INF5820

Natural Language Processing - NLP

H2009

Jan Tore Lønning

jtl@ifi.uio.no

NER and Relation detection & classification

INF5830

Lecture 14

Nov 9, 2009

Today

- Overview: Information extraction
- (NP-)chunking
- Named entity recognition
- The Constraint Grammar approach
- Relation detection and classification

Information extraction

- Goal: Extract structured data from text
- Track business news, or
- Intelligence news (possible terrorist attacks)
- Similarities to the frames from last week

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

Steps

- The bottom-up approach:
  1. Preproscessing: tokenization, segmentation
  2. Tagging
  3. Chunking
  4. Named entity recognition
  5. Reference resolution
  6. Relation detection and classification
  7. Temporal analysis
  8. Template filling

Steps

- The bottom-up approach:
  1. Preproscessing: tokenization, segmentation
  2. Tagging
  3. Chunking
  4. Named entity recognition
  5. Reference resolution
  6. Relation detection and classification
  7. Temporal analysis
  8. Template filling

Chunking

- Form of shallow parsing
- Flat structures
- Identify (some) phrases

[$_{NP}$ The morning flight] [$_{PP}$ from] [$_{NP}$ Denver] [$_{VP}$ has arrived.]

[$_{NP}$ a flight] [$_{PP}$ from] [$_{NP}$ Indianapolis][$_{PP}$ to][$_{NP}$ Houston][$_{PP}$ on][$_{NP}$ TWA].

- Non-overlapping phrases
- Compromises, cf. PP
- Sometimes only interested in NPs

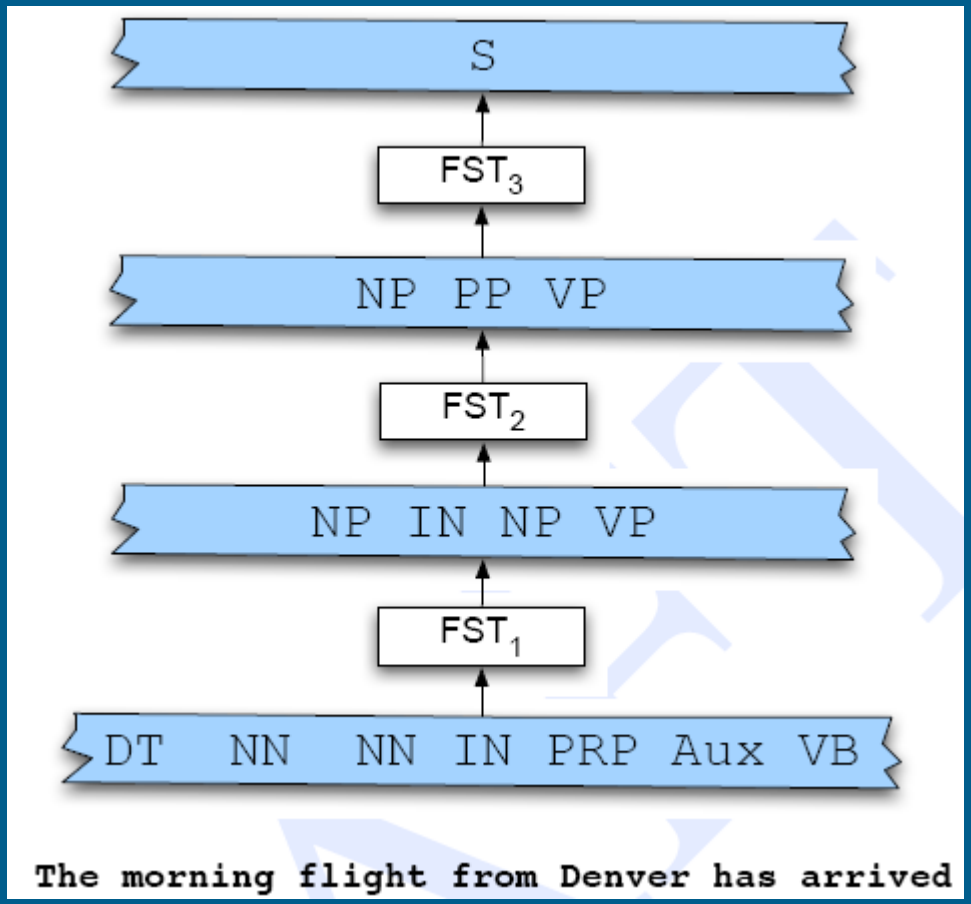
Approach 1: Cascaded FSTs

$NP \rightarrow (Det)\ Noun^*\ Noun$

$NP \rightarrow Proper\text{-}Noun$

$VP \rightarrow Verb$

$VP \rightarrow Aux\ Verb$



2: ML-approaches

- Two tasks:
  - Identify the phrase (beginning-end)
  - Classify the phrase

ML-approaches continued

- The two steps as a tagging task:

The morning flight from Denver has arrived
B_NP I_NP I_NP B_PP B_NP B_VP I_VP

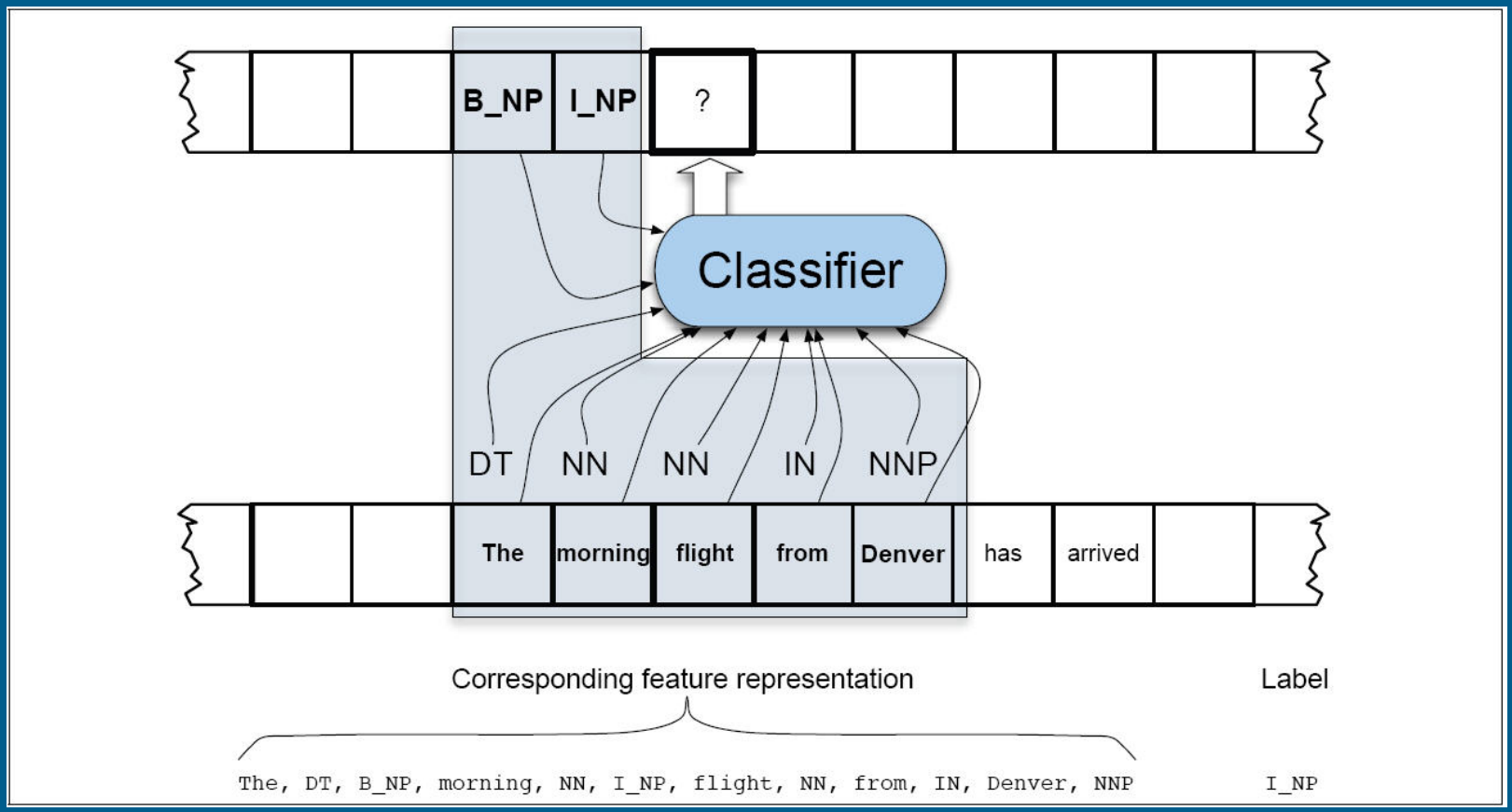
The same sentence with only the base-NPs tagged

