Information Extraction

Núria Castell Ariño

Definition

- Information extraction (IE) is the automated retrieval of specific (pre-specified) information related to a selected topic from a body or bodies of text.
- IE tools: to obtain information from text documents, databases, websites or multiple sources.
- IE is the task of automatically **extracting structured information** from unstructured, semi-structured and structured machine-readable documents.
- Usually IE is used in natural language processing to extract structured from unstructured text.
- Multimedia documents: automatic annotation and content extraction out of images/audio/video could be seen as IE.

IE tools successfully solve challenges related to

content management and knowledge discovery in the areas of:

- **Business intelligence** (for enabling analysts to gather structured information from multiple sources)
- **Financial investigation** (for analysis and discovery of hidden relationships)
- **Scientific research** (for automated references discovery or relevant papers suggestion)
- **Media monitoring** (for mentions of companies, brands, people)
- **Healthcare records management** (for structuring and summarizing patients records)
- **Pharma research** (for drug discovery, adverse effects discovery and clinical trials automated analysis)

Main tasks involved in extracting structured information from unstructured texts

- Pre-processing of the text – the text is prepared for processing with the help of computational linguistics tools such as tokenization, sentence splitting, morphological analysis, etc.
- Finding and classifying concepts – the mentions of people, things, locations, events and other pre-specified types of concepts are detected and classified.
- Connecting the concepts – this is the task of identifying relationships between the extracted concepts.
- Unifying – this subtask is about presenting the extracted data into a standard form.
- Getting rid of the noise – this subtask involves eliminating duplicate data.
- Enriching your knowledge base – the extracted knowledge is ingested in your database for further use.

Specific tasks

- Named Entity Recognition (NER)
  - coreference resolution
- Relation Extraction
- Event Extraction
  - event coreference
- Temporal expressions
  - normalization
- Template Filling

News exemple

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 (NER)

Task: Find each mention of a named entity in the text and label its type.

Named Entity: Anything that can be referred to with a proper name. Extended to non physical entities such as dates, times, temporal expressions, numerical expressions,…

Named Entity types:

- generic: people, places, organizations, times,…

- application dependent: names of drugs, diseases, symptoms,… (pharmacology)

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].

Generic Named Entity types

Type	Tag	Sample Categories
People	PER	Individuals, fictional characters, small groups
Organization	ORG	Companies, agencies, political parties, religious groups, sports teams
Location	LOC	Physical extents, mountains, lakes, seas
Geo-Political Entity	GPE	Countries, states, provinces, counties
Facility	FAC	Bridges, buildings, airports
Vehicles	VEH	Planes, trains, and automobiles

Type	Example
People	*Turing* is often considered to be the father of modern computer science.
Organization	The *IPCC* said it is likely that future tropical cyclones will become more intense.
Location	The *Mt. Sanitas* loop hike begins at the base of *Sunshine Canyon*.
Geo-Political Entity	*Palo Alto* is looking at raising the fees for parking in the University Avenue district.
Facility	Drivers were advised to consider either the *Tappan Zee Bridge* or the *Lincoln Tunnel*.
Vehicles	The updated *Mini Cooper* retains its charm and agility.

NER useful for

- Extraction of events and relationship between participants
- Sentiment analysis
- Question/Answering
- Linking documents

NER difficulties

Identification of segments of text as

proper named entities and type association

- Ambiguity of segmentation
- Type ambiguity

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Facility
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

[$_{PERS}$ 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 [$_{FAC}$ Washington] had proved to be a leaky ship, every passage I made...

NER algorithms

- Word-by-word sequence labelling task
- A sequence classifier is trained to label the tokens in a text with tags that indicate the presence of particular kinds of named entities.
- Three standard families of algorithms for NER tagging:
  - feature based (MEMM/CRF)
  - neural (bi-LSTM)
  - rule-based

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

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

Feature Based Algorithm for NER

- MEMM (Maximum entropy Markov models)
- CRF (Conditional Random Fields)

https://nlp.stanford.edu/software/CRF-NER.html

https://nlp.stanford.edu/software/crf-faq.shtml

Feature Based Algorithm for NER

Typical features

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

Feature Based Algorithm for NER

The first approach is to extract features and train an MEMM or CRF sequence Model.

