Information Extraction

PCL II, CL, UZH

June 01, 2016

Outline

Universität Zürich[UZH]

- Information extraction
- Chunking/chinking
- Named entity recognition
- Temporal expression processing
- Relation extraction
- Event extraction

Information extraction

Universität Zürich[UZH]

- Acquisition of structured data from unstructured text
- not to confuse with information retrieval
  - IR finds documents within a collection, that are relevant for a specific information need
    - input: query
    - output: set of relevant documents
  - IE actually "looks into" the documents
    - input: text
    - output: structured info from it

Information extraction

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

Information extraction

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

Named entity recognition

In 1854 **Matthew Perry**, a **U.S. Navy** commodore performed a blockade of **Japan** in the **Yokohama bay** in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the **United States** and **the Japanese Empire**. In five years, similar treaties were signed with several **other Western countries**.

what: ?
who: ?
whom: ?
where: ?
when: ?

named entity:
anything with a proper name

Named entity recognition

In 1854 **Matthew Perry**, a **U.S. Navy** commodore performed a blockade of **Japan** in the **Yokohama bay** in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the **United States** and **the Japanese Empire**. In five years, similar treaties were signed with several **other Western countries**.

what: ?
who: ?
whom: ?
where: ?
when: ?

named entity:
person / location / geo-political / ...

Temporal expr. detection/analysis

In **1854** Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted **for more than 200 years**. As a result diplomatic relations were established between the United States and the Japanese Empire. **In five years**, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

named entity:
person / location / geo-political / ...

temporal expression:
phrase indicating a moment in time or time period

Temporal expr. detection/analysis

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

named entity:
person / location / geo-political / ...

temporal expression:
absolute / relative / duration

(Co-)reference resolution

Universität Zürich[UZH]

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted **for more than 200 years**. As a result diplomatic relations were established between the **United States** and **the Japanese Empire**. **In five years**, similar treaties were signed with several **other Western countries**.

what: ?
who: ?
whom: ?
where: ?
when: ?

reference resolution:
U.S. = **United States**
Japan = **the Japanese Empire**

(Co-)reference resolution

Universität Zürich[UZH]

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

reference resolution:
U.S. = United States
Japan = the Japanese Empire

Anaphora resolution

Universität Zürich[UZH]

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

reference resolution:
U.S. = United States
Japan = the Japanese Empire

Relation detection and classification

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted **for more than 200 years**. As a result diplomatic relations were established between the **United States** and **the Japanese Empire**. **In five years**, similar treaties were signed with several **other Western countries**.

what: ?
who: ?
whom: ?
where: ?
when: ?

relation detection:
who is what of/to whom/what?
family / employment / part-whole / membership / geo-spatial..

Event detection and classification

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: ?
who: ?
whom: ?
where: ?
when: ?

event detection and classification:
detect and classify events, binding the named entities and others

Template filling

In 1854 Matthew Perry, a U.S. Navy commodore performed a blockade of Japan in the Yokohama bay in order to force it to end its policy of isolation, which has lasted for more than 200 years. As a result diplomatic relations were established between the United States and the Japanese Empire. In five years, similar treaties were signed with several other Western countries.

what: policy of isolation **where:** Japan
when: (period) before 1654 - 1854

who: Matthiew Perry **organization:** U.S. Navy
relation: member **type:** commodore

what: blockade **who:** Matthew Perry / U.S.
whom: Japan **when:** 1854
where: Yokohama bay

what: establishment of diplomatic relations
who: Japan **when:** 1854
who: U.S.

what: establishment of diplomatic relations
who: Japan **when:** (by) 1859
who: other Western countries

