Machine Learning in Natural Language Processing

Lecture 26: COMS 4771 Machine Learning

Vinodkumar Prabhakaran
vinod@cs.columbia.edu

Outline

- Motivation
- NLP Research Areas using ML
  - NLP Applications
  - Fundamental NLP steps
- NLP at Columbia
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- Conclusion

Outline

- **Motivation**
- NLP Research Areas using ML
  - NLP Applications
  - Fundamental NLP steps
- NLP at Columbia
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- Conclusion

Motivation: NLP in action

IBM Watson beating human champions in the Jeopardy! game

http://www.youtube.com/watch?v=BflW1hQ4RwE

What's the big deal?

A deeper understanding of the huge wealth of information out there in the web

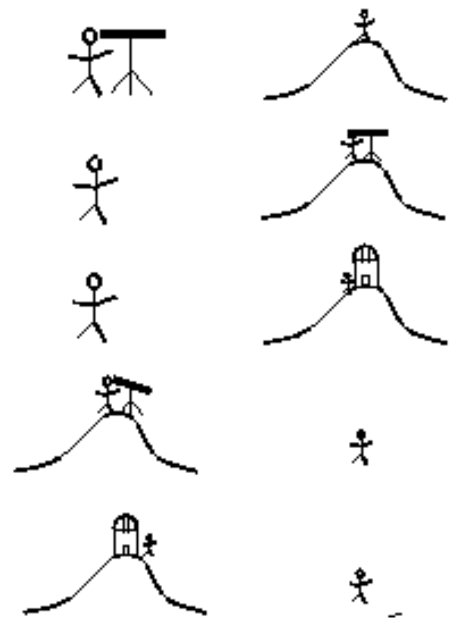
- But this "information out there" is in the free form text.
- How did Watson understand it and reason based on that understanding?

More generally,

- Can machine learn to understand language?
- Can machines perform what humans can (and more) when dealing with language?

Why is it difficult?

- Language is inherently ambiguous
  - ambiguity in words:
    - "Mary deposited the money in the bank" vs.
    - "Mary sat by the river bank".
  - ambiguity in sentences
    - I saw the man on the hill with a telescope.



- Language also expresses opinions, emotions, desires and wishes in addition to facts.

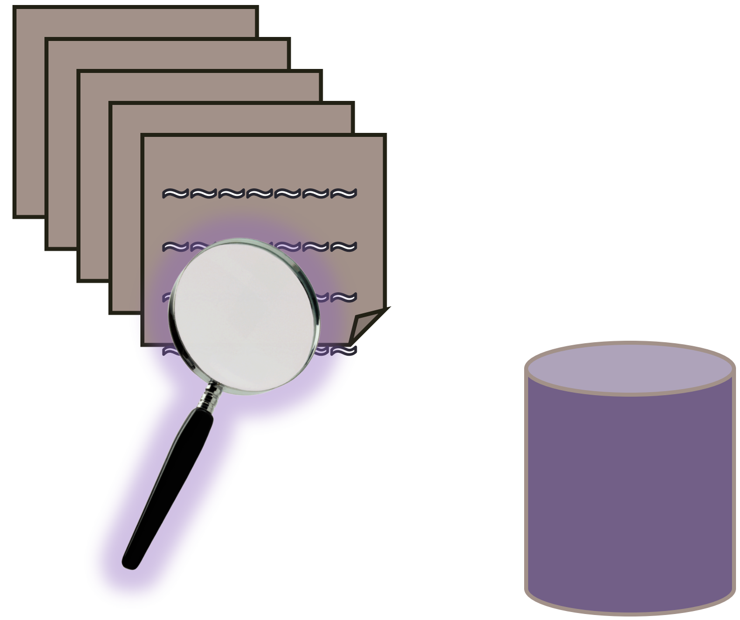
Outline

- Motivation
- **NLP Research Areas using ML**
  - **NLP Applications**
  - Fundamental NLP steps
- NLP at Columbia
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- Conclusion

NLP Applications

- **Information Extraction**

Extracting knowledge (facts, relations between entities, etc.) from unstructured text



1. Identify Entities
2. Coreference resolution
3. Identify relations

Information Extraction from Text

`Apple is headquartered in California. Tim Cook is its CEO.`

↓

`Based_in(Apple, California); CEO_of(Tim Cook, Apple)`

Information Extraction from Text

`Apple is headquartered in California. Tim Cook is its CEO.`

Apple (Org.) is headquartered in **California** (Loc.). **Tim Cook** (Per.) is its CEO.

Named Entity Tagging & Classifying

`Based_in(Apple, California); CEO_of(Tim Cook, Apple)`

Information Extraction from Text

Apple is headquartered in California. Tim Cook is its CEO.

Apple is headquartered in **California**. **Tim Cook** is its CEO.
(Org.) **(Loc.)** **(Per.)**

Coreference/Anaphora resolution

Based_in(Apple, California); CEO_of(Tim Cook, Apple)

Information Extraction from Text

`Apple is headquartered in California. Tim Cook is its CEO.`

Apple is **headquartered** in **California**. **Tim Cook** is <u>its</u> **CEO**.
(Org.) **(Loc.)** **(Per.)**

Relation Extraction

Based_in(Apple, California); CEO_of(Tim Cook, Apple)

