Chapter 3. Foundation of Computational Linguistics (NLP)

Jingbo Xia
College of Informatics, HZAU

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- Lexicon (Part of Speech)
- Syntax (Parsing Tree)
- Semantic

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- Lexicon (Part of Speech)
- Syntax (Parsing Tree)
- Semantic

What is linguistics?

- The study of language in all its manifestations
  - IT company focuses on spoken language
  - Research also depends on written language
- Borders on computer science, psychology, medicine, sociology, law, history, mathematics, philosophy, gender studies, physics, politics…
- Has many fields covering very diverse areas

What is NLP?

Definition 1:

Natural Language Processing (NLP) is a subfield of artificial intelligence and linguistics. It studies the problems inherent in the processing and manipulation of natural language, but not, generally, natural language understanding.

Definition 2:

A study of how to use computers to do things with human languages.

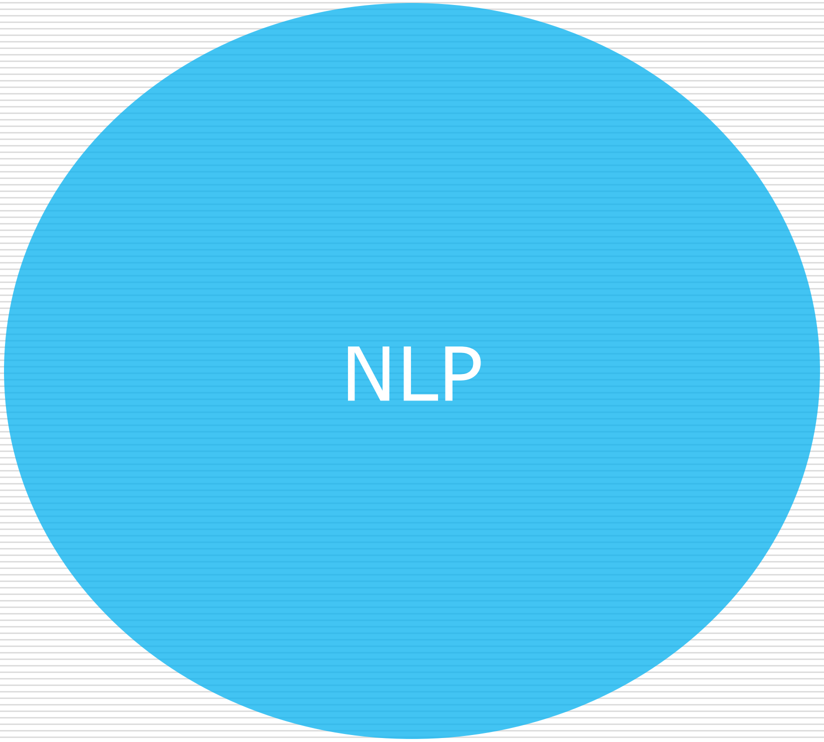
Synonyms: Language Engineering, Human Language Technology

Natural Language Process (NLP)

=

Computational Linguistics

Natural Language Process (NLP)
=
Computational Linguistics (CL)



Motivation 1

- MEDLINE: currently contains over 16 million biomedical abstracts
- 50.000 new abstracts per month

- Huge amount of biomedical knowledge
- Problem: unstructured text
  difficult to analyze automatically

40.000 abstracts á 5 min – app. 400 days (8 h a day)