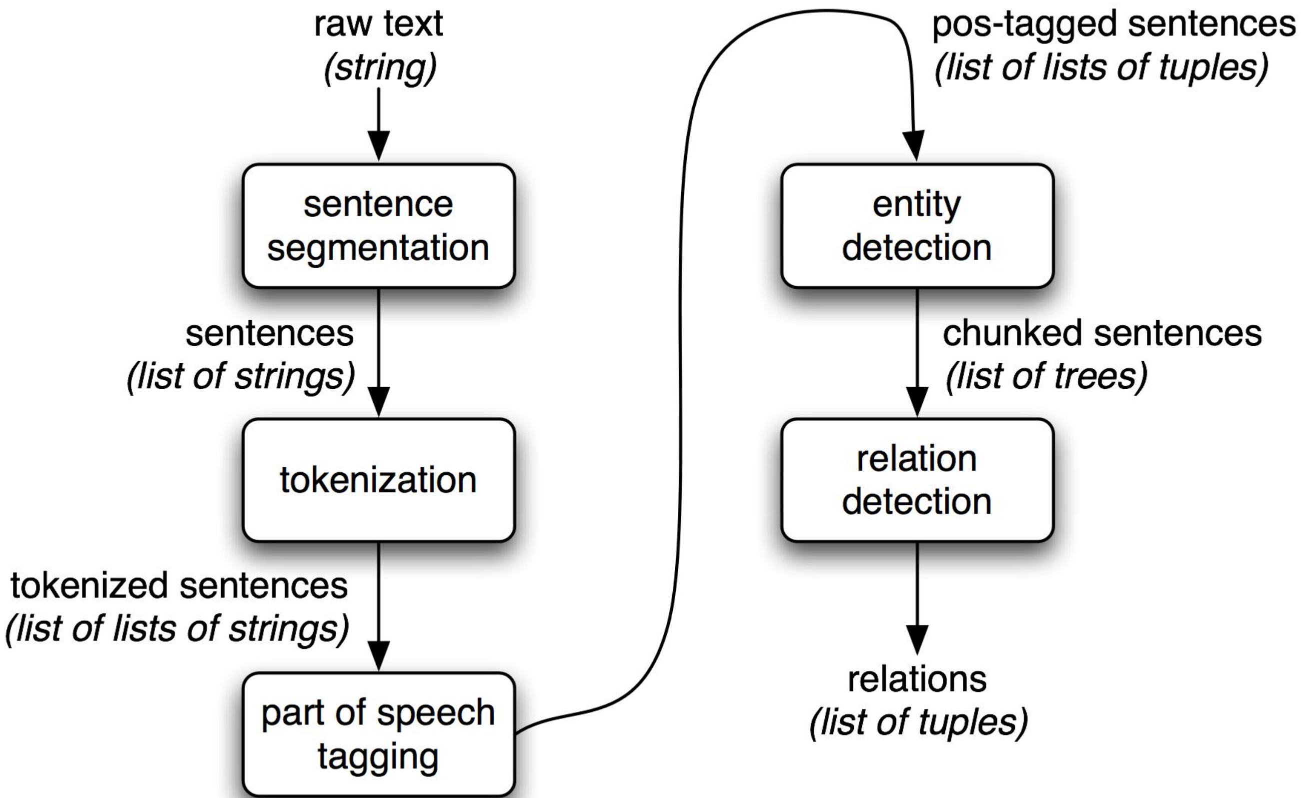
Information Extraction

- (Chunking)
- Named entity recognition
- Reference and anaphora resolution
- Relation detection/classification
- Temporal expression detection/analysis
- Event detection/classification
- Template filling

Information Extraction

- **(Chunking)**
- **Named entity recognition**
- Reference and anaphora resolution
- **Relation detection/classification**
- **Temporal expression detection/analysis**
- Event detection/classification
- Template filling

Information Extraction



Picture Source: http://www.nltk.org/book/ch07.html

Chunking

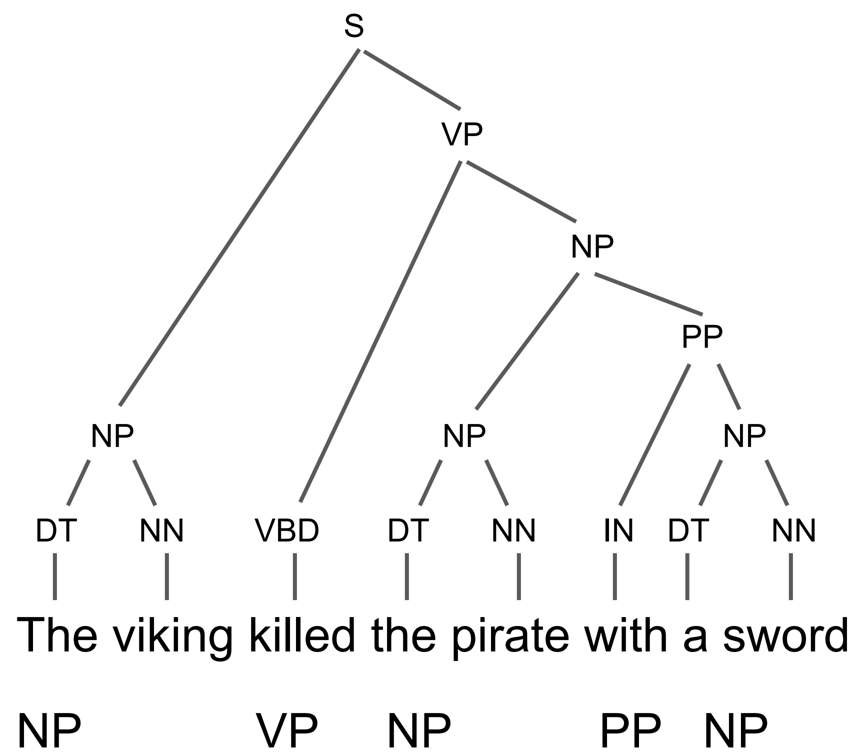
Universität Zürich[UZH]