Many unknown words are in fact named entities.

Word shape features are thus particularly important in the context of NER.

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.

word-shape (L’Hospitalet) = X’Xxxxxxxxx

short-word-shape= X’Xx

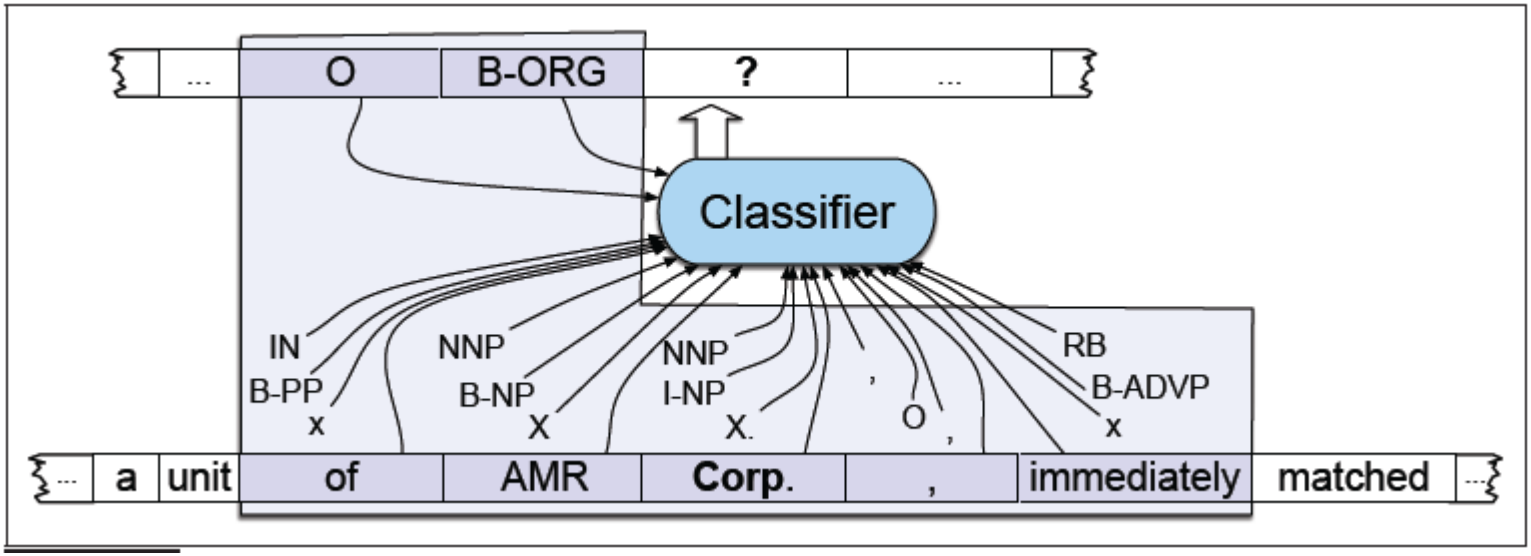
