INF5830 – 2018 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 9, 17.10

Today:

- Chunking
- Named Entity Recognition
- Relation detection

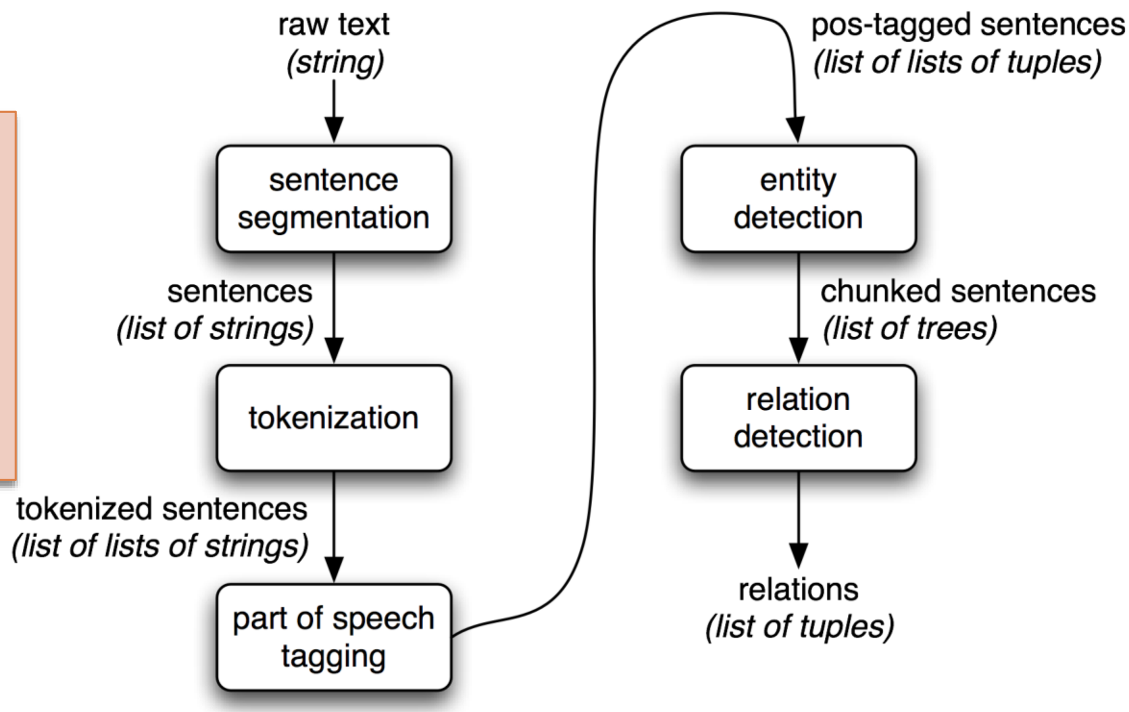
IE basics

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

- Bottom-Up approach
- Start with unrestricted texts, and do the best you can
- The approach was in particular developed by the Message Understanding Conferences (MUC) in the 1990s
- Select a particular domain and task

Steps

(Some approaches do these steps in a different order – or simultaneously)