- Detection of multi-token sequences, e.g.
    - New York
    - The big bad wolf
- Approaches:
    - based on manually defined regular expressions
    - automatically learned taggers/classifiers

Chunking vs consituency parsing

Parsing:
- Deeper Tree
- Nested Structures + Recursive Structures
- More information

Chunking:
- Shallow
- No nested Structures
- Less information (but often still enough)
- Lighter
- Often limited to one Chunk Type



RegExp Chunking

Universität Zürich[UZH]

- define regular expressions of PoS-tags
- matching sequences = chunks

```
>>> sentence = [("the", "DT"), ("little", "JJ"), ("yellow", "JJ"),
... ("dog", "NN"), ("barked", "VBD"), ("at", "IN"),  ("the", "DT"), ("cat", "NN")]

>>> grammar = "NP: {<DT>?<JJ>*<NN>}"

>>> cp = nltk.RegexpParser(grammar)
>>> result = cp.parse(sentence)
>>> print result
(S
  (NP the/DT little/JJ yellow/JJ dog/NN)
  barked/VBD
  at/IN
  (NP the/DT cat/NN))
```

RegExp Ch_i_nking

- define regular expressions of PoS-tags
- matching sequences **excluded** from chunks

```
grammar = r"""
  NP:
    {<.*>+}          # Chunk everything
    }<VBD|IN>+{      # Chink sequences of VBD and IN
  """
sentence = [("the", "DT"), ("little", "JJ"), ("yellow", "JJ"),
       ("dog", "NN"), ("barked", "VBD"), ("at", "IN"),  ("the", "DT"), ("cat", "NN")]
cp = nltk.RegexpParser(grammar)

>>> print cp.parse(sentence)
(S
  (NP the/DT little/JJ yellow/JJ dog/NN)
  barked/VBD
  at/IN
  (NP the/DT cat/NN))
```

Chunking as Tagging

Universität Zürich[UZH]

- The IOB format
  - Inside/outside/beginning
  - token-level annotation
- Example chunks:
  - Matthew Perry
  - a U.S. Navy commodore
  - a blockade
  - Japan
  - the Yokohama bay

In	O
1854	O
Matthew	B
Perry	I
,	O
a	B
U.S.	I
Navy	I
commodore	I
performed	O
a	B
blockade	I
of	O
Japan	B
in	O
the	B
Yokohama	I
bay	I
.	O

Chunking as Tagging

Universität Zürich[UZH]

- The IOB format
  - Inside/outside/beginning
  - token-level annotation
- Possible to sub-specify chunk types
  - NP / VP / …
  - tags turn into O / B-NP / I-NP / B-VP / …

In	O
1854	O
Matthew	B
Perry	I
,	O
a	B
U.S.	I
Navy	I
commodore	I
performed	O
a	B
blockade	I
of	O
Japan	B
in	O
the	B
Yokohama	I
bay	I
.	O

Chunking as Tagging

- Take a IOB-tagged corpus
- Train a tagger/classifier on it
  - IOB / IOB with chunk types

Chunking in NLTK

Universität Zürich[UZH]

