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.

person organization year

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

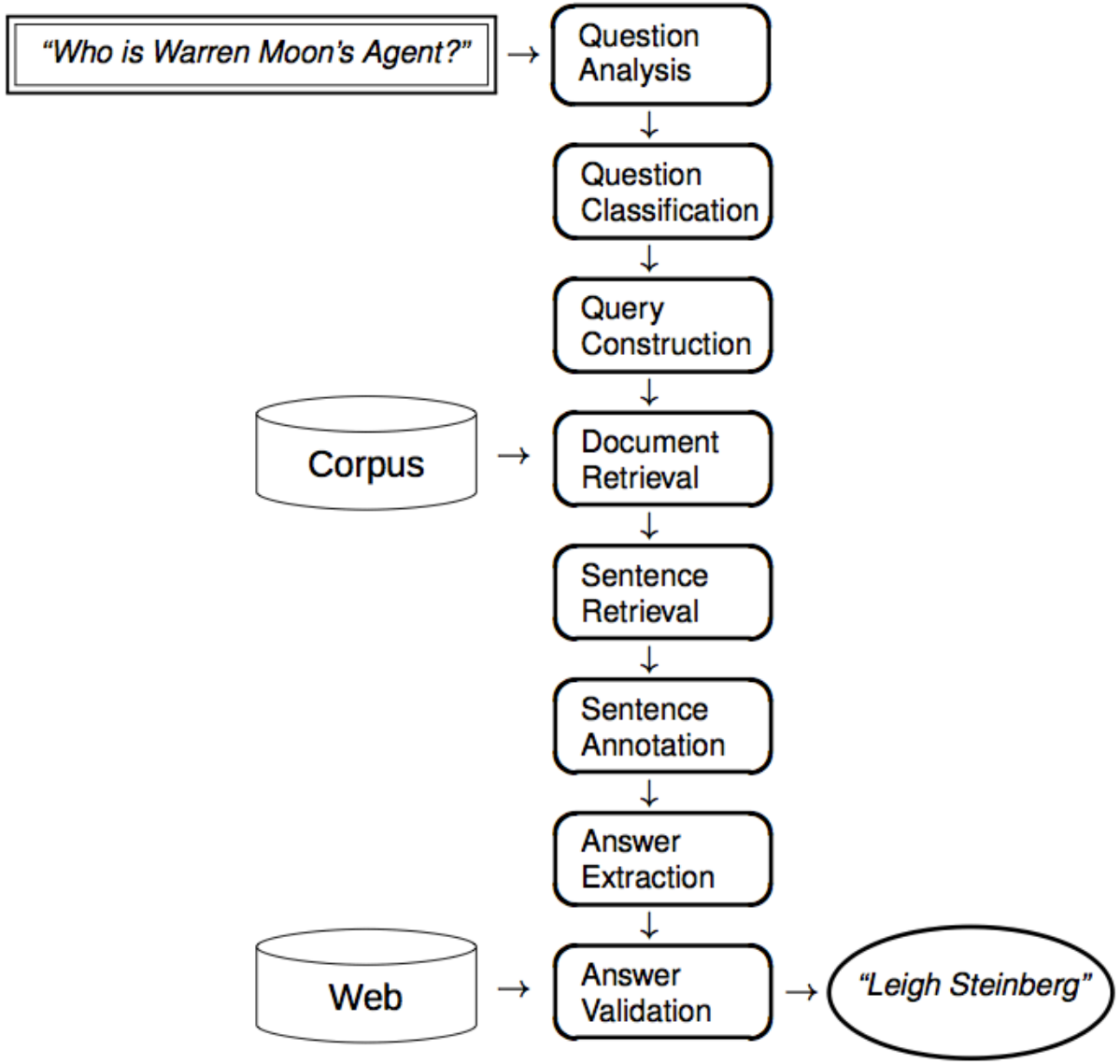
- named entity recognition and classification:
  - part of information extraction
  - unstructured → structured information
  - semantic of word/s

