APAN PS5430
Applied Text & Natural Language Analytics
Week 4: Information Extraction I

Javid Huseynov, Ph.D.
Thursday, February 13, 2019

COLUMBIA UNIVERSITY

Week 4 Agenda

- Overview of NLP Pipeline Tasks
- Information Extraction (IE)
- Named Entity Recognition & Linking (NER and NEL)
  - Approaches
  - Rule-based NER
  - Machine Learning-based NER
  - NER Evaluation Metrics
- Coreference Resolution
- IE Tool Demos
- Class Exercise: SpaCy NER Training & Entity Linking using Spark

NLP Pipeline Tasks

TEXT

Basic Text Processing

- Regular Expressions
- Tokenization Segmentation
- Stemming Lemmatization
- Part-of-Speech Tagging

Information Extraction

- Named Entity Recognition
- Named Entity Disambiguation
- Coreference Resolution
- Relationship Extraction

Natural Language Understanding

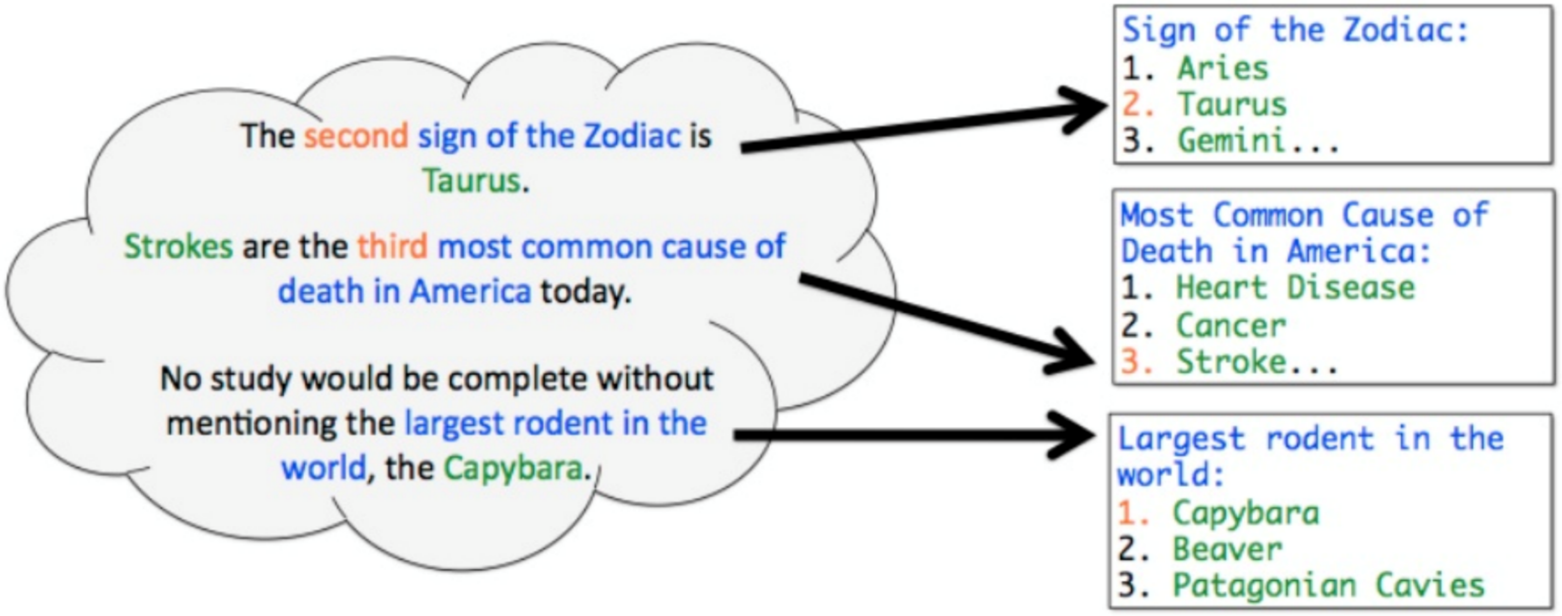
- Sentiment Analysis
- Semantic Analysis
- Question Answering
- Machine Translation

KNOWLEDGE

Information Extraction (IE)