NLP Applications

- Information Extraction
- Machine Translation

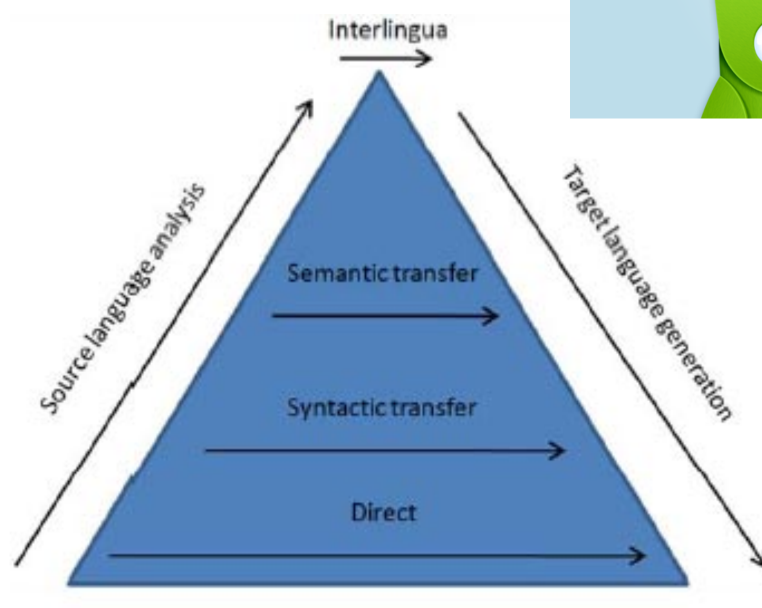


Figure 1: The Vauquois triangle

NLP Applications

- Information Extraction
- Machine Translation
- Question Answering

IBM Watson

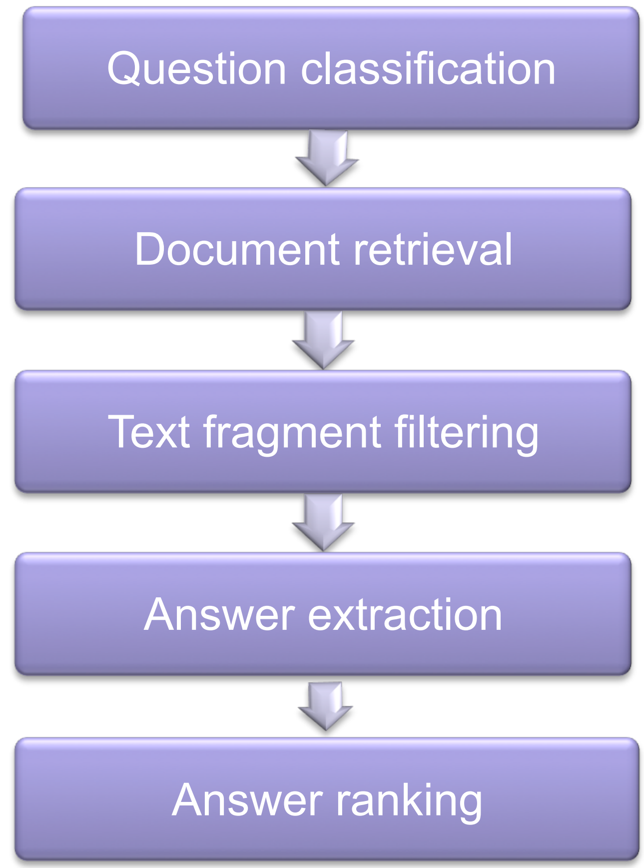
Clinical Q&A

What medical conditions does the drug "acetaminophen" contraindicate with?

Information Extraction + Reasoning

NLP Applications

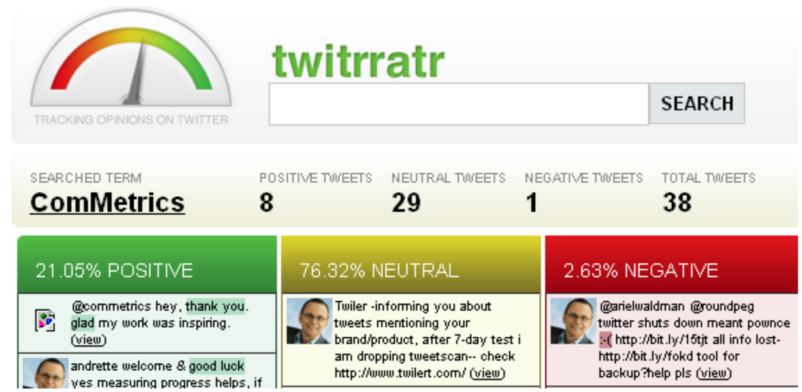
- Information Extraction
- Machine Translation
- Question Answering

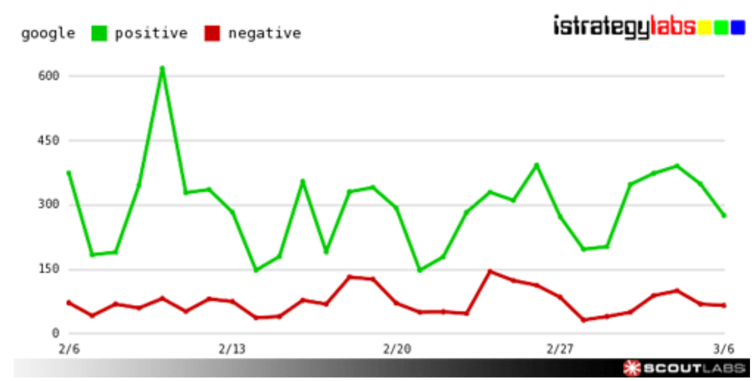


NLP Applications

- Information Extraction
- Machine Translation
- Question Answering
- Sentiment Analysis

- Determine the attitude or sentiment of the speaker/writer about a subject/topic/product





NLP Applications

- Information Extraction
- Machine Translation
- Question Answering
- Sentiment Analysis

Datasets

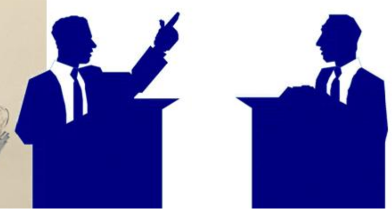
- Movie reviews (IMDB, …)
- Product reviews (Amazon etc.)
- Twitter

ML Approaches

- SVM
- Naïve Bayes
- MaxEnt
- Unsupervised approaches

NLP Applications

- Information Extraction
- Machine Translation
- Question Answering
- Sentiment Analysis
- Computational Socio-linguistics



Can we predict social relations between people based on how they interact?