Feature Based Algorithm for NER

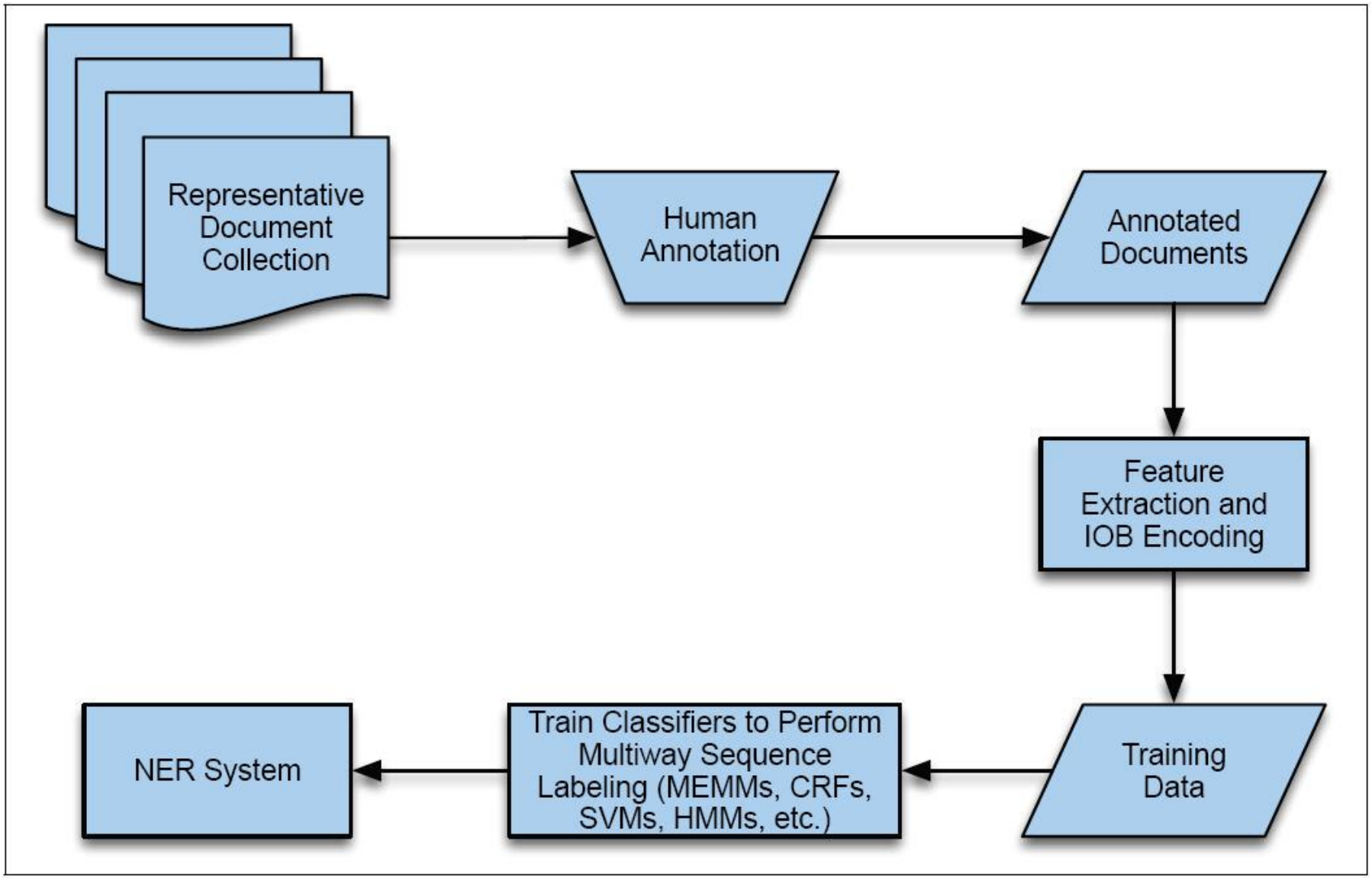
Gazetteer: list of place names, corporations, commercial products, first names, surnames,...

Difficult to create and maintain.

Usefulness depending on application, language, media,...