The NLP task of **extracting structured** (semantic) **information** from unstructured text, to enable:

- further modeling by computer algorithms
- meaning and knowledge extraction



Subtasks

- Named Entity Recognition
- Named Entity Linking
- Coreference Resolution
- Relationship Extraction

Tools

- IBM Watson NLU
- Google Cloud NL
- Amazon Comprehend
- Thomson Reuters Open Calais
- Microsoft Text Analytics
- Stanford CoreNLP
- spaCy
- Natural Language Toolkit (NLTK)

*** open-source**

Named Entity Recognition (NER)

IE subtask of *finding* and *classifying* **named entities**, e.g. person, company, organization, geolocation, etc.

IBM announced that Technology Strategist, IBM Watson Customer Engagement Lisa Seacat DeLuca will be speaking at the NAI 2017 Annual Conference on Friday, April 7 at the Marriott Longwarf Hotel in Boston.

- **COMPANY:** IBM
- **PERSON:** Lisa Seacat DeLuca
- **POSITION:** Technology Strategist
- **ORGANIZATION:** National Academy of Inventors, NAI
- **DATE:** April 7
- **FACILITY:** Marriott Longwarf Hotel
- **CITY:** Boston

Methods

- Rule-based
  - Gazetteer Lookup
- Pattern-based
  - Regular Expressions
- ML Sequence-based
  - Supervised Classifier

Uses

- Document classification
- Information retrieval
- Question answering

NER: Uses & Tools

- The uses:
  - Named entities can be indexed, linked off, etc.
  - Sentiment can be attributed to companies or products
  - Many IE relations are associations between named entities
  - For question answering, answers are often named entities.

- Concretely:
  - Many web pages tag entities with links to bio or topic pages, etc.
  - Apple/Google/Microsoft/… smart recognizers for document content

- **NER tools**
  - Stanford NER
  - IBM Watson NLU
  - Thomson Reuters OpenCalais
  - Google Cloud NL
  - Amazon Comprehend
  - Azure Text Analytics
  - SpaCy
  - NLTK
  - Evri
  - Yahoo Term Extraction

Named Entity Linking (NEL), a.k.a. Named Entity Disambiguation

Task of *identifying* and *linking off* **named entities** to a knowledge base, such as DBpedia, Dun & Bradstreet, Yago, Babel, etc.

IBM announced that Technology Strategist, IBM Watson Customer Engagement Lisa Seacat DeLuca will be speaking at the NAI 2017 Annual Conference on Friday, April 7 at the Marriott Longwarf Hotel in Boston.

- IBM - Corporation headquartered in Armonk, New York
- Lisa Seacat DeLuca - IBM Technology Strategist
- NAI – National Academy of Inventors, **not** Network Advertising Initiative or National Association for Interpretation

Methods

- Rule-based
- Machine Learning
- Knowledge Graphs

Applications

- Hotlinking / Wikifying
- Enriching knowledge base
- Linking to enterprise data

Tools

- IBM Watson NLU
- TR Open Calais
- Google Cloud NL

NER Approaches

Knowledge-driven

- Advantages
  - Higher precision
  - Simple lookup methods
  - Small amount of training data
- Disadvantages
  - Expensive development
  - Domain dependence
  - Weak scalability

Data-driven

- Advantages
  - Higher recall
  - No need for grammars
  - No need for linguistic experts
  - Availability of tagged data
- Disadvantages
  - Lower precision
  - Require a lot of training data

Rule-based NER or NEL

- Regular Expressions
  - Phone number (###-###-####)
  - Email (contains @ and .com/org/net)
  - Capitalized names
- Context patterns
  - `[PERSON] earned [MONEY]`
    `Ex. David earned $10`
  - `[PERSON] joined [ORGANIZATION]`
    `Ex. Sam joined IBM`
  - `[PERSON],[JOBTITLE]`
    `Ex. Mary, the teacher`
- Challenges
  - First word in sentence is capitalized
  - Titles in articles can be all caps
  - Nested named entities can contain non-cap words
  - All nouns in German are capitalized
  - New proper names emerge daily, i.e. movies, books, celebrities, etc.
  - Proper names can be ambiguous, i.e.
    - Jordan (river, country or person)
    - Columbia University (mixed geo and organization)