NLP Applications

- **Information Extraction**
- **Machine Translation**
- **Question Answering**
- **Sentiment Analysis**
- **Computational Socio-linguistics**

Datasets

- Enron email corpus
  - Around 500,000 messages between Enron employees
- Online discussion forums
- Twitter/Facebook
- Offline discussions such as presidential debates, supreme court hearings

ML Approaches

- Social Network Analysis
- SVM/SVR

Outline

- Motivation
- **NLP Research Areas using ML**
  - NLP Applications
  - **Fundamental NLP steps**
- NLP at Columbia
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- Conclusion

Fundamental NLP steps

Parts-Of-Speech tagging:

```
Mary thinks Paris is beautiful.
```

```
Mary/NOUN thinks/VERB Paris/NOUN is/VERB beautiful/ADJ ./.
```

Datasets

- English Penn Treebank (WSJ) (7 million words POS tagged)

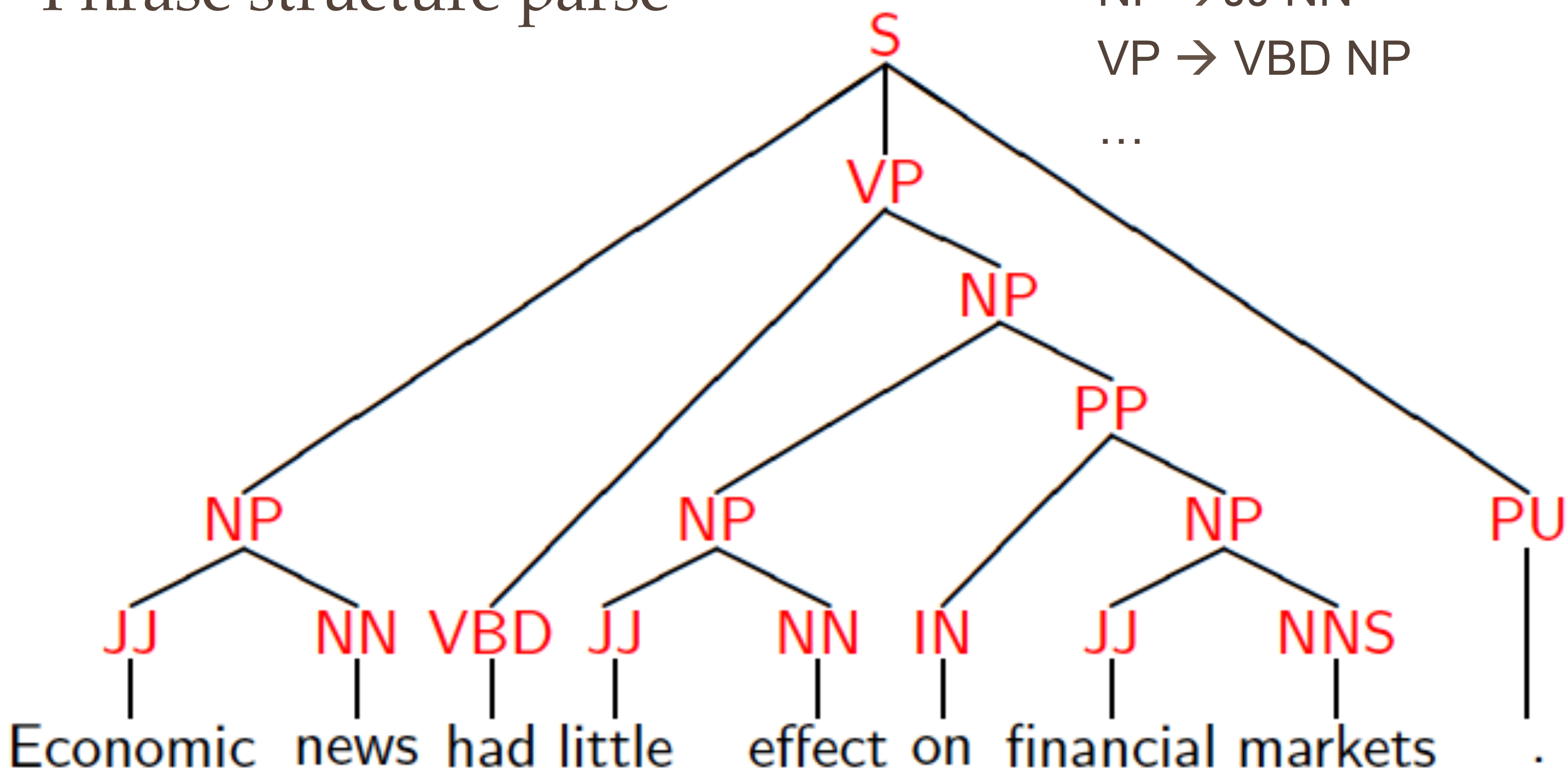
Approaches

- SVM, HMM, MEMM, Perceptron
- Maximum entropy cyclic dependency network (Stanford Tagger)
  - 97.32% accuracy on seen words; 90.79% on unseen words
  - http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)

Fundamental NLP steps

Parsing

- Phrase structure parse



Context Free Grammar (CFG)