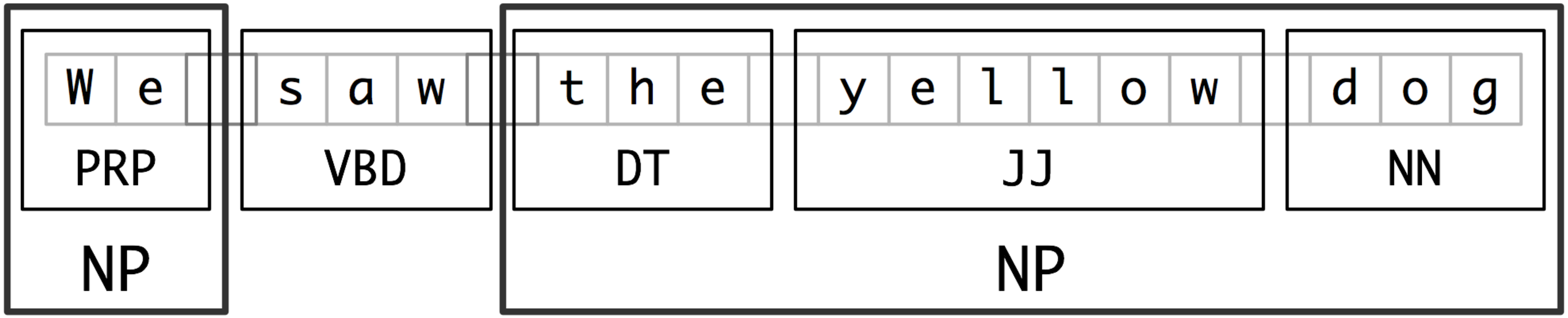


From NLTK

5

Chunking

Next steps



- Chunk together words to phrases

NP-chunks

[The/DT market/NN] for/IN
[system-management/NN software/NN]
for/IN [Digital/NNP]
['s/POS hardware/NN] is/VBZ
fragmented/JJ enough/RB that/IN
[a/DT giant/NN] such/JJ as/IN
[Computer/NNP Associates/NNPS]
should/MD do/VB well/RB there/RB ./.

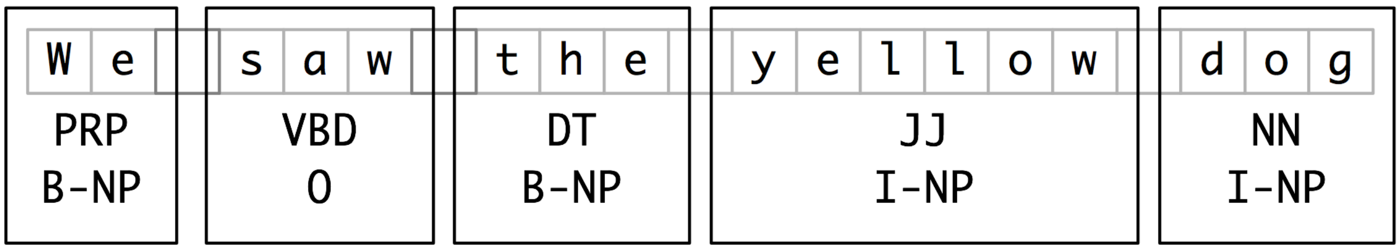
- Exactly what is an NP-chunk?
- It is an NP
- But not all NPs are chunks
- Flat structure: no NP-chunk is part of another NP chunk
- Maximally large
- Opposing restrictions

Regular Expression Chunker

- Input POS-tagged sentences
- Use a regular expression over POS to identify NP-chunks
- NLTK example:
- It inserts parentheses

```
grammar = r"""
  NP: {<DT|PP\$>?<JJ>*<NN>}
      {<NNP>+}
"""
```

IOB-tags

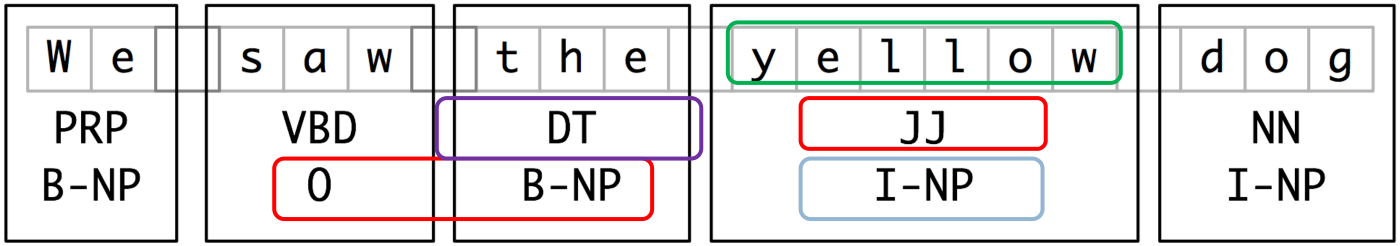


- Properties
  - One tag per token
  - Unambiguous
  - Does not insert anything in the text itself

Sequence labelling

- The IOB schema can be applied to many different tasks
- For example,
  - sentence segmentation
  - Tokenization
- can be considered IOB-labelling over characters
  - Evang et al (2013) consider the two tasks simultaneously

Assigning IOB-tags



- The process can be considered a form for tagging
  - POS-tagging: Word to POS-tag
  - IOB-tagging: POS-tag to IOB-tag
- But one may in addition use additional features, e.g. words
- Can use various types of classifiers
  - NLTK uses a MaxEnt Classifier