- CoNLL-2000 corpus:
  - NP / VP / PP

```
>>> from nltk.corpus import conll2000
>>> print conll2000.chunked_sents('train.txt')[99]
(S
  (PP Over/IN)
  (NP a/DT cup/NN)
  (PP of/IN)
  (NP coffee/NN)
  ,/,
  (NP Mr./NNP Stone/NNP)
  (VP told/VBD)
  (NP his/PRP$ story/NN)
  ./.)
```

```
>>> print conll2000.chunked_sents('train.txt',
chunk_types=['NP'])[99]
(S
  Over/IN
  (NP a/DT cup/NN)
  of/IN
  (NP coffee/NN)
  ,/,
  (NP Mr./NNP Stone/NNP)
  told/VBD
  (NP his/PRP$ story/NN)
  ./.)
```

Chunking in NLTK, RegExp:

```
>>> from nltk.corpus import conll2000

>>> test_sents = conll2000.chunked_sents('test.txt', chunk_types=['NP'])

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

>>> print cp.evaluate(test_sents)
ChunkParse score:
    IOB Accuracy:  87.7%
    Precision:     70.6%
    Recall:        67.8%
    F-Measure:     69.2%
```

Chunking in NLTK, tagging:

```
>>> from nltk.corpus import conll2000
>>> from nltk.tag.hmm import HiddenMarkovModelTagger as HmmTagger

>>> train_sents = [ [(t, c if c[-2:] == 'NP' else 'O') for w, t, c in snt]
        for snt in conll2000.iob_sents('train.txt')]

>>> defTagger = nltk.DefaultTagger('O')
>>> uniTagger = nltk.UnigramTagger(train_sents, backoff=defTagger)
>>> biTagger = nltk.BigramTagger(train_sents, backoff=uniTagger)
>>> hmm_tagger = HmmTagger.train(train_sents)

>>> test_sents_iob = [ [(t, c if c[-2:] == 'NP' else 'O') for w, t, c in snt]
        for snt in conll2000.iob_sents('test.txt')]

>>> print defTagger.evaluate(test_sents_iob)
0.434366886886604175

>>> print uniTagger.evaluate(test_sents_iob)
0.832112628490617 8

>>> print biTagger.evaluate(test_sents_iob)
0.934166367646 7484

>>> print hmm_tagger.evaluate(test_sents_iob)
0.936044916309 6017
```

Information Extraction

- (Chunking)
- **Named entity recognition**
- Reference and anaphora resolution
- Relation detection/classification
- Temporal expression detection/analysis
- Event detection/classification
- Template filling

Named entity recognition

- task: detect mentions of things with proper names (mostly)
  - people
  - locations
  - organizations
  - geo-political (country, kanton,...)
  - facility (bridge, building, airport,...)
  - vehicle

Specific NEs

- BioMed: biological entities (genes/proteins/...)
- Text+Berg: mountains/glaciers/huts/people
- Special kinds of NEs
  - temporal expressions
  - numerical expressions
  - → typically handled **separately**

Ambiguity for NEs

1. different mentions of the same entity
   - Switzerland / the Swiss Confederation
   - Martin / M. Volk / prof. dr. Martin Volk

2. polysemy

[$_{PERS}$ 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...

Figure 22.4 Examples of type ambiguities in the use of the name *Washington*.

Automatic NE recognition

Essentially:

- find phrases that are named entities
- classify into people/locations/organizations/...

In 1854 [$_{PERS}$Matthew Perry], a [$_{GPE}$U.S.] [$_{ORG}$Navy] commodore performed a blockade of [$_{GPE}$Japan] in the [$_{LOC}$Yokohama bay].

IOB labelling

In	O
1854	O
Matthew	B-pers
Perry	I-pers
,	O
a	O
U.S.	B-geo-pol
Navy	B-org
commodore	O
performed	O
a	O
blockade	O
of	O
Japan	B-geo-pol
in	O
the	O
Yokohama	B-loc
bay	I-loc
.	O

NER as tagging

- Features:
  - surface form, lemma
  - shape: orthographic pattern (lower/upper/case, initial-upper, mixed, includes numbers/punct., ...)
  - affixes of this and surrounding words
  - syntactic chunk labels
  - presence of the word in a list of places/people/organizations/etc.
  - predictive tokens present
    - **X**, president of **Y**
    - **Z** M.D. / Dr. **Z**
    - in **W, Q**
  - bag-of-words/bag-of-N-grams
  - ...