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





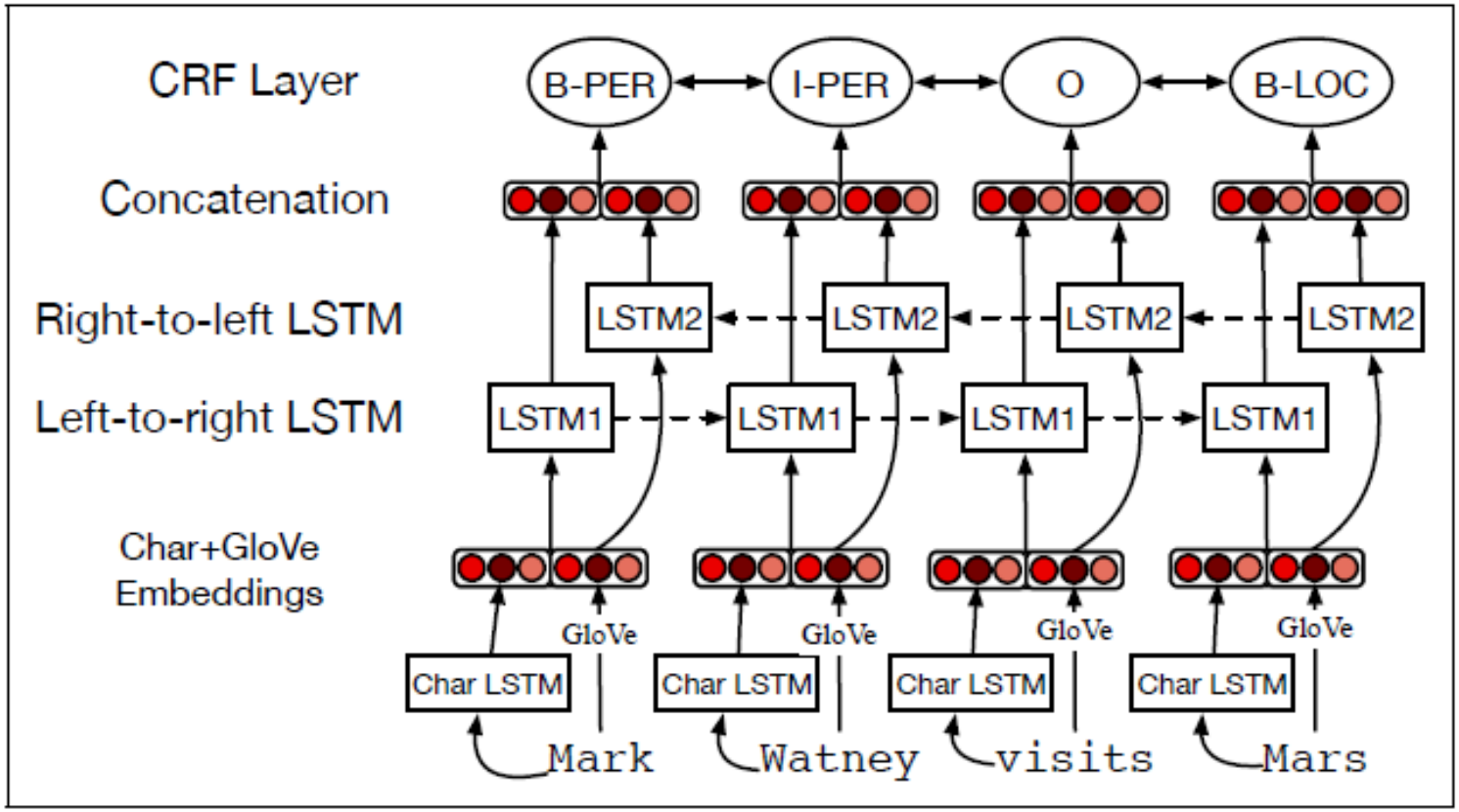
Neural Algorithm for NER

Based on bi-LSTM networks

- Long short-term memory networks
- A left-to-right LSTM and a right-to-left LSTM

For Name Entity tagging a CRF layer is normally used on top of the bi-LSTM output

Neural Algorithm for NER



After (Lample et al., 2016)

Rule Based NER

Commercial approaches to NER: based on pragmatic combinations of lists and rules, with some smaller amount of supervised machine learning

Exemple: IBM System T

Rule Based NER

Common approach

- First, use high-precision rules to tag unambiguous entity mentions.
- Then, search for substring matches of the previously detected names.
- Consult application-specific name lists to identify likely name entity mentions from the given domain.
- Finally, apply probabilistic sequence labeling techniques that make use of the tags from previous stages as additional features.

