Information extraction[1]

Victor Kitov

v.v.kitov@yandex.ru

[1] With materials used from "Speech and Language Processing", D. Jurafsky and J. H. Martin.

Table of Contents

Intro

- Information extraction - turn unstructred text into structured data
- Named entity recognition (NER) - find named entities in text and label their types
- Named entities:
    - typically: people, places, organizations
    - additionally: dates, times, prices.
    - custom: names of genes, names of college courses.
- Coreference resolution - group named entities relating to the same real-world entity.

Applications of NER

- Extract sentiment towards particular product
- Extract relations between objects
- Link text to structured data (e.g. from Wikipedia)
- Question answering

Example

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.

Example

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

Types of entities

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**.

Problems on NER

- Problems in NER:
  - Find span of text covering given entity
  - Detect entity type
- Entity with the same name can correspond to different types:

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

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

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

Solution to NER

- Solve NER with sequence labelling classification (using MEMM, conditional random fields)
  - classifier perfroms both segmentation and type identification
- BIO tagging: B=beginning, I-inside previous tag, O=outside
  - 2*n+1 classes for n entity types
  - n+1 classes for IO tagging
    - can't separate 2 neighbouring types

Example

Words	BIO 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

[$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said.

Features for NER classification

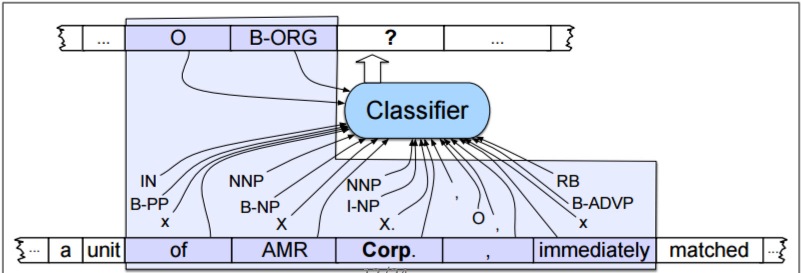
identity of w_i
identity of neighboring words
part of speech of w_i
part of speech of neighboing words
base-phrase syntactic chunk label of w_i and neighboring words
presence of w_i in a **gazeteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from 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

Feature descriptions

- For word L'Occitane the following features will be generated:
  - prefix(w) = L
  - prefix(w) = L'
  - prefix(w) = L'O
  - prefix(w) = L'Oc
  - suffix(w) = ane
  - suffix(w) = ne
  - suffix(w) = e
  - word-shape(wi) = X'Xxxxxxxx
  - short-word-shape(wi) = X'Xx
- gazetteer - list of places, names, corporations, commercial products.

Classifier visualization

Word	POS	Chunk	Short shape	Label
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



Practical NER algrotithm architecture

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

Intuition:

- once entity is named, its shortened name will be used as well.
- presense of one entity is a good feature to detect other entities

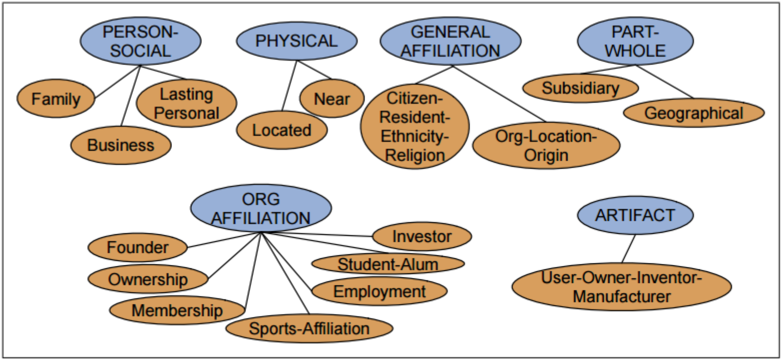
Table of Contents

Relation extraction

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

- Extracted relations:
    - [Tim Wagner] is a spokesman for [American Airlines]
    - [American Airlies] is a unit of [AMR Corp].
    - [United] is a unit of [UAL Corp.]
- Relations can be represented with RDF triples
    - e.g.: <Golden Gate Park> <location> <San Francisco>

Relation examples



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**...

Datasets with relations

Datasets with relations:

- wikipedia infoboxes
  - e.g. for Stanfrord we have state = "California", president = "John L. Hennessy"
- DBpedia, Freebase - ontologies, derived from wikipedia
- WordNet
  - hypernym relations: giraffe is mammal
  - instance-of relations: Moscow is a city

Relations extraction

Relations extraction approaches:

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

Relation extraction with patterns

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

Useful to add named entity specifications in rules:

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

Hand-built patterns have high precision and low recall.

Supervised relation extraction

```
FOR EACH sentence IN text:
   entities=FindEntities(sentence)
   for all entity pairs <e1,e2> in entities:
      if Related(e1,e2)
         relations=relations∪ClassifyRelation(e1,e2)
```

Features

- Example: **American Airlines**, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said.
- Denote M1=American Airlines, M2=Tim Wagner

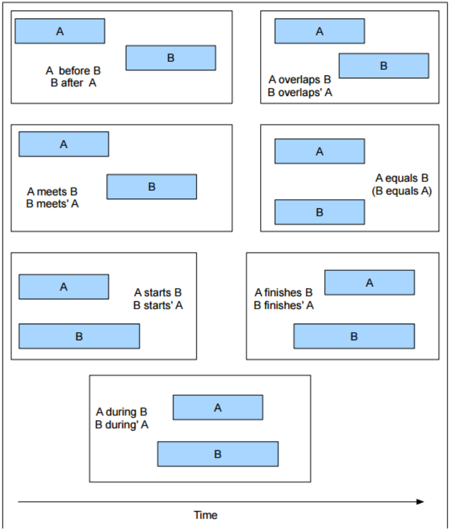
Useful features:

M1 headword	*airlines*
M2 headword	*Wagner*
Word(s) before M1	NONE
Word(s) after M2	*said*
Bag of words between	*{a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman*
M1 type	ORG
M2 type	PERS
Concatenated types	ORG-PERS
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base phrase path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Other approaches

- Semisupervized (bootstrapping)
  - semantic drift
  - confidence evaluation
- Distant Supervision
  - use existing ontology
  - find occurences in text
  - e.g. Wikipedia infoboxes and texts
  - need to add negative class representatives

Timing of events



Template filling

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