applications

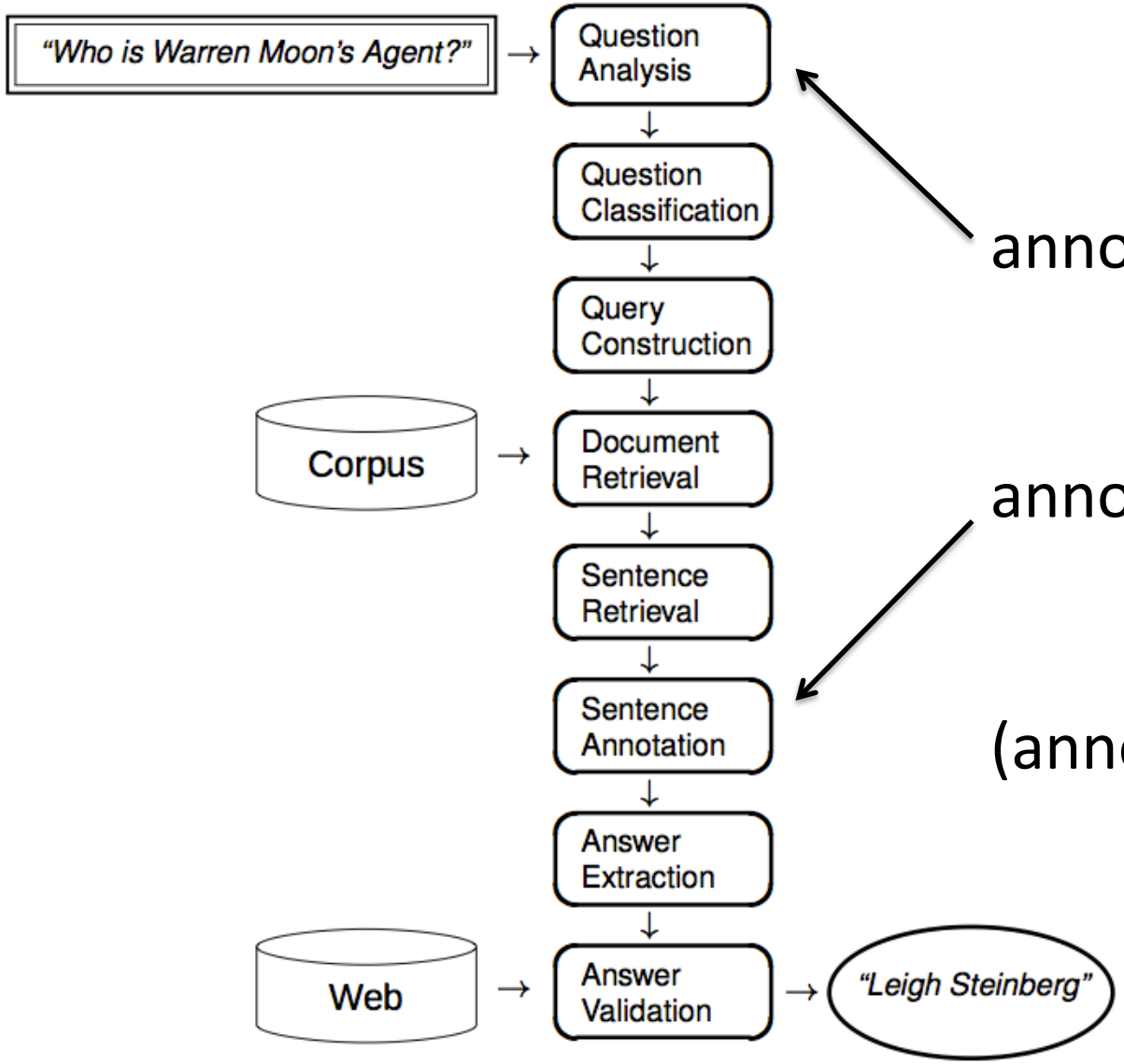
- named entity recognition and classification:
  - part of information extraction
  - unstructured → structured information
  - semantic of word/s

- usage is application dependent
  - find out the semantic
  - storing of entities and relations in databases

application – question answering



application – question answering



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

rule-based

- patterns & lexicons
- linguistic analyses
- trial and error

statistical

- probabilities
- language model
- annotate data

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

word class	example
oneDigitNum	1
containsDigitAndColon	2:34
containsAlphaDigit	A4
allCaps	KRDL
capPeriod	M.
firstCommonWordInitCap	
firstNonCommonWordIC	
CommonWordInitCap	Department
initCapNotCommonWord	David
mixedCasesWord	ValueJet
charApos	O'clock
allLowerCase	can
compoundWord	ad-hoc

statistical approach - example

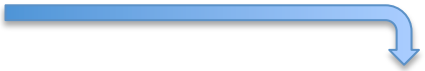
word class	example
oneDigitNum	1
containsDigitAndColon	2:34
containsAlphaDigit	A4
allCaps	KRDL
capPeriod	M.
firstCommonWordInitCap	
firstNonCommonWordIC	
CommonWordInitCap	Department
initCapNotCommonWord	David
mixedCasesWord	ValueJet
charApos	O'clock
allLowerCase	can
compoundWord	ad-hoc

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

statistical approach - example

word class	example
oneDigitNum	1
containsDigitAndColon	2:34
containsAlphaDigit	A4
allCaps	KRDL
capPeriod	M.
firstCommonWordInitCap	
firstNonCommonWordIC	
CommonWordInitCap	Department
initCapNotCommonWord	David
mixedCasesWord	ValueJet
charApos	O'clock
allLowerCase	can
compoundWord	ad-hoc

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



Organization	7525
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

example:

- 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

- W. Bruce Croft, Donald Metzler, Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison Wesley, 2010.

- David Nadeau, Satoshi Sekine. A survey of named entity recognition and classification. Linguisticae Investigaciones, 2007.

- Karën Fort, Maud Ehrmann, Adeline Nazarenko. Towards a Methodology for Named Entities Annotation. ACL workshop on Linguistic Annotation, 2009.

- William J. Black, Fabio Rinaldi, David Mowatt. FACILE: Description of the NE System used for MUC-7. Proceedings of the 7th Message Understanding Conference, 1998.

- Shihong Yu, Shuanhu Bai, Paul Wu. Description of the Kent Ridge Digital Labs System used for MUC-7. Proceedings of the 7th Message Understanding Conference, 1998.