S →NP VP PU

NP →JJ NN

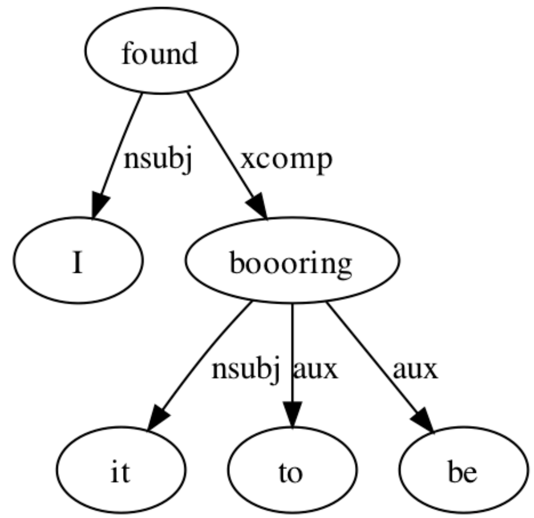
VP → VBD NP

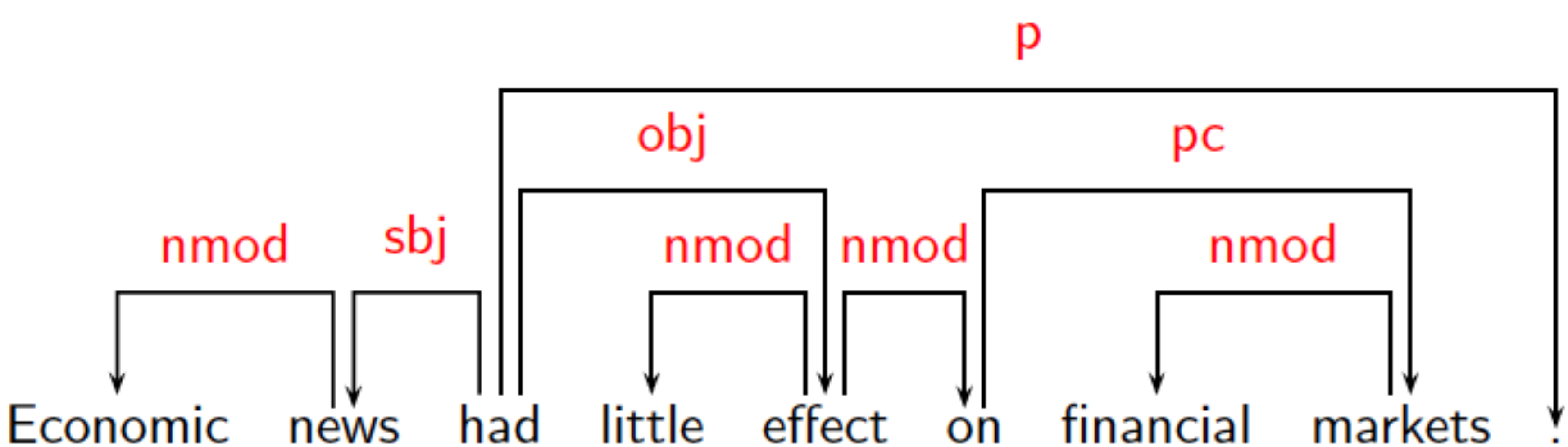
…

Fundamental NLP steps

Parsing

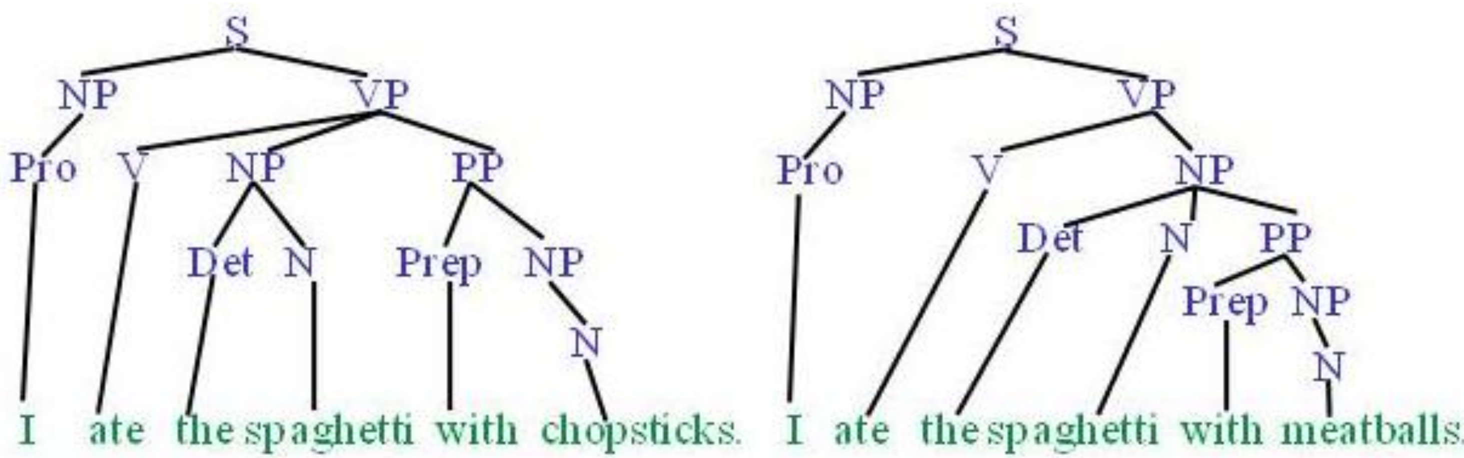
- Phrase structure parse
- Dependency Parse



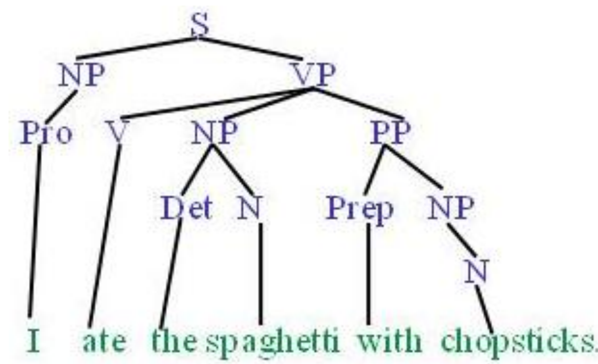


Fundamental NLP steps

Parsing ambiguity



Fundamental NLP steps



Phrase Structure Parsing

Datasets

- English Penn Treebank (WSJ)

