- **Relative temporal expressions** map to particular times through some other reference point (as in *a week from last Tuesday*).
- **Durations** denote spans of time at varying levels of granularity (seconds, minutes, days, weeks, centuries etc.).

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Figure 21.17 Examples of absolute, relational and durational temporal expressions.

Extracting Times

The temporal expression recognition task consists of finding the start and end of all of the text spans that correspond to such temporal expressions.

- **Rule-based approaches** to temporal expression recognition use cascades of automata to recognize patterns at increasing levels of complexity;
- **Sequence-labeling approaches** follow the same IOB scheme used for named entity tags:

A fare increase initiated last week by UAL Corp's...

O O O O B I O O O

Extracting Times: Difficulties

1984 tells the story of Winston Smith...

...U2's classic ***Sunday*** Bloody ***Sunday***

Extracting Times: Temporal Normalization

Temporal normalization is the process of mapping a temporal expression to either a specific point in time or to a duration. Points in time correspond to calendar dates, to times of day, or both. Durations primarily consist of lengths of time but may also include information about start and end points. Normalized times are represented with the VALUE attribute from the ISO 8601 standard for encoding temporal values.

See:

- ISO8601 (2004). Data elements and interchange formats— information interchange—representation of dates and times. Tech. rep., International Organization for Standards (ISO).
- Ferro, L., Gerber, L., Mani, I., Sundheim, B., and Wilson, G. (2005). Tides 2005 standard for the annotation of temporal expressions. Tech. rep., MITRE.
- Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D. S., Ferro, L., and Lazo, M. (2003). The TIMEBANK corpus. In Proceedings of Corpus Linguistics 2003 Conference, pp. 647–656. UCREL Technical Paper number 16.

Extracting Events and their Times

The task of **event extraction** is to identify mentions of events in texts.

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT 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 [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Extracting Events and their Times

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character level suffixes for nominalizations (e.g., <i>-tion</i>)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

Figure 21.23 Features commonly used in both rule-based and statistical approaches to event detection.

Temporal Ordering of Events

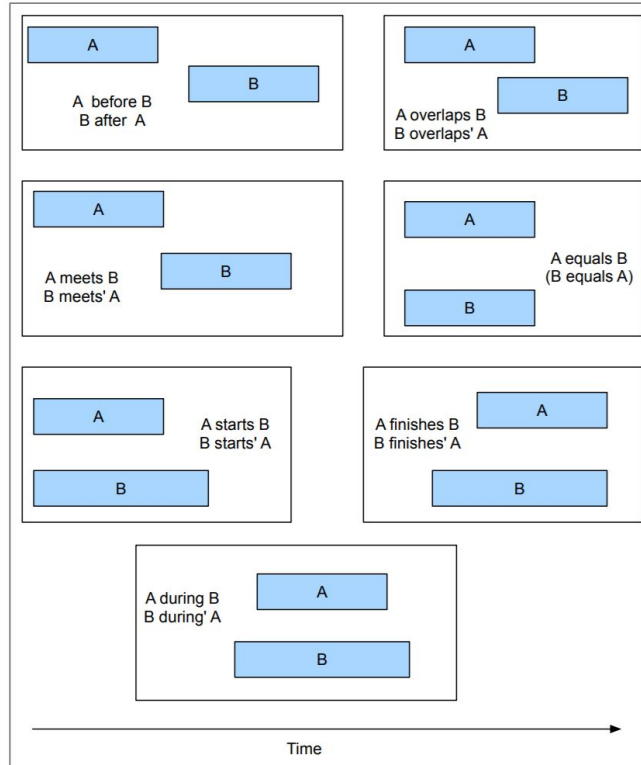


Figure 21.24 Allen's 13 possible temporal relations.

Template Filling

Many texts contain reports of events, and possibly sequences of events, that often correspond to fairly common, stereotypical situations in the world. These abstract scripts situations or stories, related to what have been called **scripts**, consist of prototypical sequences of sub-events, participants, and their roles. In their simplest form, such scripts can be represented as **templates** consisting of fixed sets of **slots** that take as values slot-fillers belonging to particular classes. The task of **template filling** is to find documents that invoke particular scripts and then fill the slots in the associated templates with fillers extracted from the text.

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	\$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES