Research Group
Databases and Information Systems

Natural Language Processing

Günther Specht
Eva Zangerle

Summer Term 2019

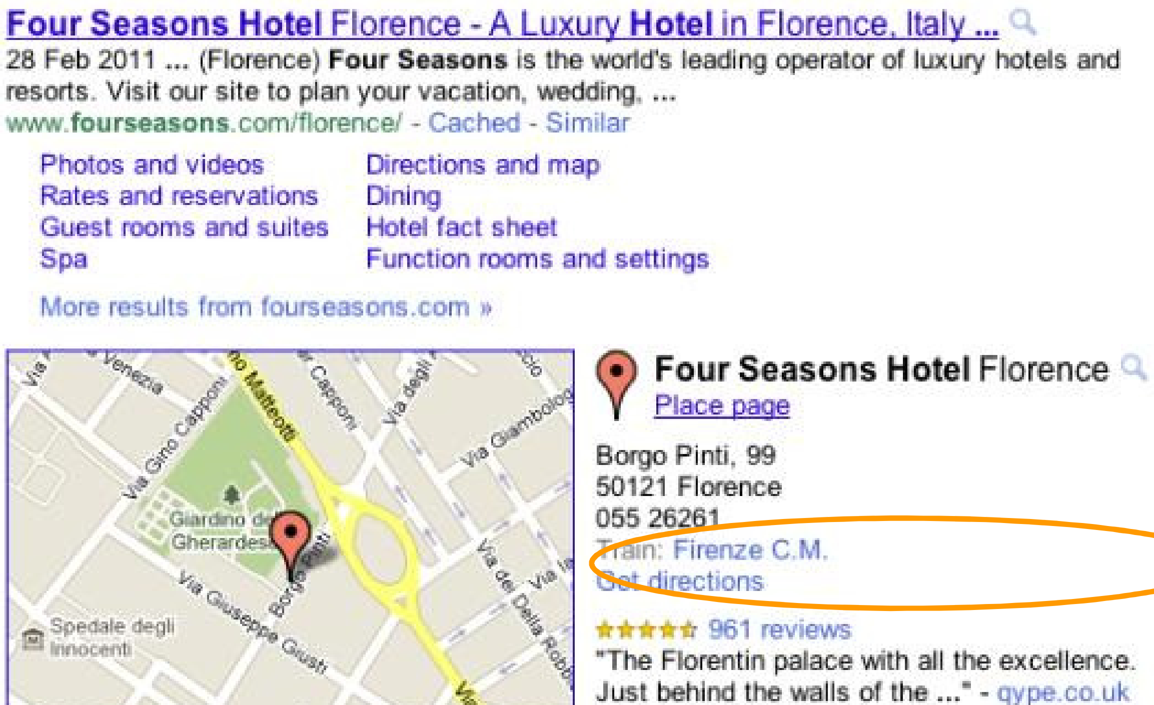
Motivation

Why Natural Language Processing?

- Huge amounts of text are available on the Internet (> 20 billion pages) and company intranets
- Processing such a large amount of text is vital to a number of applications:
  - Indexing and search
  - Text categorization
  - Information extraction, knowledge acquisition
  - Automatic translation and summarization
  - Automatic question answering
  - Text generation
  - Speech understanding, Human-computer dialog

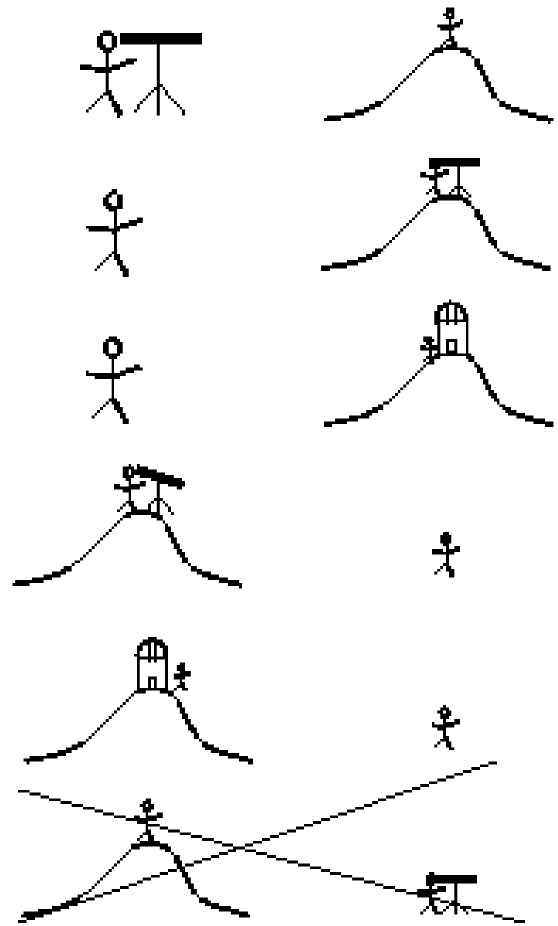
Linguistic data is ubiquitous

- Most of the information around companies & the Web comes in human languages – not traditional DB stuff!
  - reports, customer email,
  - web pages, sound, video,
  - opinions, feedback



Why is NL Understanding Difficult?

- *Ambiguity* is the primary difference between natural language (NL) and computer languages (CLs)
  - CLs are designed by grammars that produce a unique parse for each sentence in the language
- Examples of ambiguous NL wordings
  - I saw the man on the hill with a telescope.
  - I saw the Grand Canyon flying to LA.
  - Time flies like an arrow.
  - Fruit flies like banana.



Resolving Ambiguity

- The hidden structure of language is highly ambiguous at different levels: lexical, syntactic, semantic

Part of speech ambiguities

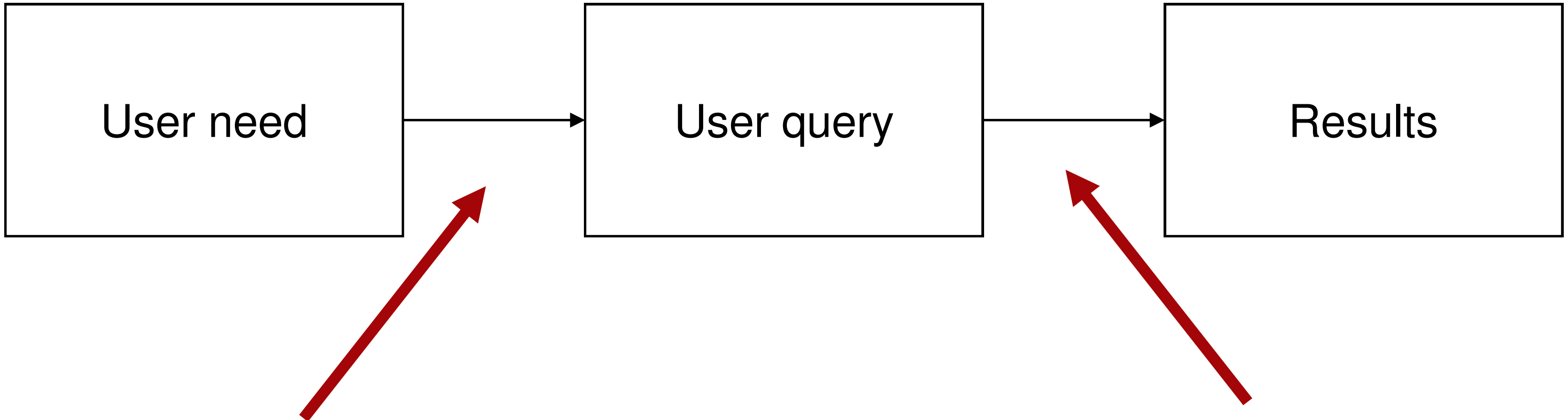
Syntactic attachment ambiguities

		VB					
	VBZ	VBP	VBZ				
NNP	NNS	NN	NNS	CD	NN		
Fed	*raises*	*interest*	*rates*	*0.5*	*%*	*in*	*effort*
						to	*control*
							inflation

Word sense ambiguities: Fed → *"federal agent"*

interest → *a feeling of wanting to know or learn more*

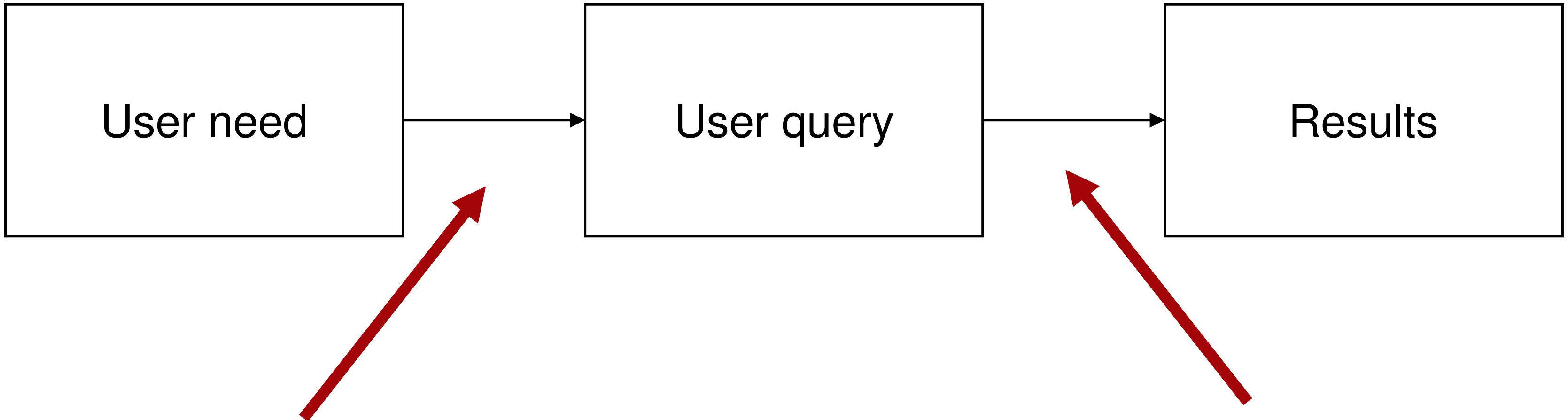
Translating User Needs: Databases



For RDB, a lot of people know how to do this correctly, using SQL or a GUI tool

The answers coming out here will then be precisely what the user wanted

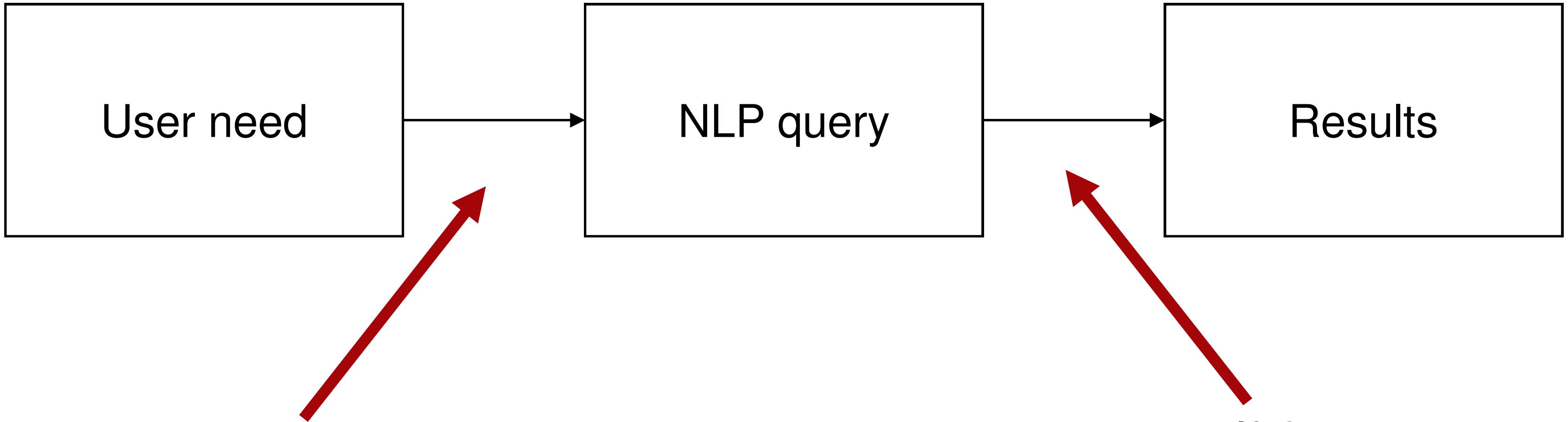
Translating User Needs: Information Retrieval



For meanings in text,
no IR-style query
gives you exactly
what you want;
it only hints at it

The answers
coming out may
be roughly what
was wanted, or
can be refined

Translating User Needs: NL Processing



For a deeper NLP analysis system, the system subtly translates the user's language

If the answers coming back aren't what was wanted, the user frequently has *no idea* how to fix the problem
Risky!

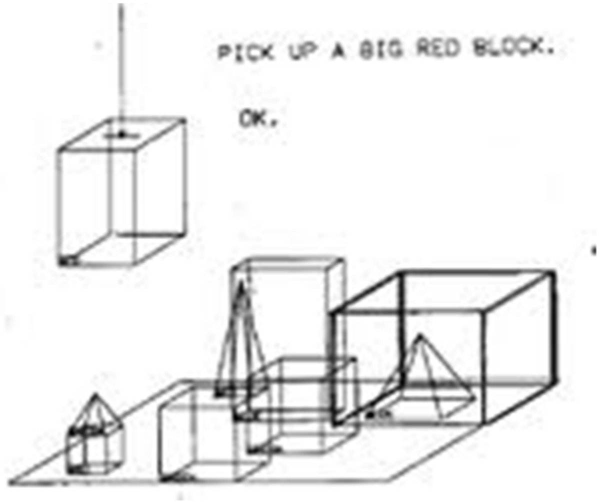
Basic concepts

Natural Language Processing

- NLP is the branch of computer science focused on the interaction between computers and natural languages
  - Born as a spin-off of Artificial Intelligence
  - Nowadays, several NLP methods and applications are very related to IR
- NLP "counterpart" in linguistics: Computational Linguistics
  - More on the linguistic/cognitive side of the problem

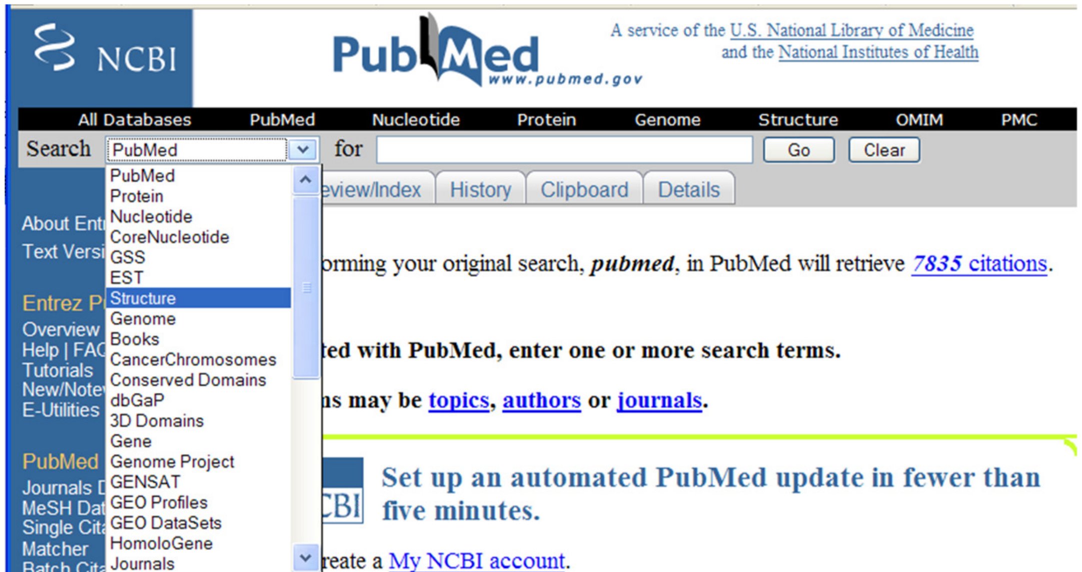
A brief history of NLP: 1950s-1970s

- 1950s: the Turing test
  - Soon enough, machines would be mistaken for humans
- 1960s-1970s:
  - Thanks to simple pattern-matching rules, ELIZA the chat-bot was able to converse in NL [Weizenbaum,1966]
  - In the SHRDLU world, a mechanical hand would receive commands in NL to move blocks around [Winograd,1971]
  - "Conceptual ontologies" to represent knowledge in restricted domains, rule-based approaches to NL understanding [Schank & Abelson, 1977]



A brief History of NLP: 1980s-1990s

- Machine Learning & statistical models emerge
  - Decision trees, Support Vector Machines, Hidden Markov Models, …
  - Syntactic parsers, Named Entity recognizers trained on large datasets [Finkel et al, 2005]
  - Large news/medical corpora made available to test algorithms using deep NL features
    - –New York Times, Wall Street Journal
    - –Medline, PubMed

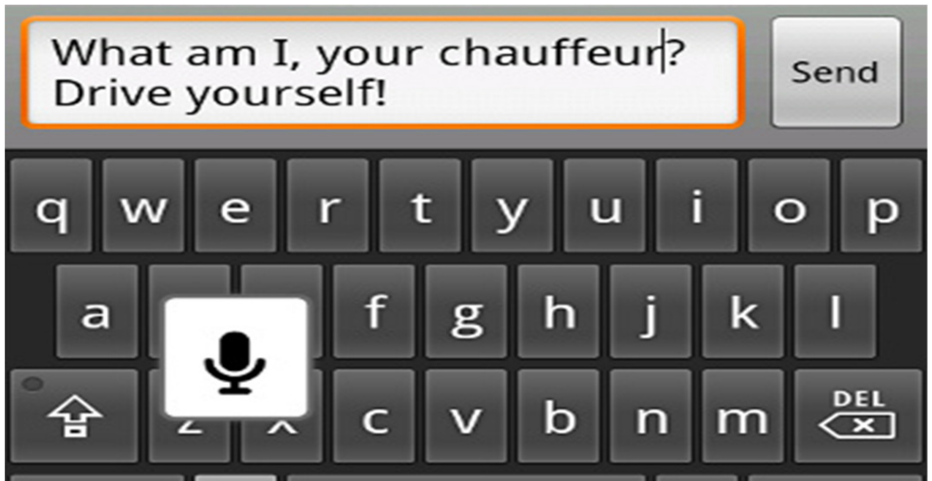


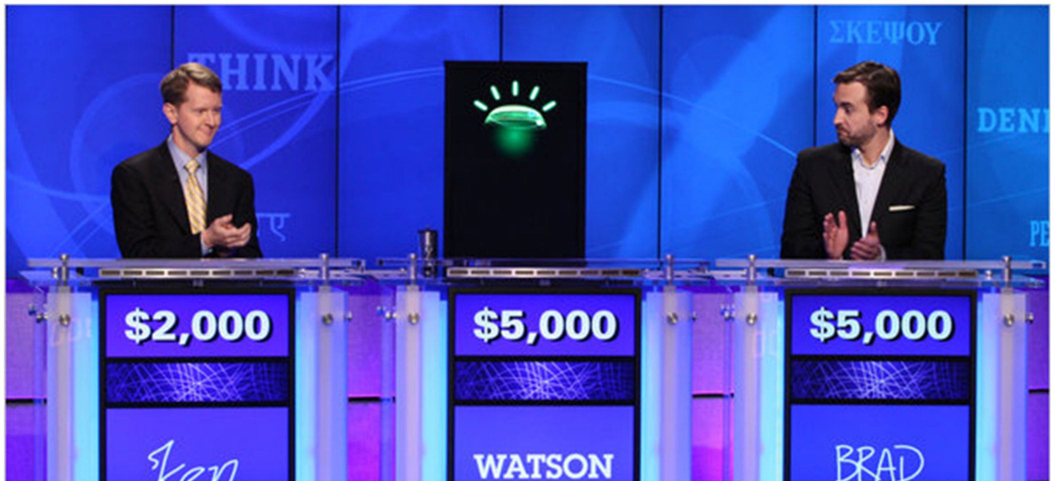
A brief History of NLP: 1990s-2000s

- Great evaluation campaigns
  - TREC (trec.nist.gov), CLEF (clef-campaign.org)
  - Challenging tasks such as word sense disambiguation, summarization, question answering, machine translation
- "Deeper" NLP:
  - Semantic Role Labeling shifts analysis from syntax to semantics [Carreras & Marquez, 2005]
- Industrial mobile NL technologies make NLP more and more robust
  - Automatic speech recognition and understanding
  - AT&T's *How May I Help You?* [Gorin et al.,1997]

A brief History of NLP: today

- Machine Learning methods are the rule
  - discriminative methods such as Support Vector Machines, Conditional Random Fields have proven their efficiency for complex tasks
- Challenging problems
  - answering complex questions (Watson wins *Jeopardy!*)
  - machine translation (cf. Google translate)
- Great effort on non-text
  - *speech* understanding now makes it possible to have you speak to your smartphone





Summary

- Definitions
- Methods
- Evaluation
- Conclusions

Levels of Natural Language Understanding

Syntax, Semantics, Pragmatics

- **Syntax** concerns the proper ordering of words and its effect on meaning.
  - *The dog bit the boy* != *The boy bit the dog*.
- **Semantics** concerns the (literal) meaning of words, phrases, and sentences.
  - "plant": a photosynthetic organism, a manufacturing facility, the act of sowing
- **Pragmatics** concerns the overall communicative and social context and its effect on interpretation.
  - *Remove the kernels from the cherries and throw them away*

Syntactic Tasks

- Morphological Analysis
- Part of Speech Tagging
- Shallow Parsing
- Deep Syntactic Parsing

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words.
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning
  - E.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried $\Rightarrow$ carry + ed (past tense)
  - independently $\Rightarrow$ in + (depend + ent) + ly
  - Googlers $\Rightarrow$ (Google + er) + s (plural)
  - unlockable $\Rightarrow$ un + (lock + able) ? (un + lock) + able ?

Part Of Speech (POS) Tagging

- Part of Speech (POS): a lexical category determined by morphological behavior of the word
- POS tagging: annotation of each word in a sentence with a POS

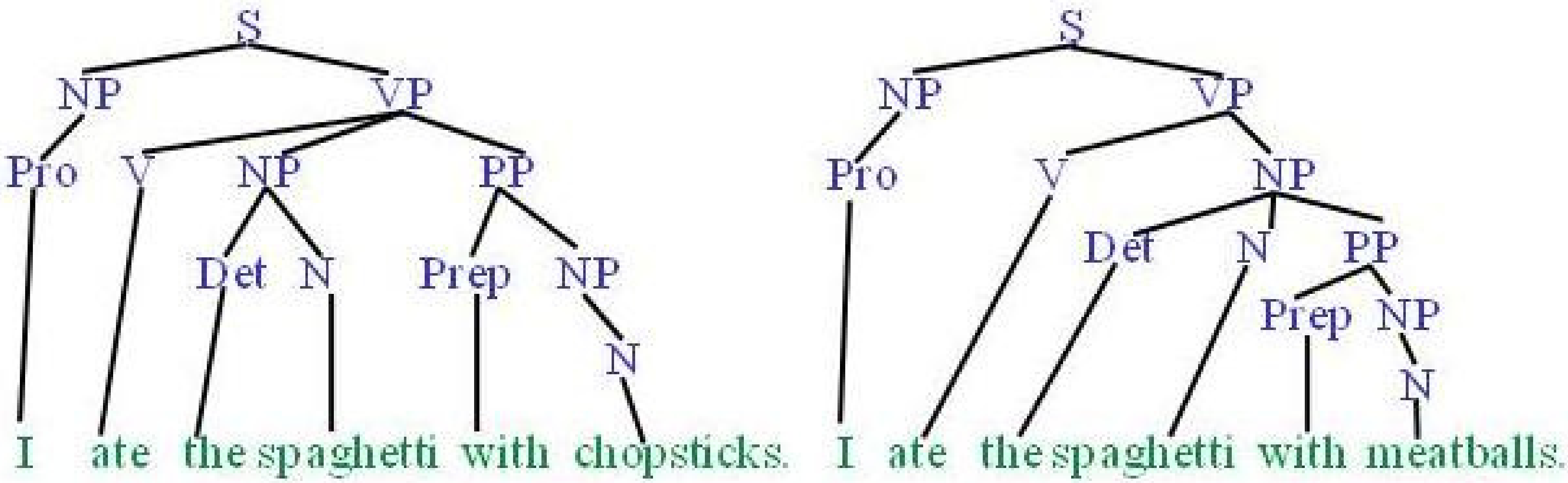
  I$_{\text{Pronoun}}$ **ate**$_{\text{Verb}}$ **the**$_{\text{Determiner}}$ **spaghetti**$_{\text{Noun}}$ **with**$_{\text{Preposition}}$ **meatballs**$_{\text{Noun}}$**.**
- Useful for subsequent tasks such as word sense disambiguation
- POS taggers exist for most languages, even least researched ones
- POS tagging is considered to be a “solved problem”, with > 90% accuracy using data-driven techniques
  - “Classic” approach: decision trees [Schmid’94]

Phrase Chunking (aka Shallow Parsing)

- A shallow subdivision of a sentence into its main constituents or phrases
- Main phrase types:
  - noun phrases (NPs), verb phrases (VPs), prepositional phrases (PPs)
- Example of chunked output:
  - [I]$_{NP}$ [drink]$_{VP}$ [my coffee]$_{NP}$ [with two sugars]$_{PP}$.
  - [He]$_{NP}$ [reckons]$_{VP}$ [the current account deficit]$_{NP}$ [will narrow]$_{VP}$ [to only $1.8 billion]$_{PP}$ [in September]$_{PP}$
- Many libraries exist for chunking, e.g. OpenNLP from Stanford (opennlp.sourceforge.net)
- Usually require a segmentation part (separate phrases from each other) and a tagging part (tag separated phrases)

Syntactic Parsing

- Produce a representation of the syntactic roles played by words in a sentence (generally in the form of a *parse tree*).
- Typical method: associate grammar rules with probabilities to decide among different interpretation options



Semantic Tasks

- Word Sense Disambiguation
- Semantic Role Labeling
- Recognizing Textual Entailment

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- WSD is the task of automatically assigning the most likely meaning of each word in a sentence
- Useful in many NLP applications such as automatic question answering, machine translation
- Methods:
  - Dictionaries (Lesk method: similar meaning based on overlap in dictionary definitions)
  - Machine learning methods, both supervised (e.g. Support Vector Machines) and unsupervised (clustering word occurrences in sentence)

Semantic Role Labeling (SRL)

- SRL: labeling phrases of a sentence with semantic roles with respect to a target word (generally the verb)
  - Also called "shallow semantic parsing"
- Examples of semantic roles:
  agent patient source destination instrument
- Examples of parsed output:
  - [John]$_{agent}$ drove [Mary]$_{patient}$ from [Austin]$_{source}$ to [Dallas]$_{destination}$ in [his Toyota Prius]$_{instrument}$.
  - [The hammer]$_{instrument}$ broke [the window]$_{patient}$.
- This task is deeply dependent on the understanding of syntactic relations between words in the sentence (syntactic parsing) [Carreras & Marquez'05]

Textual Entailment

- Determine whether one natural language sentence implies another under an ordinary interpretation.
- Example:

TEXT	HYPOTHESIS	ENTAILMENT
Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc. last year.	*Yahoo bought Overture.*	TRUE
Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.	*Microsoft bought Star Office.*	FALSE
The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.	*Israel was established in May 1971.*	FALSE
Since its formation in 1948, Israel fought many wars with neighboring Arab countries.	*Israel was established in 1948.*	TRUE

Pragmatic/Discourse tasks

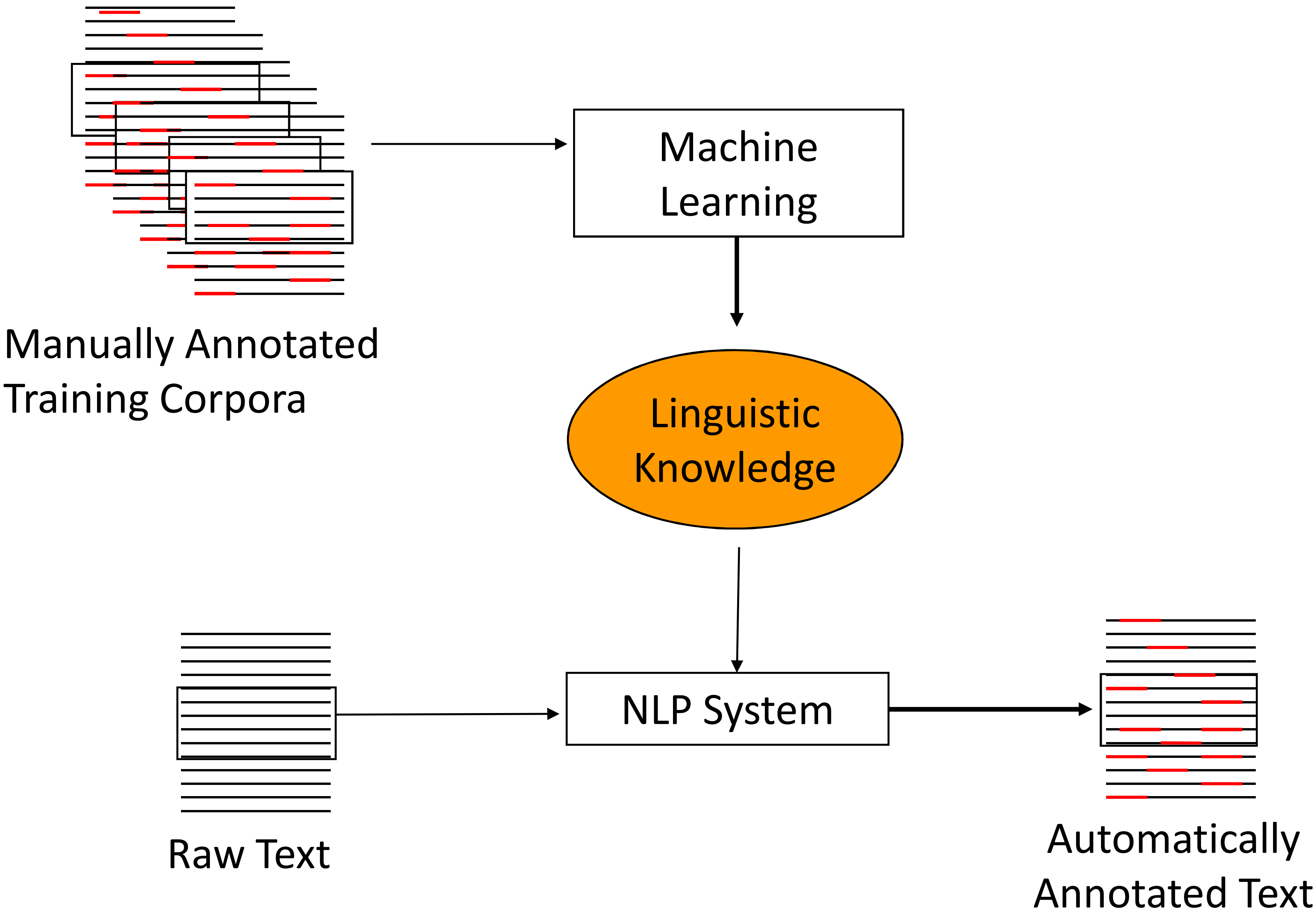
- Anaphora: an instance of an expression referring to another
- Co-reference occurs when multiple expressions in a sentence or document have the same referent (i.e. refer to the same phrase).
- Anaphora/co-reference resolution consists in determining which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Ellipsis is the omission or suppression of parts of words or sentences when these can be inferred from the context
  1. Wise men talk because they have something to say; fools because they have to say something (Plato)
  2. Wise men talk because they have something to say; fools talk because they have to say something (Plato)

Methods

NLP Methods

- Two (not necessarily competing) ways to process natural language
- Rule-based: follow linguistically-motivated rules/apply manually acquired resources (e.g. dictionaries) to classify and interpret NL
- Machine learning: use data to drive the inference of patterns and regularities in NL
  - Data description often derives from linguistically-motivated
  - Allow to discover rules!

Machine Learning Approach



Advantages of the Learning Approach

- Larger and larger amounts of text are available
- Annotating corpora is easier and easier and requires less expertise than manual knowledge engineering
- Algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge
- The probabilistic knowledge acquired is robust enough to handle linguistic regularities as well as exceptions

Another Method:

Natural Language Parsing
using
Logic Programming

NL Parsing via LP

Natural Language Parsing via Logic Programming
(as easy Specification, in Prolog, in Deductive DBS,..)

1. Syntactic Analysis via Position Identifiers

Example:

John	saw	a	man	with	a	mirror
0 1	1 2	2 3	3 4	4 5	5 6	6 7
noun	verb	det	noun	prep	det	noun

- Grammer rules:

s(X, Y, s(S1,S2)) :- np(X, Z, S1), vp(Z, Y, S2).

np(X, Y, np(S)) :- noun(X, Y, S).

np(X, Y, np(S1,S2)) :- det(X, Z, S1), noun(Z,Y,S2).

vp(X, Y, vp(S1,S2)) :- ver(X, Y, S1), np(Z, Y, S2).

…

Using Position Identifier

- Syntactic Analysis via Position Identifiers

- Classes of primary words:

  noun(From, To, noun(X)) :- word(From, To, X), db_noun(X, C, G, N)).

  …

  - Input:

    word(0,1,john).

    word(1,2,saw).

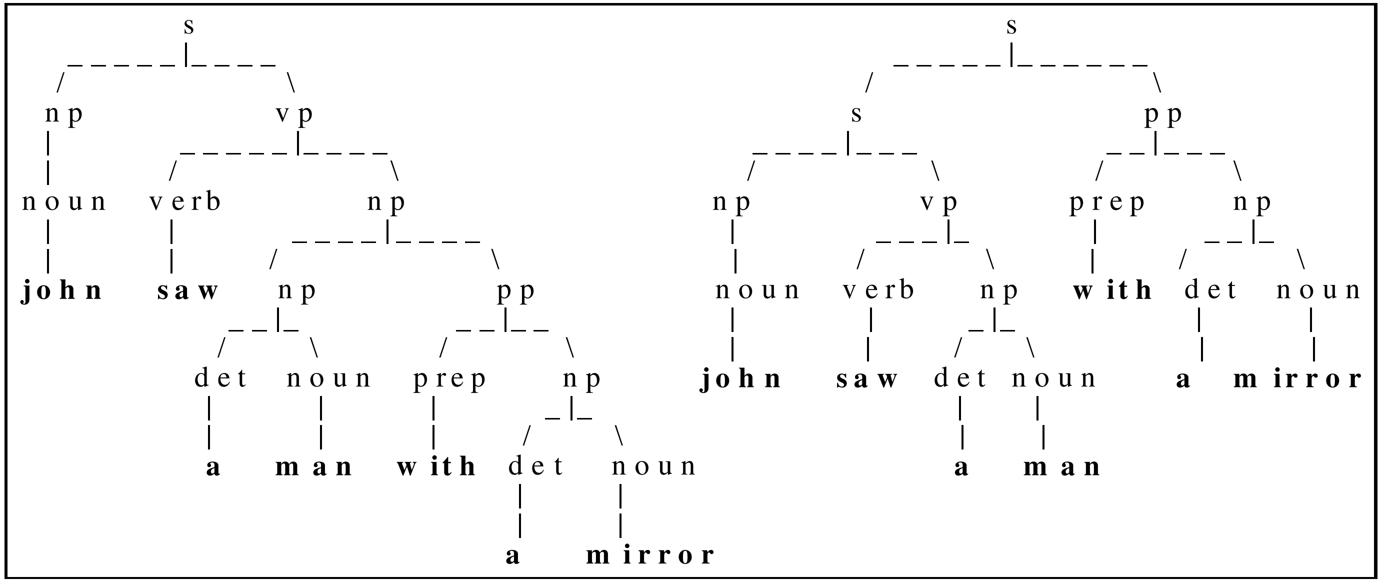
    word(2,3,a).