Machine Learning-based NER or NEL

- Supervised Learning for NER
  - Label training data (POS and IOB tags)
  - methods: Hidden Markov Models, k-Nearest Neighbors, Decision Trees, AdaBoost, SVM, …
  - steps: NE recognition, POS tagging, Parsing

- Unsupervised Learning
  - labels must be automatically discovered
  - method: clustering
  - example: NE disambiguation, text classification

- **IOB2 Tagging Format**

```
Alex     B-PER
is       O
going    O
to       O
Los      B-LOC
Angeles  I-LOC
```

ML Approaches: *k*-Nearest Neighbor or Distance-based

- Given two objects X and Y:
  - $X = (x_1, x_2, …., x_n)$
  - $Y = (y_1, y_2, …, y_n)$

Calculate Euclidean distances

- $d(X, Y) = \sqrt{\sum_{i=0}^{n} |xi - yi|2}$

Higher Similarity ~ Lower Distance

- Pros:
  - Robust, simple, fast training
- Cons:
  - Depends on *distance* and *k*
  - Susceptible to noise

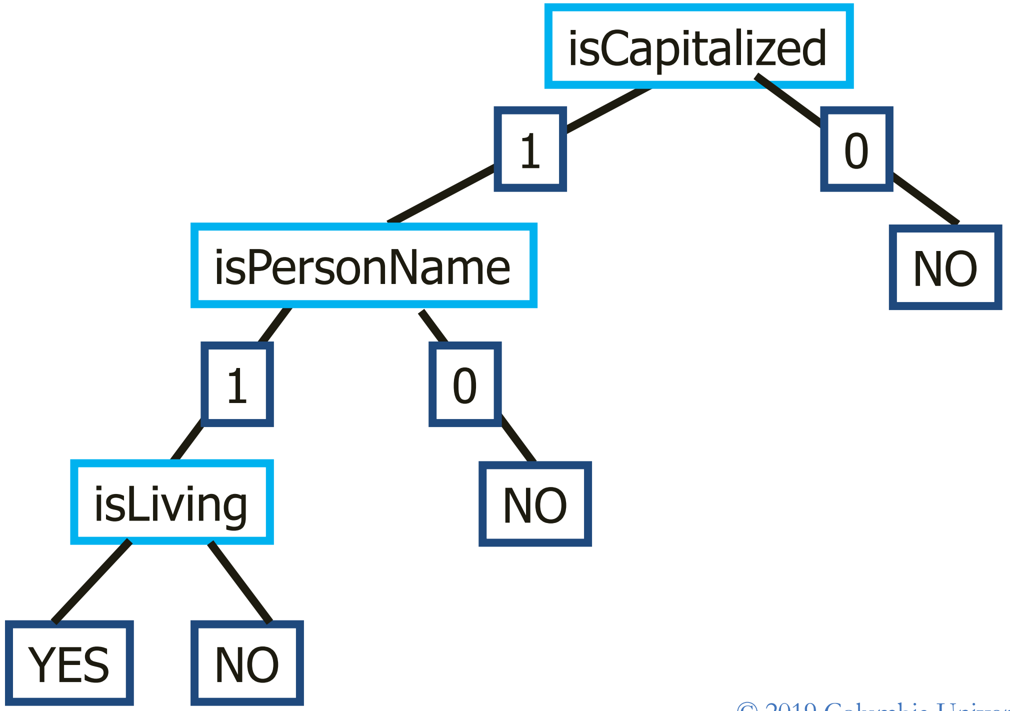
	Person	Capitalized	Living	NBA
Michael Jordan	1	1	1	1
Jordan	0	1	0	0
Kobe Bryant	1	1	1	1
Chicago Bulls	0	1	0	1
Los Angeles Lakers	0	1	0	1

- d (“Michael Jordan”, “Jordan”) = $\sqrt{1^2 + 0^2 + 1^2 + 1^2} = \mathbf{1.73}$
- d (“Michael Jordan”, “Kobe Bryant”) = $\sqrt{0^2 + 0^2 + 0^2 + 0^2} = \mathbf{0}$
- d (“Michael Jordan”, “Chicago Bulls”) = $\sqrt{1^2 + 0^2 + 1^2 + 0^2} = \mathbf{1.41}$