Evaluating (IOB-)chunkers

- cp = nltk.RegexpParser("")
- test_sents = conll ('test', chunks=['NP'])
- IOB Accuracy: 43.4%
- Precision: 0.0%
- Recall: 0.0%
- F-Measure: 0.0%

- What do we evaluate?
  - IOB-tags? or
  - Whole chunks?
  - Yields different results
- For IOB-tags:
  - Baseline:
    - majority class O,
    - yields > 33%
- Whole chunks:
  - Which chunks did we find?
  - Harder
  - Lower numbers

Evaluating (IOB-)chunkers

- cp = nltk.RegexpParser("")
- test_sents = conll ('test', chunks=['NP'])
- IOB Accuracy: 43.4%
- Precision: 0.0%
- Recall: 0.0%
- F-Measure: 0.0%

>> cp = nltk.RegexpParser(r"NP: {<[CDJNP].*>+}")

- IOB Accuracy: 87.7%
- Precision: 70.6%
- Recall: 67.8%
- F-Measure: 69.2%

Named Entity Recognition

Named entities

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:
  - Anything you can refer to by a proper name
  - i.e. not all NP (chunks):
    - *high fuel prices*
  - Maybe longer than NP than just chunk:
    - *Bank of America*
- Find the phrases
- Classify them

Types of NE

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

- The set of types vary between different systems
- Which classes are useful depend on application

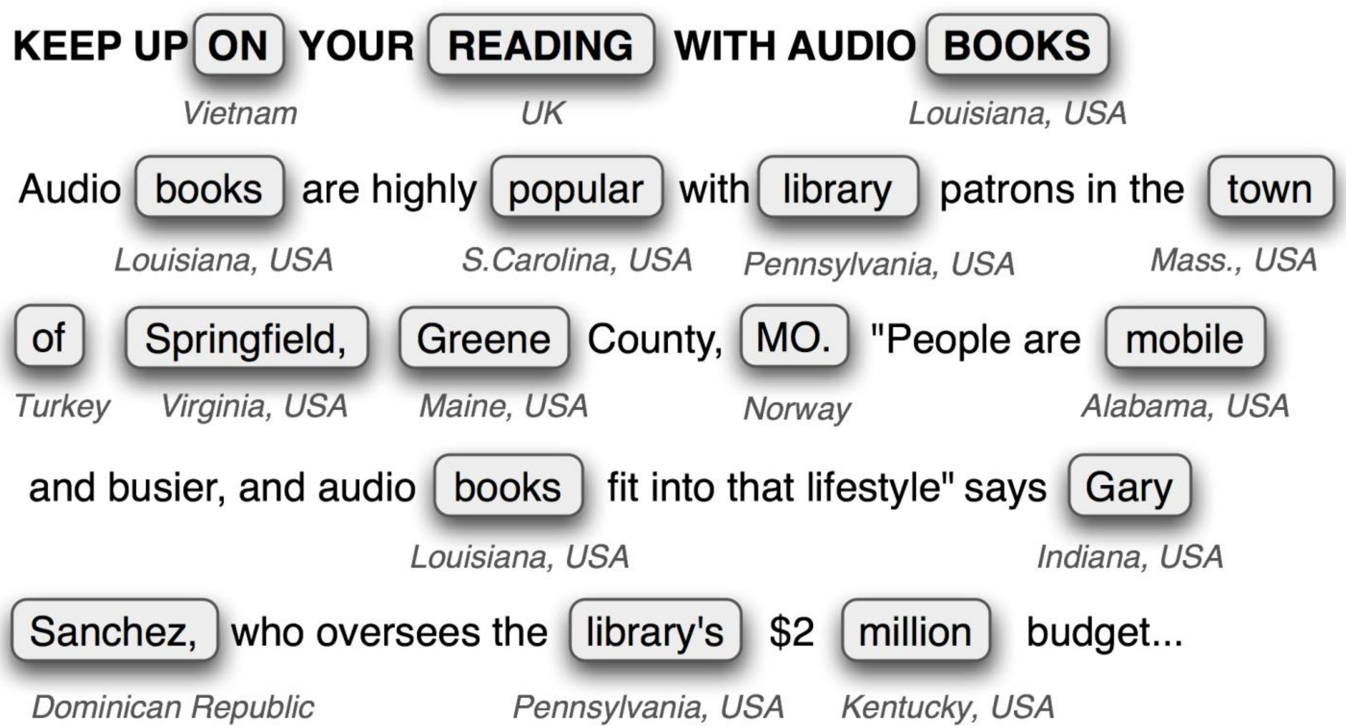
Ambiguities

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...