Evaluation of NER

The usual metrics:

Recall: ratio of the number of correctly labeled responses to the total that should have been labeled

Precision: ratio of the number of correctly labeled responses to the total labeled

F_1: the harmonic mean of the two

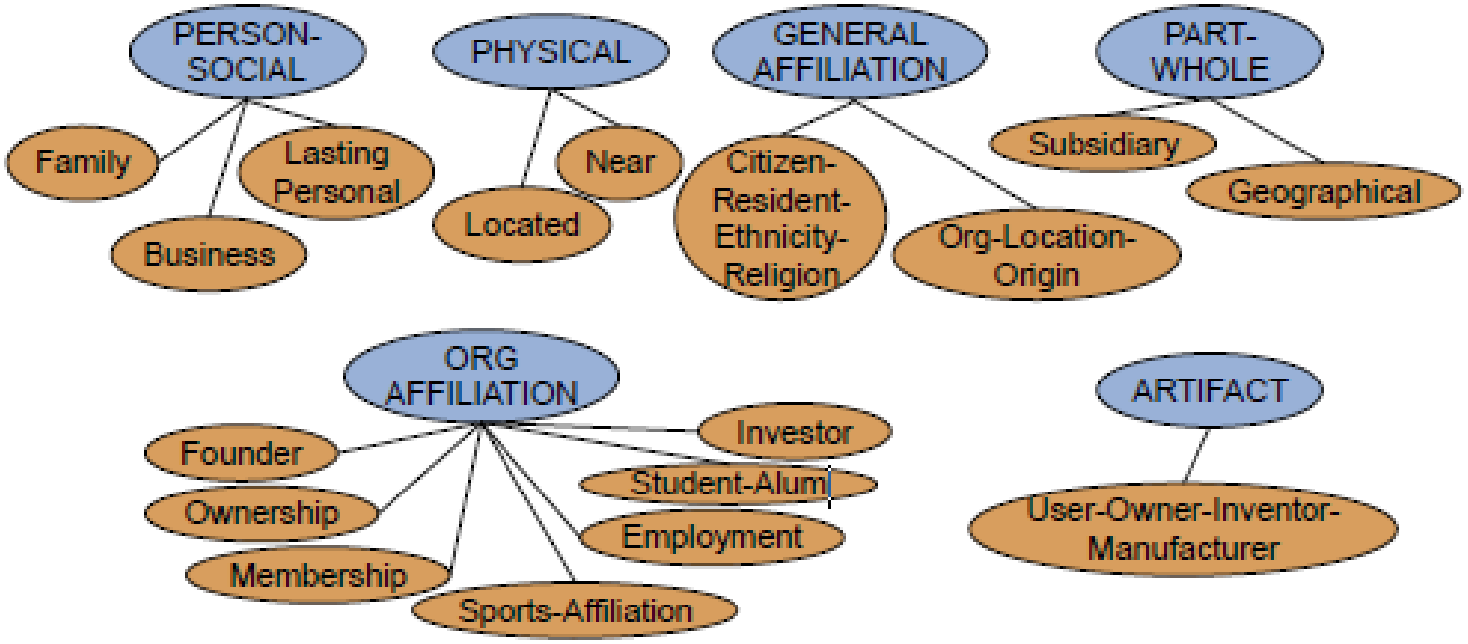
Responses: entities (no words)

Relation Extraction

After NER we need to detect relationships that exist among the detected entities.

RE is an association task: checking if groups of entities are instances of a relation.

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].