The morning flight from Denver has arrived.
B_NP I_NP I_NP O B_NP O O

B_NP : begin NP
I_NP : inside NP
B_VP : begin VP
I_VP: inside VP

O: not part of a phrase
Etc.

ML: Classifier



ML: more

- Training data from Penn treebank
- Evaluation on found chunks
  - compared to test set
  - Recall and precision
    - (typo in hardcover book)

Today

- Overview: Information extraction
- (NP-)chunking
- Named entity recognition
- The Constraint Grammar approach
- Relation detection and classification

Named Entity Classes

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.

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

- Choice of types is/should be application specific

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...

Named entity recognition

- Two tasks:
  - Identify the phrase (beginning-end)
  - Classify the phrase
- Similar to chunking but
  - Different/more fine-grained classification

Words	Label
American	B_{ORG}
Airlines	I_{ORG}
,	O
a	O
unit	O
of	O
AMR	B_{ORG}
Corp.	I_{ORG}
,	O
immediately	O
matched	O
the	O
move	O
,	O
spokesman	O
Tim	B_{PERS}
Wagner	I_{PERS}
said	O
.	O

Features

Feature	Explanation
Lexical items	The token to be labeled
Stemmed lexical items	Stemmed version of the target token
Shape	The orthographic pattern of the target word
Character affixes	Character-level affixes of the target and surrounding words
Part of speech	Part of speech of the word
Syntactic chunk labels	Base-phrase chunk label
Gazetteer or name list	Presence of the word in one or more named entity lists
Predictive token(s)	Presence of predictive words in surrounding text
Bag of words/Bag of N-grams	Words and/or *N*-grams occurring in the surrounding context

Shape	Example
Lower	cummings
Capitalized	Washington
All caps	IRA
Mixed case	eBay
Capitalized character with period	H.
Ends in digit	A9
Contains hyphen	H-P

Training data

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

Today

- Overview: Information extraction
- (NP-)chunking
- Named entity recognition
- The Constraint Grammar approach
- Relation detection and classification

CG-approach to NER

- See OB-tagger
  - (text example: dn)
- Same approach as to tagging:
  - Start with all possible classes
  - Write rules which remove alternatives
  - May end with more than one answer

Today

- Overview: Information extraction
- (NP-)chunking
- Named entity recognition
- The Constraint Grammar approach
- Relation detection and classification

Relation detection and classification

- Two steps
  - Detection:
    Is there a relation between two entities?
  - Classification: What kind of 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 [*PERS* 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].

As logical relations

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 American is a unit of AMR	$PartOf = \{\langle a, b\rangle, \langle c, d\rangle\}$
Tim Wagner works for American Airlines	$OrgAff = \{\langle c, e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a, f\rangle, \langle a, g\rangle, \langle a, h\rangle, \langle a, i\rangle\}$

Supervised learning

- Corpus marked with NEs and relations

Features, examples

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	$Airlines \leftarrow_{subj} matched \leftarrow_{comp} said \rightarrow_{subj} Wagner$