Gazetteer

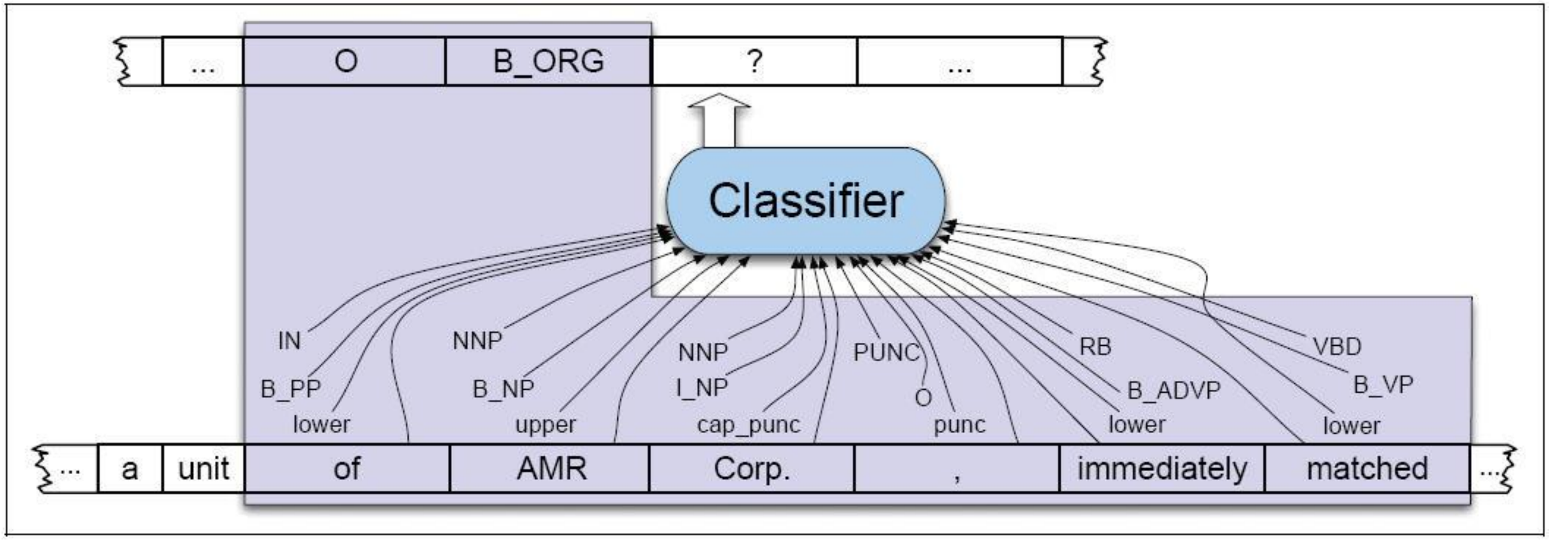
- Useful: List of names, e.g.
  - Gazetteer: list of geographical names
- But does not remove all ambiguities



Representation (IOB)

Features				Label
American	NNP	B_{NP}	cap	B_{ORG}
Airlines	NNPS	I_{NP}	cap	I_{ORG}
,	PUNC	O	punc	O
a	DT	B_{NP}	lower	O
unit	NN	I_{NP}	lower	O
of	IN	B_{PP}	lower	O
AMR	NNP	B_{NP}	upper	B_{ORG}
Corp.	NNP	I_{NP}	cap_punc	I_{ORG}
,	PUNC	O	punc	O
immediately	RB	B_{ADVP}	lower	O
matched	VBD	B_{VP}	lower	O
the	DT	B_{NP}	lower	O
move	NN	I_{NP}	lower	O
,	PUNC	O	punc	O
spokesman	NN	B_{NP}	lower	O
Tim	NNP	I_{NP}	cap	B_{PER}
Wagner	NNP	I_{NP}	cap	I_{PER}
said	VBD	B_{VP}	lower	O
.	PUNC	O	punc	O

Classification



- Similar to tagging and chunking
- You will need features from several layers
- Features may include
  - Words, POS-tags, Chunk-tags, Graphical prop.
  - and more (See J&M, 3.ed)