Approaches

- PCFG (Probabilistic CFG), Reranking
- Lexicalized PCFG + self training on 2 million raw sentences
  - 92% accuracy
  - http://aclweb.org/aclwiki/index.php?title=Parsing_(State_of_the_art)

Outline

- Motivation
- NLP Research Areas using ML
  - NLP Applications
  - Fundamental NLP steps
- **NLP at Columbia**
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- Conclusion

NLP at Columbia

Research labs

- NLP lab
- Speech Lab
- CCLS (Center for Computational Learning Systems)

Faculty

- Prof. Kathy McKeown, Prof. Julia Hirschberg, Prof. Michael Collins (CS Dept.)
- Dr. Owen Rambow, Dr. Nizar Habash, Dr. Becky Passonneau (CCLS)

Courses

- COMS 4705 – Natural Language Processing (mostly in the Fall)
- COMS 6998 – ML for NLP (mostly in Spring)
- COMS 6998 – Machine Translation (mostly in Spring)

NLP at Columbia

Research areas

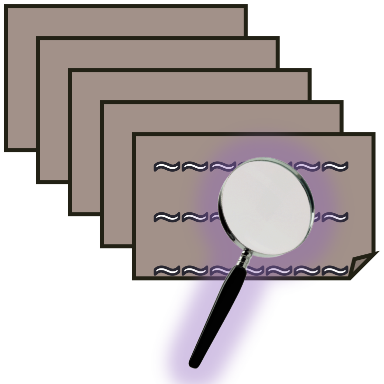
- ML methods for Parsing/Tagging etc.
- Semantics
- Machine Translation
- Arabic NLP
- Social/Interaction analysis (WISR)
- Speech analysis – transcription, analysis
- Text summarization, generation

Outline

- Motivation
- NLP Research Areas using ML
  - NLP Applications
  - Fundamental NLP steps
- NLP at Columbia
- **Relation Extraction**
  - **Supervised Relation Extraction**
  - **Distant Supervision**
- Conclusion

Relation Extraction – the what?

Given a pair of entities *e1* and *e2* and a corpus C of documents/sentences, what is the relation between *e1* and *e2*?



Given a sentence s containing two entities e1 and e2, what relation between e1 and e2 is expressed in s?

"Apple is headquartered in California"

- *based_in(Apple, California)*

"IBM was incorporated in the State of New York on June 16, 1911, as the Computing-Tabulating-Recording Co. (C-T-R)…"

- *founding_year(IBM, 1911)*
- *founding_location(IBM, New York)*

Relation Extraction – the why?

- Converting the "huge wealth of information" out there in the web in unstructured form → structured data (building knowledge bases)
- Extending existing knowledge bases
  - Freebase
  - DBPedia
  - UMLS
- Aid question answering systems (Watson, Medical expert systems etc.)
  - The granddaughter of which actor starred in the movie "E.T."?
  - acted-in(?x,"E.T.") & is-a(?y, actor) & granddaughter-of(?x,?y)
    - x: Drew Barrymore; y: John Barrymore

What kind of relations?

ACE Annotations

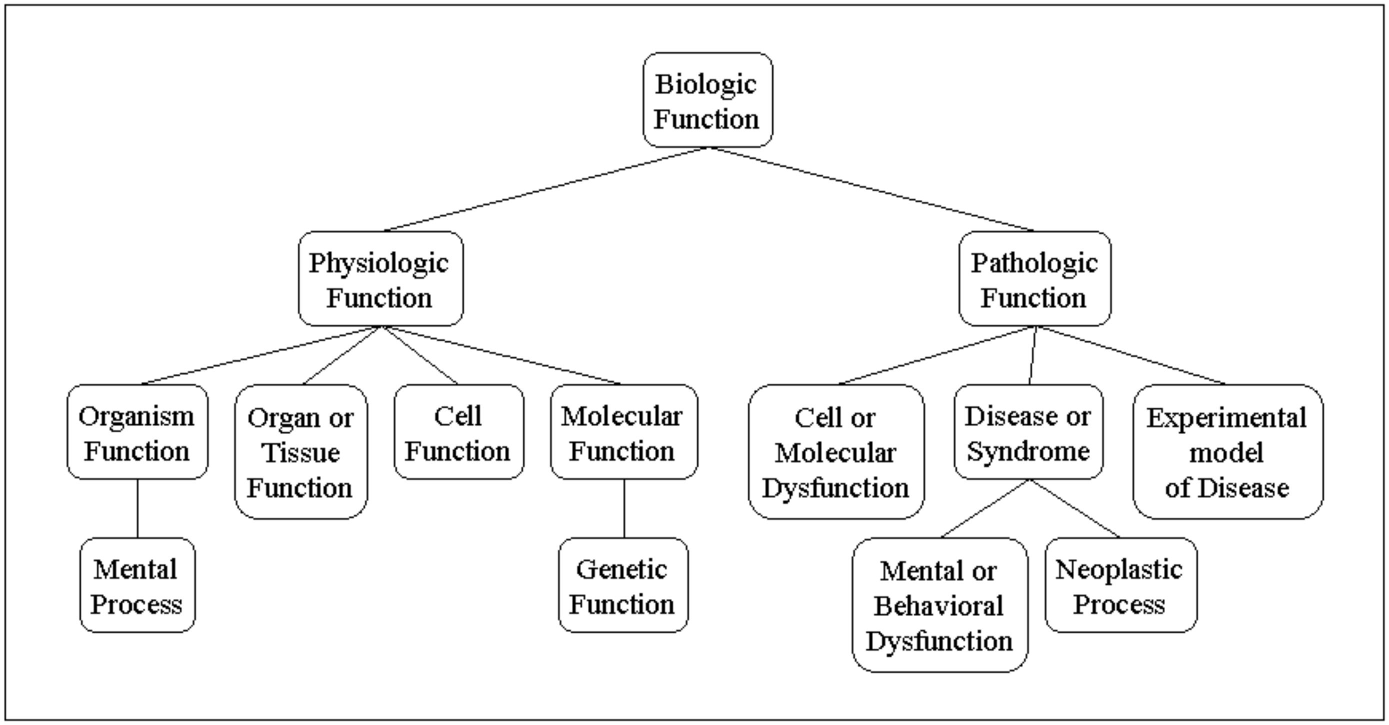
- Captures relations between 5 types of entities --- Person, Organization, Geo Political Entity, Location, Facility.
- 24 different relations in 5 categories
- Around 100K-300K words per language (English/ Chinese/ Arabic) in ACE2005

AT	NEAR	PART	ROLE	SOCIAL
Based-In Located Residence	Relative-location	Part-of Subsidiary Other	Affi liate, Founder Citizen-of, Management Client, Member Owner, Other, Staff	Associate, Grandparent Parent, Sibling Spouse, Other-professional Other-relative, Other-personal

What kind of relations?

UMLS (Unified Medical Language System)

- 134 types of entities --- Drug, Disease, Treatment, Enzyme, etc.



- 54 different relations --- DIAGNOSES, TREATS, PREVENTS, etc.

What kind of relations?

Open domain relations

- DBPedia / Wikipedia Info boxes
  - over 1 billion relation instances
- Freebase relations
  - politics, biology, films, business
  - over 116 million instances, 7300 relations, 9 million entities

RE Approaches

Rule based Systems

Extracting patterns using lexical/ syntactic regular expressions

Patterns capturing "is_a(X,Y)" relation

- Y such as X
- such Y as X
- X (and | or) other Y
- Y including X
- Y, especially X
- …

Issues:

- High precision, but low recall
- Manual labor in collecting patterns

RE Approaches

Rule based Systems

Extracting patterns using lexical/ syntactic regular expressions

Supervised Learning

1. Feature based methods
2. Kernel based methods

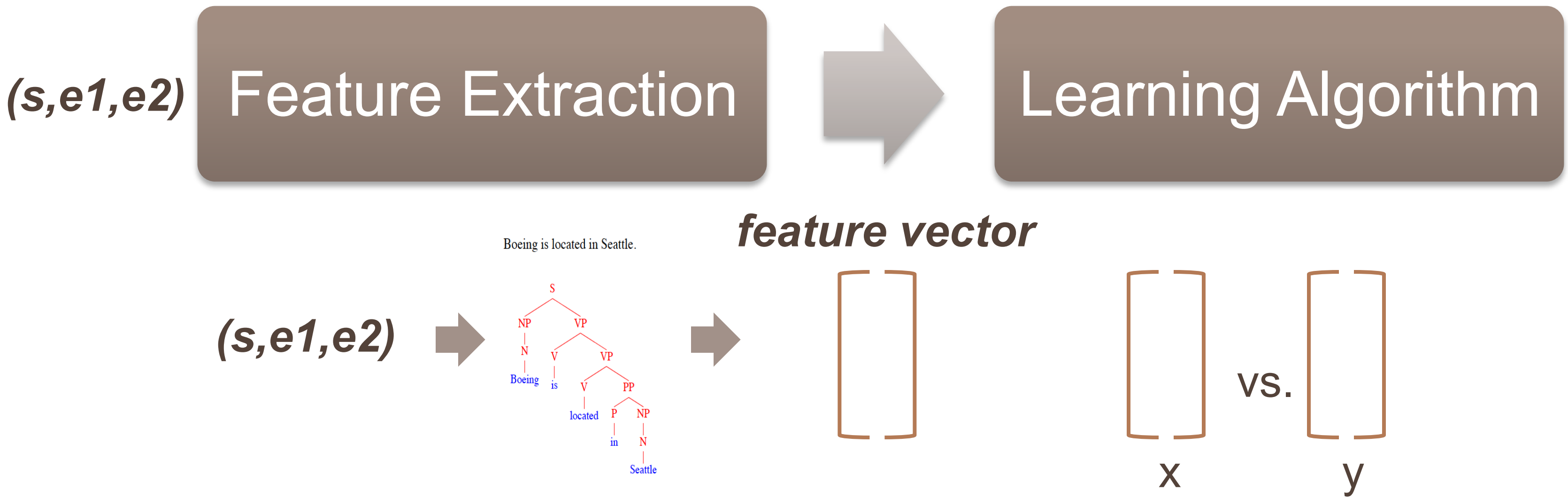
Traditional supervised approaches

- Choose a set of types of entities and relations to capture. Choose appropriate dataset and label relations
- Convert each relation instance to an appropriate representation (e.g. feature vector)
- Apply an appropriate learning algorithm to build a classifier. E.g. MaxEnt, Naïve Bayes, SVM

Issues

- Expensive to label data
- Do not generalize well across genres

Feature based approaches



- F_{based_in}(T("Apple is headquartered in Cupertino", Apple, Cupertino)) = +1
- F_{based_in}(T("Apple is based out of California", Apple, California)) = +1
- F_{based_in}(T("Apple did not break California law ", Apple, California)) = -1

Kernel Functions

- Kernel function K(x,y) finds the similarity between x and y
- If x, y are represented as feature vectors $\Phi(x)$, $\Phi(y)$
  - E.g., linear kernel → $\Phi(x).\Phi(y)$