The 17 relations used in the ACE relation extraction task.
(Automatic Content Extraction)

Relations		Examples	Types
Affiliations			
	Personal	*married to, mother of*	PER → PER
	Organizational	*spokesman for, president of*	PER → ORG
	Artifactual	*owns, invented, produces*	(PER | ORG) → ART
Geospatial			
	Proximity	*near, on outskirts*	LOC → LOC
	Directional	*southeast of*	LOC → LOC
Part-Of			
	Organizational	*a unit of, parent of*	ORG → ORG
	Political	*annexed, acquired*	GPE → GPE

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...

Domain	$\mathscr{D} = \{a,b,c,d,e,f,g,h,i\}$
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	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Sets of relations

- Per domain:
  - UMLS, the Unified Medical Language System from the US National Library of Medicine
- Wikipedia offers a large supply of relations, drawn from *infoboxes*, structured tables associated with certain Wikipedia articles.
- WordNet or other ontologies offer useful ontological relations that express hierarchical relations between words or concepts.
  - WordNet has the is-a or hypernym relation between classes.

Relation Extraction algorithms

Main classes of algorithms for RE:

- hand-written patterns
- supervised machine learning
- semi-supervised via bootstrapping
- semi-supervised via distant supervision
- unsupervised

Relation Extraction using patterns

The earliest and still common algorithm for relation extraction is lexico-syntactic patterns, first developed by M.A.Hearst (1992).

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

Hand-built lexico-syntactic patterns for finding hypernyms, { } mark optionality

Relation Extraction using patterns

Modern versions of the pattern-based approach extend it by adding named entity constraints.

“Who holds what office in which organization?”

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 using patterns

Hand-built patterns advantages:

- high-precision
- they can be tailored to specific domains.

Hand-built patterns disadvantages:

- often low-recall
- lot of work to create them for all possible patterns.

Relation Extraction via supervised learning

Scheme:

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

Required: lots a of labeled data

Relation Extraction via supervised learning

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

Finding and classifying the relations among entities in a text.

Classifiers: RELATED?, CLASSIFYRELATION
Use of positive and negative examples.
Techniques: logic regression, neural networks, SVM,..

Semisupervised Relation Extraction via Bootstrapping

We have a few high-precision **seed patterns**, or perhaps a few **seed tuples**. That's enough to bootstrap a classifier.

Bootstrapping proceeds by taking the entities in the seed pair, and then finding sentences (on the web, or whatever dataset we are using) that contain both entities.

From all such sentences, we extract and generalize the context around the entities to learn new patterns.

Semisupervised Relation Extraction via Bootstrapping

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples ← Gather a set of seed tuples that have relation *R*
iterate
sentences ← find sentences that contain entities in *tuples*
patterns ← generalize the context between and around entities in *sentences*
newpairs ← use *patterns* to grep for more tuples
newpairs ← *newpairs* with high confidence
tuples ← *tuples* + *newpairs*
return tuples

Bootstrapping from seed entity pairs to learn relations.

New tuples receive **confidence values** to avoid erroneous patterns.

Distant Supervision for Relation Extraction

The distant supervision method of Mintz et al. (2009) combines the advantages of bootstrapping supervision with supervised learning.

Distant supervision uses a large database to acquire a huge number of seed examples, creates lots of noisy pattern features from all these examples and then combines them in a supervised classifier.

Distant Supervision for Relation Extraction

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*
  foreach tuple (*e1*,*e2*) of entities with relation *R* in *D*
    sentences ← Sentences in *T* that contain *e1* and *e2*
    f ← Frequent features in *sentences*
    observations ← observations + new training tuple (*e1*, *e2*, *f*, *R*)
C ← Train supervised classifier on *observations*
return *C*

A neural classifier might not need to use the feature set f .

Unsupervised Relation Extraction

The goal of unsupervised relation extraction is to extract relations from the web when we have no labeled training data, and not even any list of relations.

This task is often called open information extraction or Open IE.

Evaluation of Relation Extraction

Supervised RE systems: using test sets with human-annotated, gold-standard relations and computing **precision**, **recall** and **F-measure**.

Semi-supervised and unsupervised systems: more difficult because they extract new relations. It is possible to approximate **precision** using a random sample of relations from the output and having a human check the accuracy of each relation.