Machine learning methods

- "Word-by word"
  - Logistic regression (MaxEnt)
- Sequence labelling:
  - Conditional random fields
    - Preferred approach until recently
- Lately: Various deep-learning approaches

Relation detection

Goal

- Extract the relations that exist between the (named) entities in the text
- A fixed set of relations (normally)
  - Determined by application:
    - Jeopardy
    - Preventing terrorist attacks
    - Detecting illness from medical record
    - …

- Born_in
- Date_of_birth
- Parent_of

- Author_of
- Winner_of

- Part_of
- Located_in

- Acquire
- Threaten

- Has_symptom
- Has_illness

Examples

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

Methods for relation extraction

1. Hand-written patterns
2. Machine Learning (Supervised classifiers)
3. Semi-supervised classifiers and bootstrapping

Hand-written patterns

- Example: acquisitions
- [ORG]…(buy(s)| bought| aquire(s|d))…[ORG]

- Hand-write patterns like this
- Properties:
  - High precision
  - Will only cover a small set of patterns
  - Low recall
  - Time consuming
- (Also in NLTK, sec 7.6)

Example

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

Figure 18.11 Hand-built lexico-syntactic patterns for finding hypernyms, using {} to mark optionality (Hearst, 1992a, 1998).

2. Supervised classifiers

- A corpus
- A fixed set of entities and relations
- The sentences in the corpus is hand annotated:
  - Entities
  - Relations between them
- Split the corpus into parts for training and testing
- Train a classifier:
  - Choose learner:
    Naive Bayes, Logistic regression (Max Ent), SVM, …
  - Select features

The classification task

function FINDRELATIONS(*words*) **returns** *relations*

relations ← *nil*
entities ← FINDENTITIES(*words*)
forall entity pairs $\langle e1, e2 \rangle$ **in** *entities* **do**
if RELATED?($e1, e2$)
relations ← *relations*+CLASSIFYRELATION($e1, e2$)

Examples of features

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

Entity-based features	
Entity$_1$ type	ORG
Entity$_1$ head	*airlines*
Entity$_2$ type	PERS
Entity$_2$ head	*Wagner*
Concatenated types	ORGPERS
Word-based features	
Between-entity bag of words	{ *a, unit, of, AMR, Inc., immediately, matched, the, move, spokesman* }
Word(s) before Entity$_1$	NONE
Word(s) after Entity$_2$	*said*
Syntactic features	
Constituent path	$NP \uparrow NP \uparrow S \uparrow S \downarrow NP$
Base syntactic chunk path	$NP \rightarrow NP \rightarrow PP \rightarrow NP \rightarrow VP \rightarrow NP \rightarrow NP$
Typed-dependency path	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

Properties

- The bottleneck is the availability of training data
- To hand label data is time consuming
- Mostly applied to restricted domains
- Does not generalize well to other domains

3. Semisupervised, bootstrapping

Patterns:
[ORG]…bought…[ORG]

Pairs:
IBM – AlchemyAPI
Google – YouTube
Facebook - WhatsApp

Relation
ACQUIRE

- If we know a pattern for a relation we can determine whether a pair stands in the relation
- Conversely: If we know that a pair stands in a relationship, we can find patterns that describe the relation

Example

- (IBM, AlchemyAPI): ACQUIRE
- Search for sentences containing IBM and AlchemyAPI
- Results (Web-search, Google, btw. first 10 results):
  - *IBM's Watson makes intelligent acquisition of Denver-based AlchemyAPI* (Denver Post)
  - *IBM is buying machine-learning systems maker AlchemyAPI Inc. to bolster its Watson technology as competition heats up in the data analytics and artificial intelligence fields.* (Bloomberg)
  - *IBM has acquired computing services provider AlchemyAPI to broaden its portfolio of Watson-branded cognitive computing services.* (ComputerWorld)

Example contd.