NER as tagging

Universität Zürich[UZH]

- Methods:
  - Hidden Markov Models
  - Support Vector Machines
  - Transformation-based Learning
  - Decision trees
  - Conditional Random Fields
  - …
- Evaluation:
  - precision, recall, F-score, …
- `nltk.corpus.conll2002` (Dutch)

Information Extraction

Universität Zürich[UZH]

- (Chunking)
- Named entity recognition
- Reference and anaphora resolution
- **Relation extraction**
- Temporal expression detection/analysis
- Event detection/classification
- Template filling

Relation Extraction

Universität Zürich[UZH]

- extract relations between (typically) named entities
- specified as $NE_1\ x\ NE_2$, where
  - x is a connecting word, such as “in”:

```
>>> IN = re.compile(r'.*\bin\b(?!\b.+ing)')
>>> for doc in nltk.corpus.ieer.parsed_docs('NYT_19980315'):
...     for rel in nltk.sem.extract_rels('ORG', 'LOC', doc,
...                                      corpus='ieer', pattern = IN):
...         print nltk.sem.show_raw_rtuple(rel)
[ORG: 'WHYY'] 'in' [LOC: 'Philadelphia']
[ORG: 'McGlashan &AMP; Sarrail'] 'firm in' [LOC: 'San Mateo']
[ORG: 'Freedom Forum'] 'in' [LOC: 'Arlington']
[ORG: 'Brookings Institution'] ', the research group in' [LOC: 'Washington']
...
```

Relation Extraction

- additional info: PoS-tags for fillers

```
>>> from nltk.corpus import conll2002
>>> vnv = """
... (
... is/V|     # 3rd sing present and
... was/V|    # past forms of the verb zijn ('be')
... werd/V|   # and also present
... wordt/V   # past of worden ('become')
... )
... .*        # followed by anything
... van/Prep # followed by van ('of')
... """
>>> VAN = re.compile(vnv, re.VERBOSE)
>>> for doc in conll2002.chunked_sents('ned.train'):
...     for r in nltk.sem.extract_rels('PER', 'ORG', doc,
...                                    corpus='conll2002', pattern=VAN):
...         print  nltk.sem.show_clause(r, relsym="VAN")
VAN("cornet_d'elzius", 'buitenlandse_handel')
VAN('johan_rottiers', 'kardinaal_van_roey_instituut')
VAN('annie_lennox', 'eurythmics')
```

Information Extraction

- (Chunking)
- Named entity recognition
- Reference and anaphora resolution
- Relation detection/classification
- **Temporal expression detection/analysis**
- Event detection/classification
- Template filling

Temporal expressions

- a phrase indicating a point of time
  - absolute
    - May 13, 2009 / summer of '69 / 10am
  - relative
    - yesterday / next week / 2 years before
- or a duration: time span
  - absolute anchor/reference point
    - 1900-1936 / from 2pm to 4pm
  - relative anchor/reference point
    - for the next two weeks / for the past millennium
  - no anchor/reference point
    - 3 days / 4 hours / 2 years

Temporal expressions

- grammatical constructions that have temporal lexical triggers as their heads
- triggers: (proper)nouns, adjectives, adverbs

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*

Figure 20.18 Examples of temporal lexical triggers.

Picture taken from: Jurafsky, Daniel, and James H. Martin

Temporal Expressions

- Rule-Based Approaches
- Sequence Labelling Approaches
- Hybrid (constituent-based) approaches

Temporal expression recognition

- Rule-based approaches
  - use a set of rules or patterns to detect temporal expressions:

```
# yesterday/today/tomorrow
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?(($OT+the$CT+\s+)?$OT+day$CT+\s+
$OT+(before|after)$CT+\s+)?$OT+$TERelDayExpr$CT+(\s+$OT+(morning|afternoon|evening|night)
$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gio;

$string =~ s/($OT+\w+$CT+\s+)
<TIMEX2 TYPE=\"DATE\"[^>]*>($OT+(Today|Tonight)$CT+)<\/TIMEX2>/$1$2/gso;

# this/that (morning/afternoon/evening/night)
$string =~ s/(($OT+(early|earlier|later?)$CT+\s+)?$OT+(this|that|every|the$CT+\s+
$OT+(next|previous|following))$CT+\s*$OT+(morning|afternoon|evening|night)
$CT+(\s+$OT+thereafter$CT+)?)/<TIMEX2 TYPE=\"DATE\">$1<\/TIMEX2>/gosi;
```

Figure 22.19 Fragment of Perl code from MITRE's TempEx temporal tagging system.