Typical features for Relation Extraction

Apple is headquartered in **Cupertino**.
(Org.) **(Loc.)**

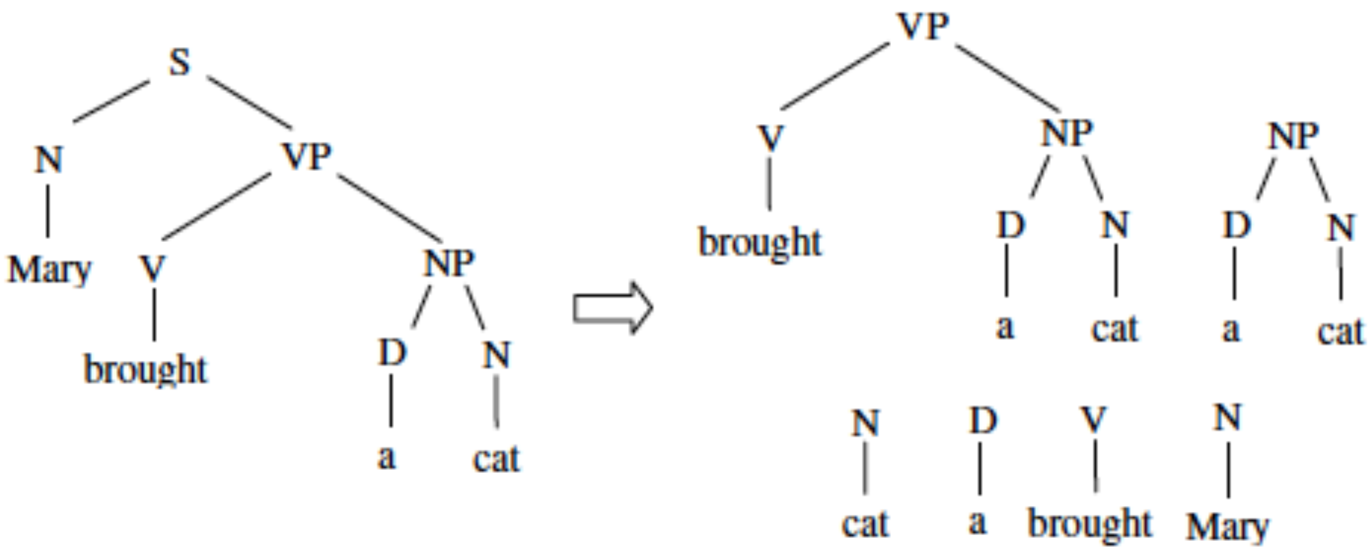
T(s,e1,e2) =

is e1 before e2?	***1***
type of e1?	***ORG***
type of e2?	***LOC***
# words in between?	***3***
words between?	***{is, headquartered, in}***
words before?	***{}***
words after?	***{}***
…	

Kernel Functions

- Kernel function K(x,y) finds the similarity between x and y
- If x, y are represented as feature vectors Φ(x), Φ(y)
  – E.g., linear kernel → Φ(x).Φ(y)
- A better way since the x and y have underlying structure – tree, graph etc.?
  – Perform feature engineering to find best set of features Φ()
    – "have_a_VERB_parent", "have_an_ADJ_child" etc.
  – Define new kernel functions to directly apply on x and y
    – Convolution kernels: string kernels, tree kernels etc.

Tree Kernels

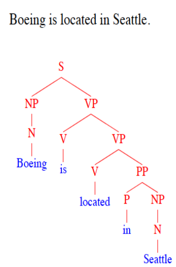


- Kernel function K(Tx,Ty) can be designed to find similarities that is relevant to the task at hand.
  - E.g.: counting the common subtrees with a decay factor associated with the subtree size

Feature based approaches

(s,e1,e2) **Feature Extraction** → **Learning Algorithm**

feature vector

(s,e1,e2) → 

→ [ ]

[ ] vs. [ ]
x y

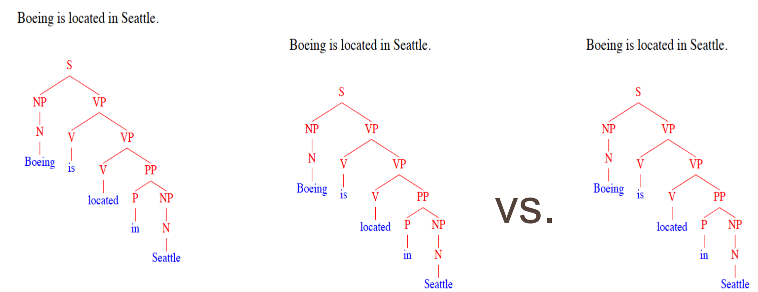
- F_{based_in}(T("Apple is headquartered in Cupertino", Apple, Cupertino)) = +1
- F_{based_in}(T("Apple is based out of California", Apple, California)) = +1
- F_{based_in}(T("Apple did not break California law ", Apple, California)) = -1

Kernel based approaches

(s,e1,e2) Input transformation → Learning Algorithm

structured representation

(s,e1,e2) →



Tx Ty

- F_{based_in}(T("Apple is headquartered in Cupertino", Apple, Cupertino)) = +1
- F_{based_in}(T("Apple is based out of California", Apple, California)) = +1
- F_{based_in}(T("Apple did not break California law ", Apple, California)) = -1