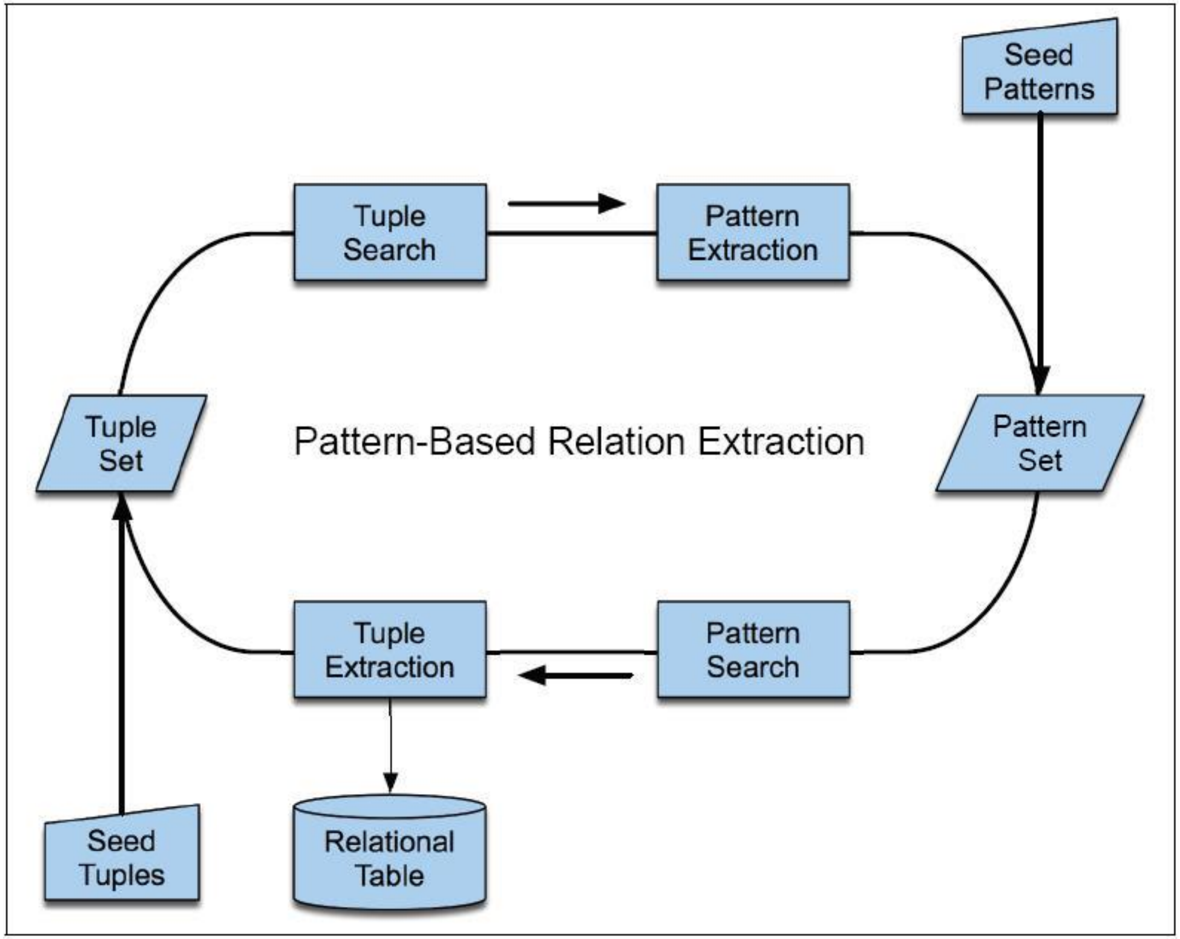
- Extract patterns
  - *IBM's Watson makes intelligent acquisition of Denver-based AlchemyAPI* (Denver Post)
  - *IBM is buying machine-learning systems maker AlchemyAPI Inc. to bolster its Watson technology as competition heats up in the data analytics and artificial intelligence fields.* (Bloomberg)
  - *IBM has acquired computing services provider AlchemyAPI to broaden its portfolio of Watson-branded cognitive computing services.* (ComputerWorld)

Procedure

- From the extracted sentences, we extract patterns
- Use these patterns to extract more pairs of entities that stand in these patterns
- These pairs may again be used for extracting more patterns, etc.