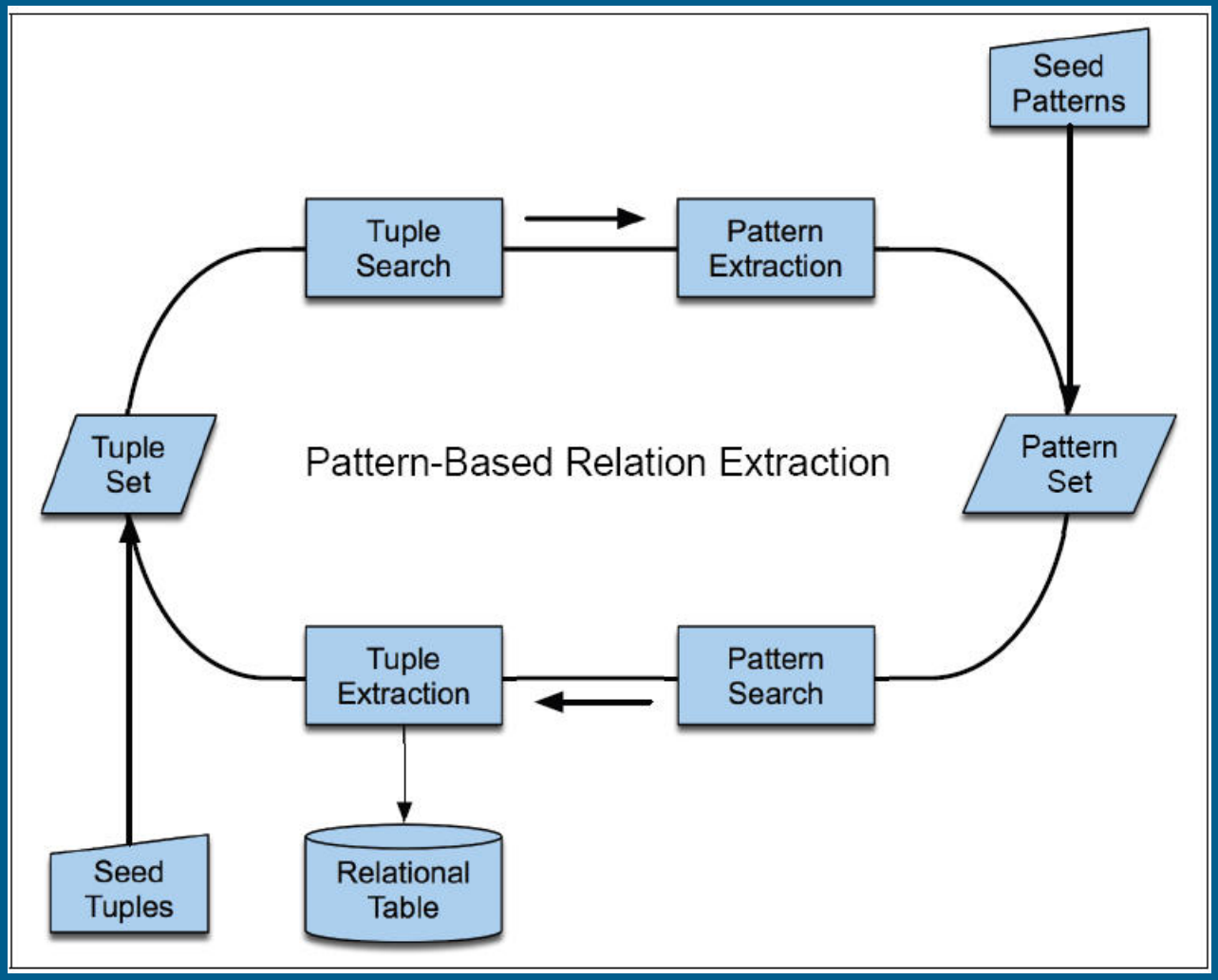
Pattern-matching

1. Choose a pattern, e.g.
   - * has a hub at +
2. Find pairs in the construction, e.g.
   - Milwaukee-based Midwest has a hub at KCI
   - Bulgaria Air has a hub at Sofia Airport
3. Extend/refine pattern, e.g.
   - [ORG] has a ADJ* hub at [LOC]

Bootstrapping

1. Choose a pattern
2. Find pairs in the construction
3. Find other occurrences of these pairs
4. Extract patterns from this, e.g.
   - [ORG] which uses [LOC] as hub
   - [ORG]’s hub at [LOC]
   - [LOC] a ADJ* hum for [ORG]
5. Repeat from (2)

Bootstrapping loop



Beware

- Check that bootstrapping does not drift away
  - Control against original tuples
  - Skip details
- Evaluate, either
  - Count occurrences of relations in sentences in corpus and evaluate against them
  - Count whether relaionship is entered into data base

Information extraction

- The information extraction approach shaped by
  - MUC: message understanding conferences
  - Competition: make the best system
  - 1987-97
  - DARPA
- See
  - Wikipedia
  - Bibliographical and historical notes J&M, 22
- Overlap with other tasks/approaches
  - But comparison not always immediate