Tree Kernel based approaches

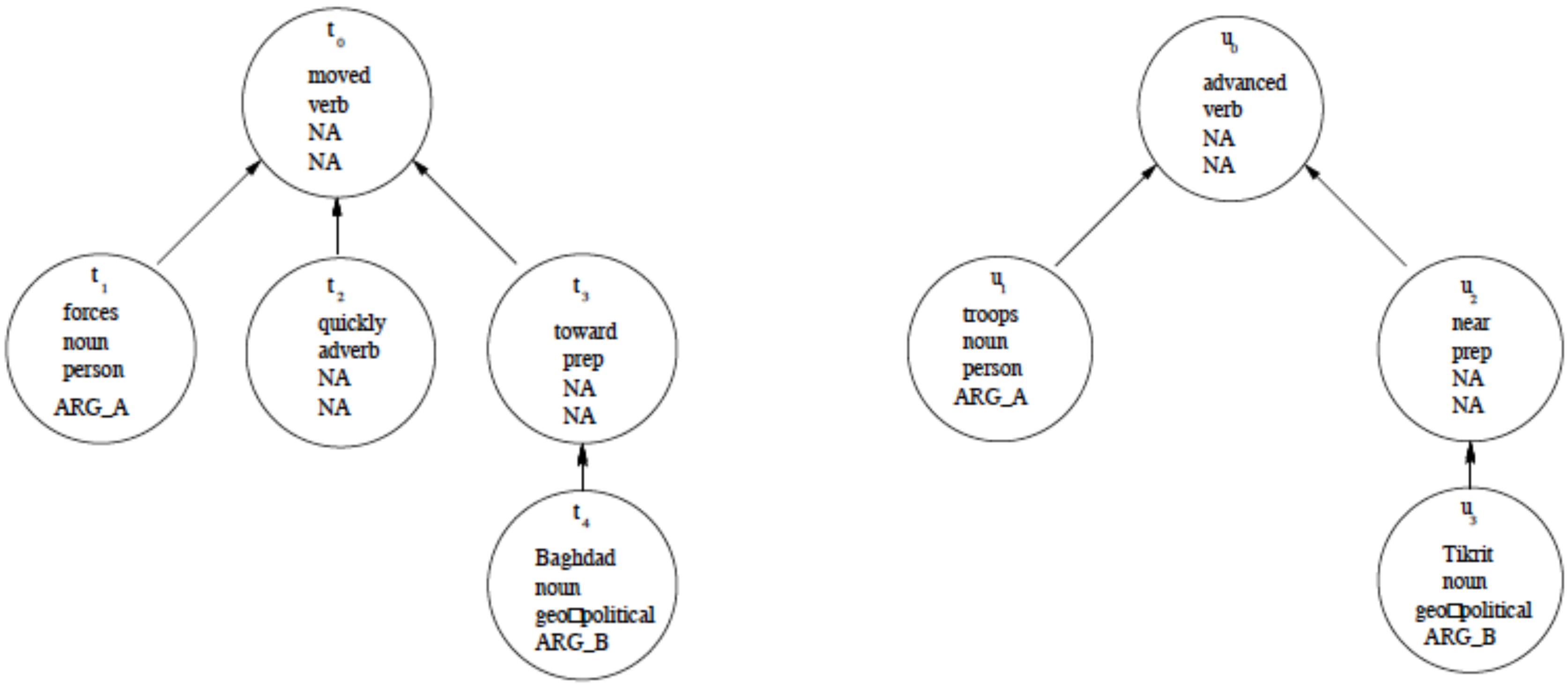
Tree Kernels in NLP

- Collins and Duffy 2002 (Parsing)
- Cumby and Roth 2003 (NER)
- Moschitti 2004 (Semantic Parsing)

Tree Kernels in Relation Extraction

- Parse (shallow) tree kernel (Zelenko et al. '03)
- **Dependency tree kernel (Culotta and Sorenson, 2004)**
- Shortest dependency path kernel (Bunescu & Mooney '05)

Culotta and Sorenson, 2004



K(Tx,Ty) = 0, if root node's POS & TYPE & ARG does not match

= sim(r1, r2) + Kc(children(r1), children(r2))

Kc(children(r1), children(r2)) is found by summing over K(c1,c2) over all children recursively, with a decay factor

Issues with supervised approaches

- Expensive to label data with relations
- Difficult to extend to new relation types and domains

Other alternatives?

- Unsupervised approaches?
- Semi supervised approaches?
  - Distance supervision

Distance supervision

- For each relation r in R (e.g.: may_treat)
- For each entity pair (e1, e2) such that r(e1,e2) in D
  - (e.g. <hypertension, acebutolol>; <fever, acetaminophen>; …)
- Extract the set of sentences containing both e1 and e2
  - Acebutolol in the treatment of patients with hypertension
  - After treatment with acetaminophen, fever subsided
  - Either acetaminophen or ibuprofen can be given to treat the fever
  - …
- Use features from all sentences to build the training/test instance
- E.g.: **Mintz et al. 2009** (Freebase relations; about 100 relations)

Distance supervision - Issues

- What about negative examples?
  - All sentences with entity pairs that are not related by r?
  - All sentences with entity pairs that are not related at all?
  - Exponentially large negative examples; How to sample?
- What about the distant supervision base assumption?
  - "Tylenol **treats** acute pain" vs.
  - "Its a pain to get Tylenol"
- What about multiple relations?
  - "Barack Obama was born in the US" vs.
  - "Obama was reelected as US President in 2012" vs.
  - "Obama proposed a new US healthcare bill"

More recent approaches

Riedel et al 2010

- Multiple Instance Learning in Distant Supervision
  - If two entities participate in a relation, **at least one sentence** that mentions these two entities might express that relation.

Hoffmann et al. 2011, Surdeanu et al. 2012

- Modeling multiple instance multi label (overlapping) relations

Wang et al 2011, EMNLP

- Relation Extraction with Relation Topics

Outline

- Motivation
- NLP Research Areas using ML
  - NLP Applications
  - Fundamental NLP steps
- NLP at Columbia
- Relation Extraction
  - Supervised Relation Extraction
  - Distant Supervision
- **Conclusion**

Conclusion

- NLP Applications
  - Information Extraction
  - Machine Translation
  - Question Answering
  - …
- Relation Extraction
  - Supervised Feature based methods
  - Supervised Kernel based methods
  - Semi supervised distant supervision

Questions?