ML Approaches: Decision tree-based

- The classifier has a tree structure, where each node is either:
  - a leaf node which indicates the value of the target attribute (class) of examples
  - a decision node which specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test
- An instance x_p is classified by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance
- Pros:
  - Understandable Rules, Feature Extraction
- Cons:
  - Error-prone for multi-class labeling, requires a lot of training data

	Person	Capitalized	Living	isPerson?
Michael Jordan	1	1	1	YES
Jordan	0	1	0	NO
Chicago Bulls	0	1	0	NO



NER Evaluation Metrics

	Actual Entity	Actual Not Entity
Predicted Entity	True Positive	False Positive
Predicted Non Entity	False Negative	True Negative

```
Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

Accuracy = (TP + TN) / (TP + TN + FP + FN)
```

Coreference Resolution

Task of *finding* **expressions** that refer to the same entity in text:

"If I have learned nothing else in all my years here, my biggest lesson is you have to constantly reinvent this company", IBM CEO Ginni Rometty said.

- Ginni Rometty, IBM – *antecedents*
- I, my, this company – *anaphors*
- Antecedents and anaphors – *markables*

Methods

- Heuristics
  - Syntactic, Semantic, or Pragmatic (topic) rules
- Supervised Learning
  - Binary Classification (SVM)
  - Ranking
  - Anaphoricity
- Unsupervised Learning
  - Bayesian w/ Dirichlet distrib.
  - Expectation Maximization

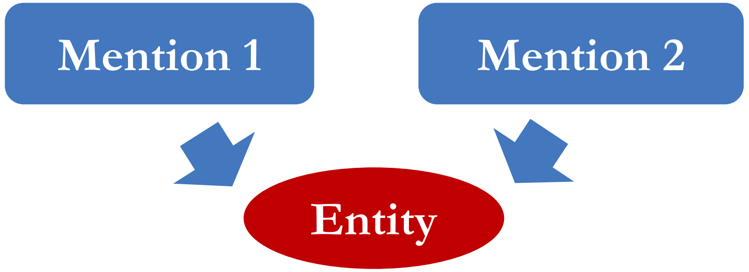
Applications

- Document Summarization
- Question Answering
- Relevance & Sentiment

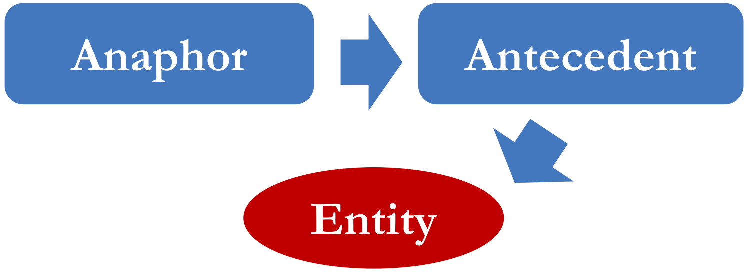
Coreference vs Anaphoricity

- **Coreference** is when two mentions refer to the same entity in the world
- **Anaphoricity** is when a term (anaphor) refers to another term (antecedent) and the interpretation of the anaphor is in some way determined by the interpretation of the antecedent
- Not all anaphoric relations are coreferential, e.g.
  - “We went to see **a concert** last night. **The tickets** were really expensive.”
- Conversely, multiple identical full noun-phrase (NP) references are typically coreferential but not anaphoric.