- …*makes intelligent acquisition* …
- … *is buying* …
- … *has acquired* …

Bootstrapping



A little more

- We could
  - either extract pattern templates and searching for these
  - or features for classification and build a classifier
- If we use patterns we should generalize
  - *makes intelligent acquisition* → *(make(s)|made) JJ* acquisition*
- During the process we should evaluate before we extend:
  - Does the new pattern recognize other pairs we know stand in the relation? (Recall)
  - Does the new pattern return pairs that are not in the relation? (Precision)

Evaluating relation extraction

- Supervised methods can be evaluated on each of the examples in a test set.
- For the semi-supervised method:
  - we don't have a test set.
  - we can evaluate the precision of the returned examples

- Beware the difference between
  - Determine for a sentence whether an entity pair is in a particular relation
  - Determine from a text:
    - We may use several occurrences of the pair in the text

Methods for relation extraction

1. Hand-written patterns
2. Machine Learning (Supervised classifiers)
3. Semi-supervised classifiers and bootstrapping

Other methods:

4. Distant supervision
   - Use large knowledge bases as basis for classification
5. Unsupervised (no predefined class of relations)

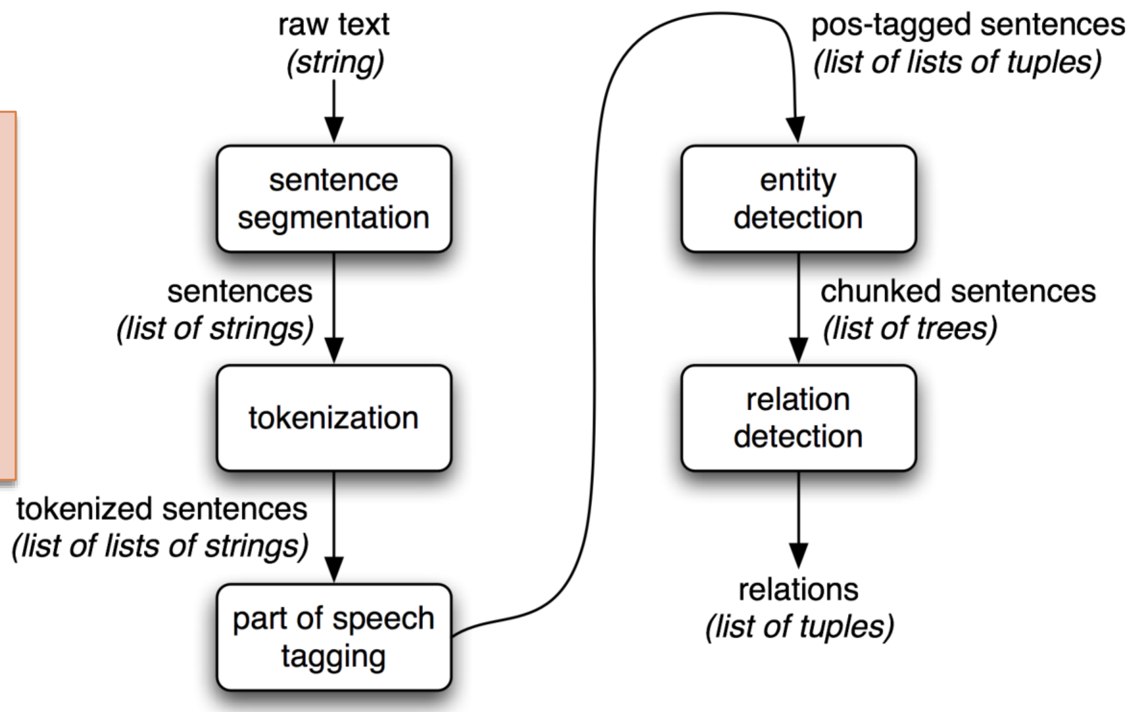
- We will not go into details
- Consider original sources when you want to use it

More fine grained IE

- Tokenization+tagging
- Identifying the "actors"
  - Chunking
  - Named-entity recognition
  - Co-refrence resolution
- Relation detection
- Eventdetection
  - Co-reference resolution of events
- Temporal extraction
- Template filling

Steps

(Some appro-aches do these steps in a different order – or simultaneously)



From NLTK

Some example systems

- Stanford core nlp
  - http://corenlp.run/
- IBM
  - https://www.ibm.com/watson/services/natural-language-understanding/
- For download also
  - SpaCy (Python)
  - OpenNLP
  - GATE (Java)

Summary

- Similarities – and differences – between
  - Tokenization
  - Tagging
  - Chunking
  - Named Entity Recognition

- Relation Extraction
  1. Pattern matching
  2. Supervised machine learned classifier
  3. Bootstrapping