sentence annotation:

named entity annotation

Hasso Plattner Institute, Potsdam
question answering seminar

stefan klauck
agenda

• definition

• applications

• challenges

• approaches
definition

• named entity
  – word or sequence of words
  – used to refer to something of interest in a particular application
definition

• named entity
  – *word* or sequence of *words*
  – used to *refer to something* of interest in a particular *application*
definition

• named entity
  – *(word or sequence of words)*
  – used to *refer to something* of interest in a particular *application*
definition

• named entity *annotation*

• prerequisite:
  – recognition
  – classification
definition

• named entity *annotation*

• prerequisite:
  – recognition
  – classification

example:
Steven Paul Jobs, co-founder of Apple, was born in 1955.
definition

• named entity *annotation*

• prerequisite:
  – recognition
  – classification

example:
*Steven Paul Jobs*, co-founder of *Apple*, was born in *1955*.  

<table>
<thead>
<tr>
<th>person</th>
<th>organization</th>
<th>year</th>
</tr>
</thead>
</table>
definition

• named entity *annotation*

• prerequisite:
  – recognition
  – classification

example:

*<person>Steven Paul Jobs</person>**, co-founder of *<organization>Apple</organization>**, was born in *<year>1955</year>*.
applications

• named entity recognition and classification:
  – part of information extraction
  – unstructured ➔ structured information
  – semantic of word/s
applications

• named entity recognition and classification:
  – part of information extraction
  – unstructured ➔ structured information
  – semantic of word/s

• usage is application dependent
  – find out the semantic
  – storing of entities and relations in databases
application – question answering

“Who is Warren Moon’s Agent?”

Corpus

Web

Question Analysis

Query Construction

Document Retrieval

Sentence Retrieval

Sentence Annotation

Answer Extraction

Answer Validation

“Leigh Steinberg”
application – question answering

“Who is Warren Moon’s Agent?”

annotate question

annotate sentence

(annotate corpus)
challenges

• kind of data to annotate
  – here: (primary) unstructured text
  – language

• kind of application
  – types of entities
  – maximize precision, recall or both
challenge – entity type

• “something of interest”

• based on “rigid designator” defined by S. Kripke
  – philosophical term
  – denote unambiguous things
challenge – entity type

• “enamex” (MUC-6)
  – persons, locations and organizations

• date and time
• other numeral types (percentages, quantities)
• ...

challenge – entity type

• “enamex” (MUC-6)
  – persons, locations and organizations

• date and time

• other numeral types (percentages, quantities)
  • ...

• domain dependent
## approaches

<table>
<thead>
<tr>
<th>rule-based</th>
<th>statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>patterns &amp; lexicons</td>
<td>probabilities</td>
</tr>
<tr>
<td>linguistic analyses</td>
<td>language model</td>
</tr>
<tr>
<td>trial and error</td>
<td>annotate data</td>
</tr>
</tbody>
</table>
approaches

rule-based
- patterns & lexicons
- linguistic analyses
- trial and error

statistical
- probabilities
- language model
- annotate data

machine learning
rule-based approach

• main work: linguistic analysis
  ➔ lexicons & patterns/rules

building blocks of rules

• entity types
• regular expressions
• features
statistical approach

• main work: annotate training data
  ➔ large annotated corpus, statistics

use of features
statistical approach - example

<table>
<thead>
<tr>
<th>word class</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>oneDigitNum</td>
<td>1</td>
</tr>
<tr>
<td>containsDigitAndColon</td>
<td>2:34</td>
</tr>
<tr>
<td>containsAlphaDigit</td>
<td>A4</td>
</tr>
<tr>
<td>allCaps</td>
<td>KRDL</td>
</tr>
<tr>
<td>capPeriod</td>
<td>M.</td>
</tr>
<tr>
<td>firstCommonWordInitCap</td>
<td></td>
</tr>
<tr>
<td>firstNonCommonWordIC</td>
<td></td>
</tr>
<tr>
<td>CommonWordInitCap</td>
<td>Department</td>
</tr>
<tr>
<td>initCapNotCommonWord</td>
<td>David</td>
</tr>
<tr>
<td>mixedCasesWord</td>
<td>ValueJet</td>
</tr>
<tr>
<td>charApos</td>
<td>O’clock</td>
</tr>
<tr>
<td>allLowerCase</td>
<td>can</td>
</tr>
<tr>
<td>compoundWord</td>
<td>ad-hoc</td>
</tr>
</tbody>
</table>
### statistical approach - example

<table>
<thead>
<tr>
<th>word class</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>oneDigitNum</td>
<td>1</td>
</tr>
<tr>
<td>containsDigitAndColon</td>
<td>2:34</td>
</tr>
<tr>
<td>containsAlphaDigit</td>
<td>A4</td>
</tr>
<tr>
<td>allCaps</td>
<td>KRDL</td>
</tr>
<tr>
<td>capPeriod</td>
<td>M.</td>
</tr>
<tr>
<td>firstCommonWordInitCap</td>
<td>Department</td>
</tr>
<tr>
<td>firstNonCommonWordIC</td>
<td>David</td>
</tr>
<tr>
<td>CommonWordInitCap</td>
<td>ValueJet</td>
</tr>
<tr>
<td>initCapNotCommonWord</td>
<td>O’clock</td>
</tr>
<tr>
<td>mixedCasesWord</td>
<td>can</td>
</tr>
<tr>
<td>charApos</td>
<td>ad-hoc</td>
</tr>
</tbody>
</table>

- **CommonWordInitCap**
  - capitalized words
- **no** first words of sentence
# statistical approach - example

<table>
<thead>
<tr>
<th>word class</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>oneDigitNum</td>
<td>1</td>
</tr>
<tr>
<td>containsDigitAndColon</td>
<td>2:34</td>
</tr>
<tr>
<td>containsAlphaDigit</td>
<td>A4</td>
</tr>
<tr>
<td>allCaps</td>
<td>KRDL</td>
</tr>
<tr>
<td>capPeriod</td>
<td>M.</td>
</tr>
<tr>
<td>firstCommonWordInitCap</td>
<td></td>
</tr>
<tr>
<td>firstNonCommonWordIC</td>
<td></td>
</tr>
<tr>
<td>CommonWordInitCap</td>
<td>Department</td>
</tr>
<tr>
<td>initCapNotCommonWord</td>
<td>David</td>
</tr>
<tr>
<td>mixedCasesWord</td>
<td>ValueJet</td>
</tr>
<tr>
<td>charApos</td>
<td>O’clock</td>
</tr>
<tr>
<td>allLowerCase</td>
<td>can</td>
</tr>
<tr>
<td>compoundWord</td>
<td>ad-hoc</td>
</tr>
</tbody>
</table>

- **CommonWordInitCap**
- **capitalized words**
- **no first words of sentence**

| None of the named entities        | 8493                      |
| Location                          | 896                       |
| Person                            | 195                       |
| Date                              | 8                         |
| Money                             | 2                         |
machine learning

• possible using both approaches
• iterative process
  • 1. start with set of seeds
     – named entities (examples) and/or rules (start rules)
  • 2. find new named entities
  • 3. generate rules based on new entity set
features

• descriptors or characteristic attributes of words

eexample:
• boolean variable denoting whether a word is capitalized or not

• selection of features forms vector
features classification

- word-level feature
- list lookup feature
- document and corpus feature
questions?
references