Temporal expression recognition

- Sequence labelling approaches
  - IOB labelling
  - statistical
  - features:
    - surface form, lemma
    - bag of words in the window around the target
    - shape
    - PoS tags of the word and context
    - chunking/syntactic tags
    - word and context presence in a list of lexical triggers

Temporal expression recognition

- Hybrid (constituent-based) approaches
  - do a full syntactic parse
  - apply a binary classifier to every constituent
  - use very similar features as with sequence labelling

Temporal normalization

- turn the ambiguous expression into a date and time
  - relative -- against an anchor/reference point

TimeML

- Tagging by the TIDES standard / TimeML:

  <TIMEX3>Last week</TIMEX3> it was <TIMEX3>10 years</TIMEX3> since we got our cat

- Format supports normalization for points in time and durations:

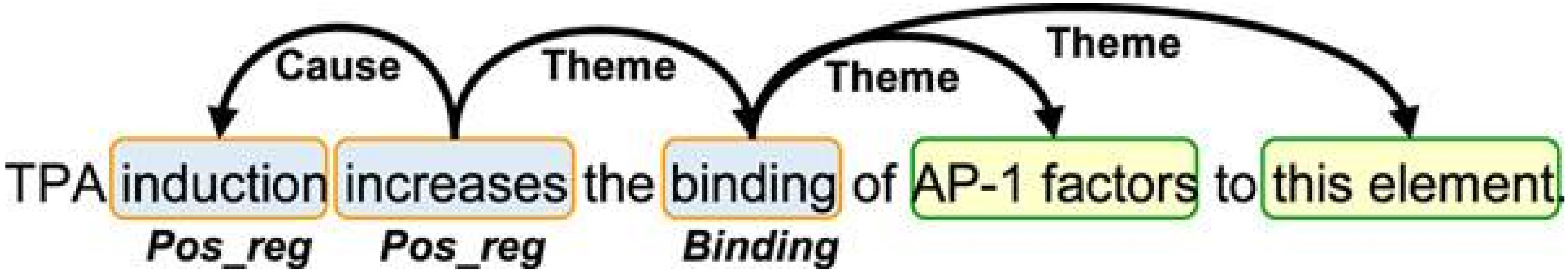
```
<TIMEX3 id=t1 type="DATE" value="2007-07-02" functionInDocument="CREATION_TIME">
July 2, 2007 </TIMEX3>  A fare increase initiated <TIMEX3 id="t2" type="DATE"
value="2007-W26"  anchorTimeID="t1">last week</TIMEX3> by UAL Corp's United  Airlines was
matched by competitors over <TIMEX3 id="t3" type="DURATION" value="P1WE" anchorTimeID="t1">
the weekend </TIMEX3>, marking the second successful fare increase in
<TIMEX3 id="t4" type="DURATION"  value="P2W" anchorTimeID="t1"> two weeks </TIMEX3>.
```

Figure 22.21 TimeML markup include normalized values for temporal expressions.

Temporal normalization

- predominantly rule-based approaches
    - manually encode interpretations for *last/one-before-last/later/before,* weekdays, months, years, all possible time and date formats and basically everything else one can think of

Event Extraction



To conclude

Information extraction:

Raw text → structured data (events, relations, …)

Exam

- 24h
- Practical tasks, similar to labs
  - most likely one simple task in the beginning
  - most likely one on machine learning
  - most likely one on creating a pipeline using NLTK's linguistic pre-processing (tokenization, sentence splitting, pos-tagging, …)
- Analyze output and/or report results

To conclude PCL2

Universität Zürich[UZH]

- some programming concepts
  - classes, packages
  - file I/O, XML
  - external processes
- some organizational things
  - code style and comments
  - code sharing
  - testing, debugging
- algorithms and approaches
  - probabilities
  - machine learning, classification
  - dynamic programming

To conclude PCL2

Universität Zürich UZH

- NLP tasks
  - handling files and corpora
  - n-grams, language modelling
  - classification (documents, words)
  - tagging
  - longest common subsequence, edit/Levenshtein distance
  - parallel sentence alignment
  - syntax and parsing, CFG, PCFG
  - semantics: logic, WSD
  - information extraction

To conclude PCL2

In other words:

- what you can do
  - NLP
  - programming

but most importantly:

- how stuff works
  - to know what you’re dealing with
  - more importantly:

    to understand algorithms and learn to design and create your own


That's all Folks!

Except...

… course evaluation