Extracting Times

Times and dates are very important in

- question/answering
- calendar and personal assistant applications
- events relations

First, extraction of temporal expressions

Second, temporal normalisation

Temporal Expression Extraction

Absolute points in time

- Mapped directly to calendar dates, times of day

Relative times

- Map to particular times using other reference point

Durations

- Spans of time at different levels of granularity (seconds, days, years, centuries,….)

Sets of these

Temporal Expression Extraction

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

Temporal expressions are grammatical constructions that have temporal lexical triggers as their heads. Full temporal expressions consist of their phrasal projections.

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Temporal Expression Extraction

TE recognition task: finding the start and end of the text spans that corresponds to a TE.

- Rule Based approaches (POS tagging + trigger words)
- Sequence-labelling approaches (IOB scheme)

TE recognizers are evaluated with recall, precision and F-measures.

Temporal Normalization

The process of mapping a TE to either

- a specific point (calendar dates, times of day, both)
- a duration (lengths of time, start and end points)

Normalized times are represented with the VALUE attribute from ISO 8601 standard.

Unit	Pattern	Sample Value
Fully specified dates	YYYY-MM-DD	1991-09-28
Weeks	YYYY-nnW	2007-27W
Weekends	PnWE	P1WE
24-hour clock times	HH:MM:SS	11:13:45
Dates and times	YYYY-MM-DDTHH:MM:SS	1991-09-28T11:00:00
Financial quarters	Qn	1999-Q3

Some ISO patterns

Temporal Normalization

Fully qualified date expressions are quite rare in real texts.

Temporal expressions in news articles are incomplete and only implicitly anchored, often with respect to the dateline of the article (document's temporal anchor).

Even simple expressions are difficult to normalize (ambiguity).

Extracting Events and their Times

Event extraction: identify mentions of events.

Event mention: any expression denoting an event or state that can be assigned to a particular point, or interval, in time.

Event detection: rule-based and machine learning approaches.

Extracting Events and their Times

[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 [EVENTapplies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

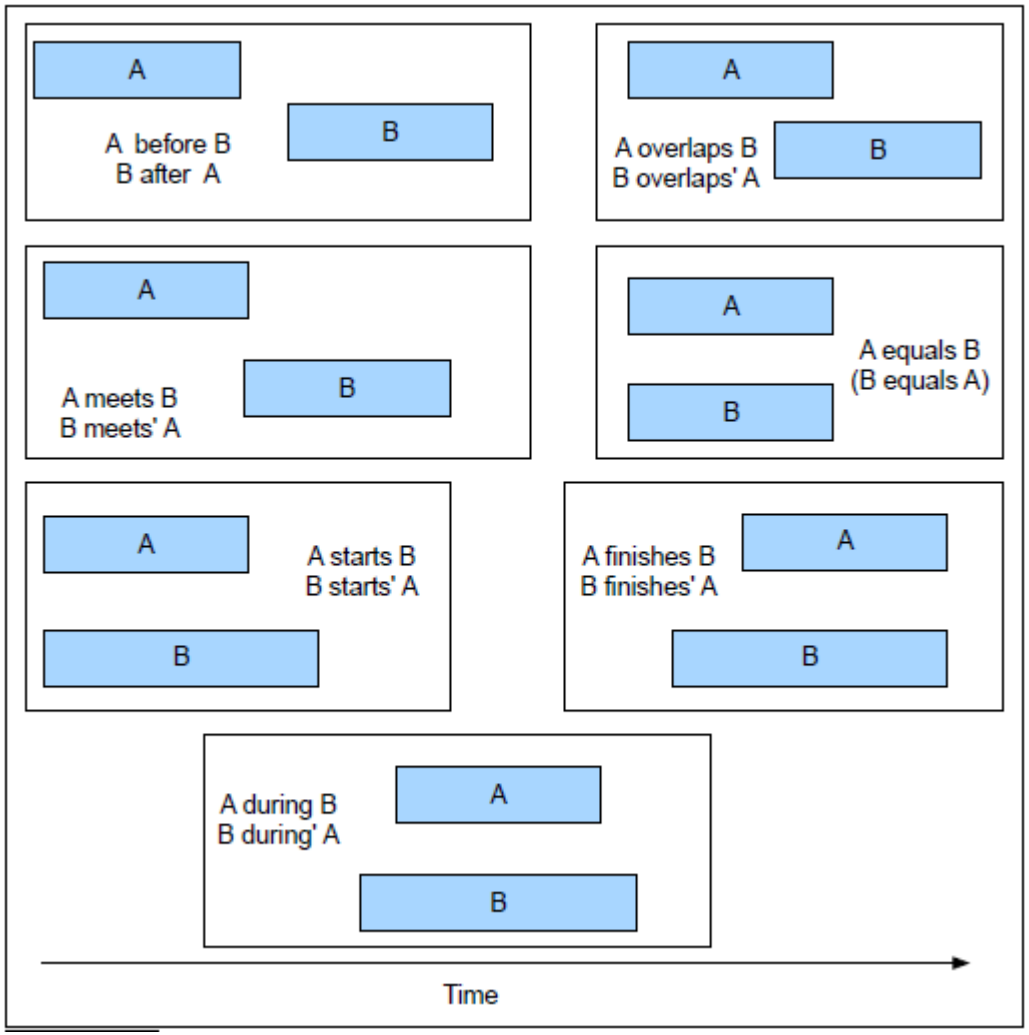
Usual: Verbs/events Also: Noun Phrases/events