    …

Using Position Identifier

- Query:
  :-s(0, _,S).

- Result:
  s(0, 7, s(np (noun (john))) vp (verb (…) np(…))).
  s(0, 2, s (…)).



Using Difference Lists

2. Alternative Calculation by Difference Lists

- Idea: Representation of input sentences and position by a list of words and a remainder list

  :- s([john, saw, aman, with, a,mirror],[],S)

- Same system of rules

```
s(X, Y,  s(S1,S2)) :- np(X, Z, S1),   vp(Z, Y, S2).
np(X, Y, np(S)) :- noun(X, Y, S).
vp(X, Y, vp(S1,S2)) :- verb(X, Y, S1),   np(Z, Y, S2).
```

- Only changed alignment of the primary words

```
noun([X | R], R,  noun(X)) :- db_noun(X, C, G, N).
verb( [X | R], R,  verb(X)) :- db_verb( X, P, N).
...
```

- For comparison, position-ids

```
noun(From,To, det(X)) :- word(From, To, X), db_noun(X, C, G, N).
```

Using Difference Lists

- Alternative Calculation by Difference Lists

Advantage

Easier way to input queries

Disadvantages

Violation of "Range Restriction"

⇨ Magic set transformation necessary!

Using DCG Grammar

3. DCG – Syntax

- Instead of:

  s(X, Y, s(S1,S2)) :- np(X, Z, S1),   vp(Z, Y, S2).

- Automatic generation of position attributes

  s(s(S1,S2)) --> np(S1),   vp(S2).

  np(np(S1,S2)) --> det(S1),   noun(S2), { <add. predicates outside DCG> }.

- Alignment either by position attributes

  noun(noun(X)) --> word(X), { db_noun(X, C, G, N) }.

- or by difference lists:

  noun([X | R], R, noun(X)) :- db_noun(X, C, G, N).

Allways: Recursive Rules

Recursive Rules:

- Right recursive
- Left recursive
- Quadrativ recursive!

Recursive Rules

- Example of a Simplified Rule
  (with quadratic recursion)
  - Connections of appositions:
    - ⇨ Defines composite primary words and composite collocations of primary words

```
1.  appP(Sentence,X, Y, appP( N_TREE1, N_TREE2))  :-
        noun(Sentence, X, Z, N_TREE1, absolutus, _, _),
        noun(Sentence, Z, Y, N_TREE2,_ , _, _).
2. appP(Sentence, X, Y, appP( N_TREE, A_TREE))  :-
        noun(Sentence, X, Z, N_TREE, absolutus, _, _),
        appP(Sentence, Z, Y, A_TREE).
3. appP(Sentence, X, Y, appP( A_TREE1, A_TREE2))  :-
        appP(Sentence, X, Z, A_TREE1),
        appP(Sentence, Z, Y, A_TREE2).
4. ...
```

Example of Ambiguities

- Example: The beginning of Ecclesiastes:

Worte	Kohelets	Sohn	David	König	von	Jerusalem
|	|	|	|	|	|	|
noun	noun	noun	noun	noun	preposition	noun

- This can be parsed on several different ways:
  - Words of Kohelet David's Son [and David was] King of Jerusalem.
  - Words of Kohelet Son of David [and each Son of David was] King of Jerusalem.
  - Words of Kohelet David's Son and [Kohelet was] King of Jerusalem.
  - Words of Kohelet David's Son [and word of the] King of Jerusalem.

Example of Ambiguities

The final form listed above includes a quadratic recursion:

```
                         CSV/R
                           |
          _________________|_________________
         /                                   \
       SUB                                 APPV/R
        |                                    |
        |                   _________________|_________________
   %d$A%b$A%r$e            /                                   \
                       APPV/R                                 APPV/R
                         |                                      |
                  _______|_______                        _______|_______
                 /               \                      /               \
               EN                CSV                  SUB               PV
                |                 |                    |                 |
                |              ___|___                 |         ________|________
                |             /       \                |        /                 \
               PN           SUB        EN          %m%a%l%k  PRAEP                EN
                |            |          |                       |                  |
                |            |          |                       |                  |
           %Q%H%L%T      %b%i%n        PN                     %b%.                 ON
                                        |                                          |
                                     %D%W%D                                %Y%R%W$S%L%M
```

Recursive Cycles

- Recursive cycles (in AMOS):