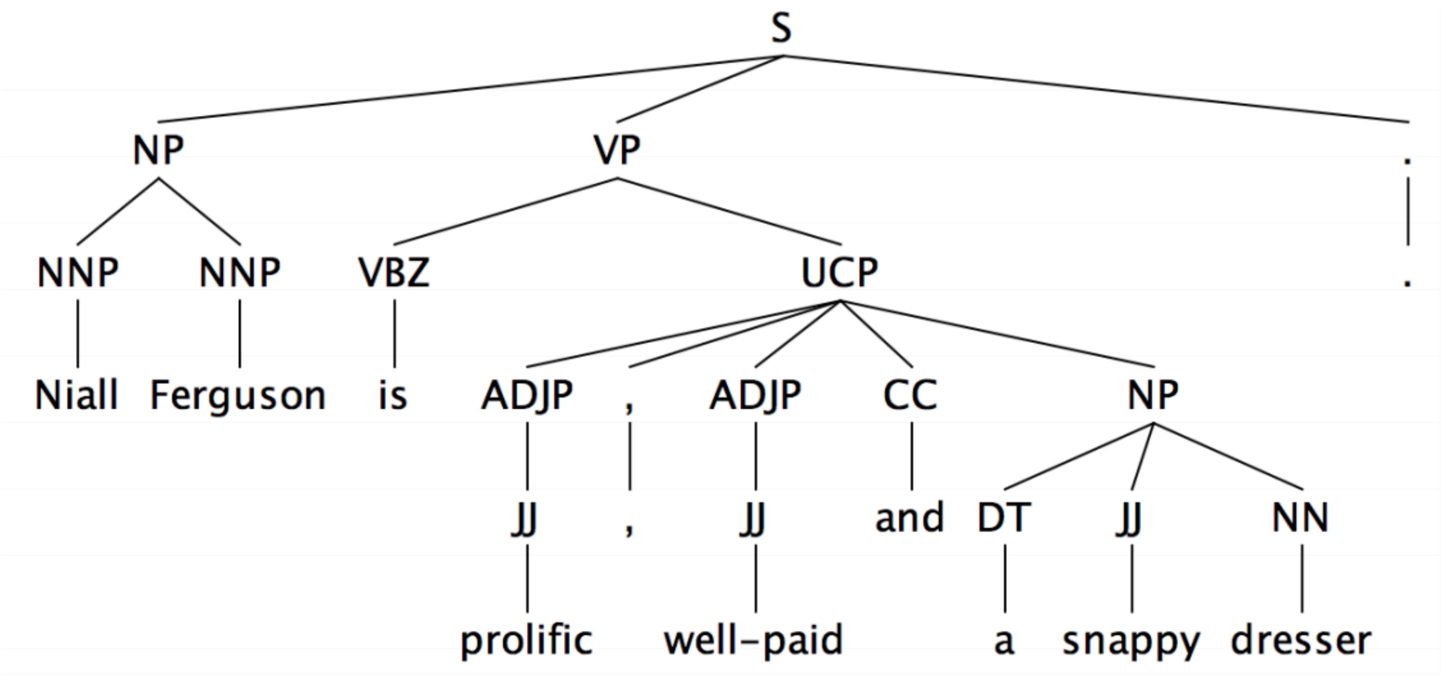
- Coreference

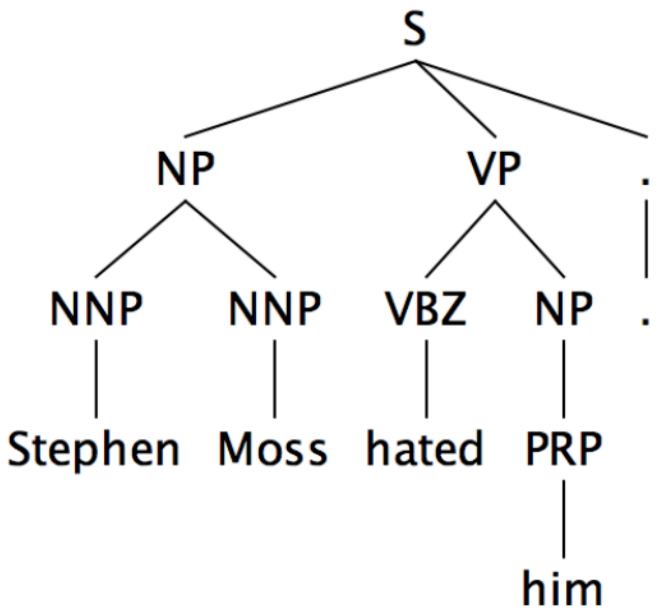


- Anaphoricity



Pronomial Anaphora Resolution: Hobbs' Naïve Algorithm





Information Extraction Tool Demos

IBM Watson NLU:

- https://natural-language-understanding-demo.ng.bluemix.net/

Thomson Reuters Open Calais:

- https://permid.org/onecalaisViewer

Google Natural Language Processing API:

- https://cloud.google.com/natural-language/

Amazon Comprehend:

- https://aws.amazon.com/comprehend/