Temporal Ordering of Events

- After detection of events and temporal expressions, we could fit the events into a complete timeline.
- This ambitious task is the subject of considerable current research but is beyond the capabilities of current systems.
- A somewhat simpler, but still useful, task is to impose a partial ordering on the events and temporal expressions mentioned in a text.

Resource: TimeBank corpus

Temporal Relations from J. Allen (1984)



Template Filling

Many texts contain reports of events, and possibly sequences of events, that often correspond to fairly common, stereotypical situations in the world.

scripts: consist of prototypical sequences of sub-events, participants, and their roles.

Simple representation: a **template** with a fixed set of slots. The values slot-fillers belong to particular classes.

Template Filling

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 (or inferred concepts).

- Machine Learning approaches (simple templates)
- Older system (FASTUS): cascades of finite-state transducers (complex templates)

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 (or inferred concepts).

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

Template Filling

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

Tie-up-1		Activity-1:	
RELATIONSHIP	tie-up	COMPANY	Bridgestone Sports Taiwan Co.
ENTITIES	Bridgestone Sports Co.	PRODUCT	iron and "metal wood" clubs
	a local concern	START DATE	DURING: January 1990
	a Japanese trading house		
JOINT VENTURE	Bridgestone Sports Taiwan Co.		
ACTIVITY	Activity-1		
AMOUNT	NT$20000000		

Example from FASTUS system (Hobbs et al., 1997).

A slot filler is itself a template.

System inspired by the results of MUC (Message Understanding Conferences, 1991-1995).

After that period: Machine learning approaches, Neural algorithms

References

- "Speech and Language Processing"
  Jurafsky, Dan and Martin, James H.
  Version 2: http://cataleg.upc.edu/record=b1508835~S1*cat
  Draft Version 3: https://web.stanford.edu/~jurafsky/slp3/

- https://ontotext.com/knowledge-hub/

- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In NAACL HLT 2016.

- Allen, J. (1984). Towards a general theory of action and time. Artificial Intelligence, 23(2), 123–154.

- Hobbs, J. R., Appelt, D. E., Bear, J., Israel, D., Kameyama, M., Stickel, M. E., and Tyson, M. (1997). FASTUS: A cascaded finite-state transducer for extracting information from natural-language text. In Roche, E. and Schabes, Y. (Eds.), Finite-State Language Processing, pp. 383–406.MIT Press.