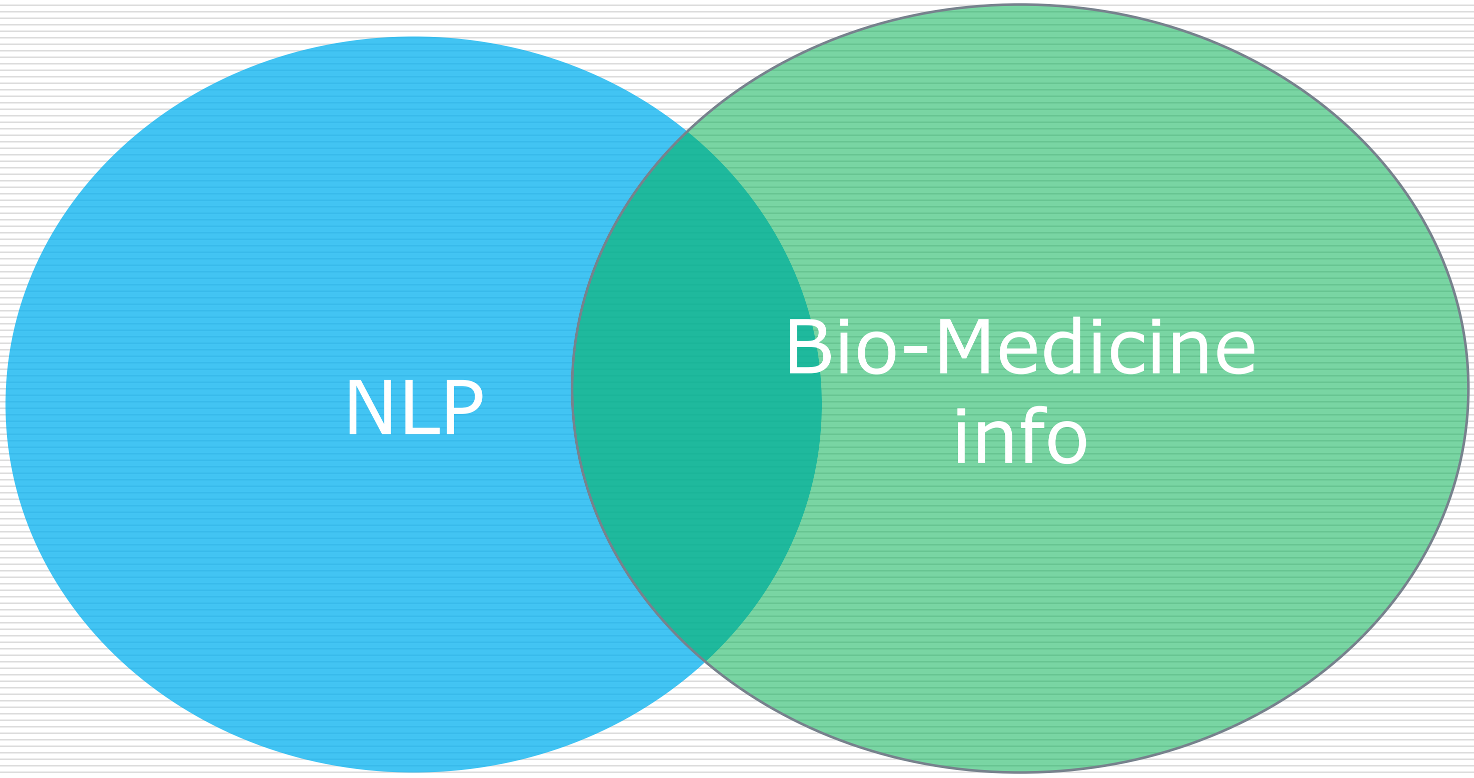
Solution: NLP – Information Extraction

Motivation 2

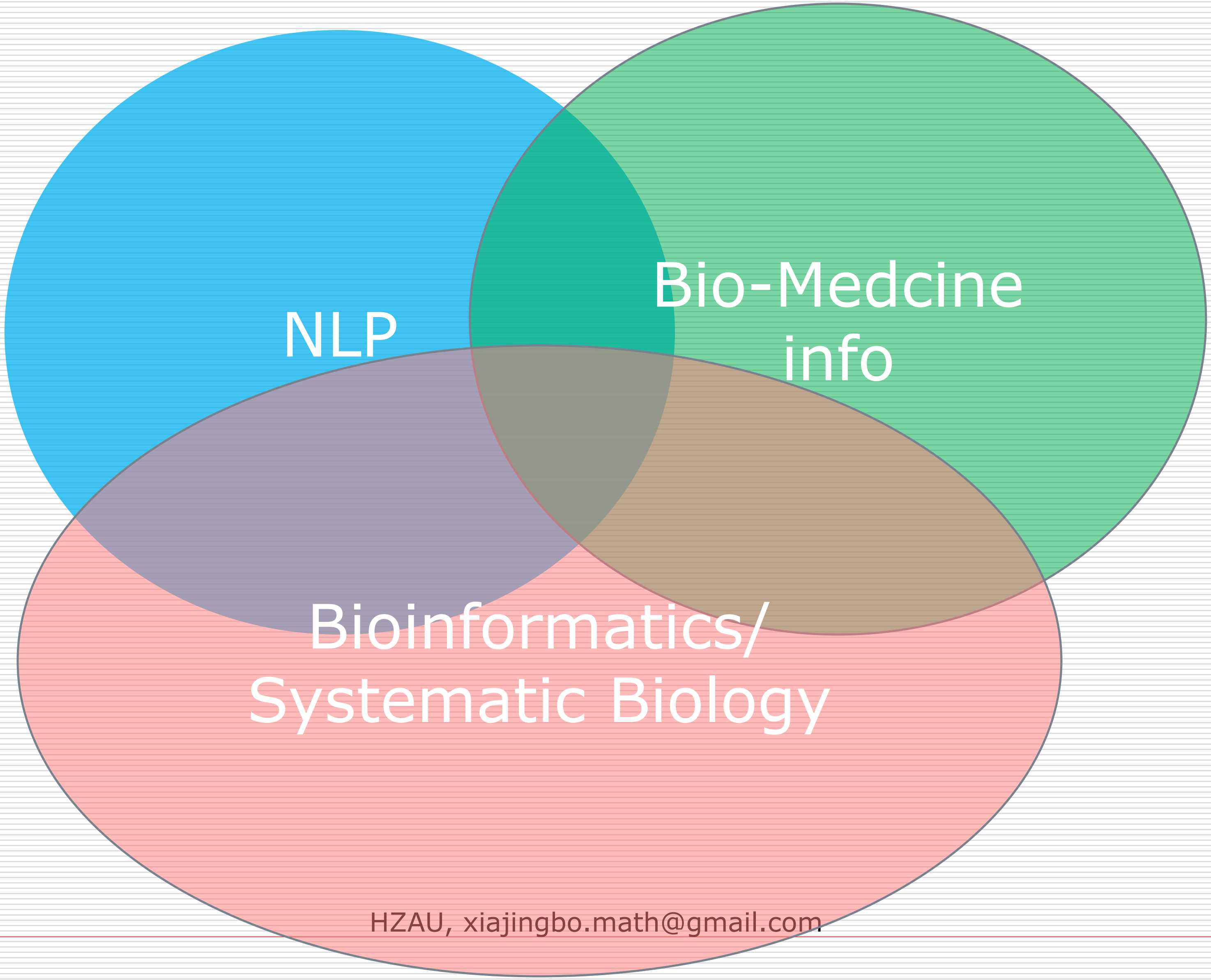
- EHR: Electronic Health Record

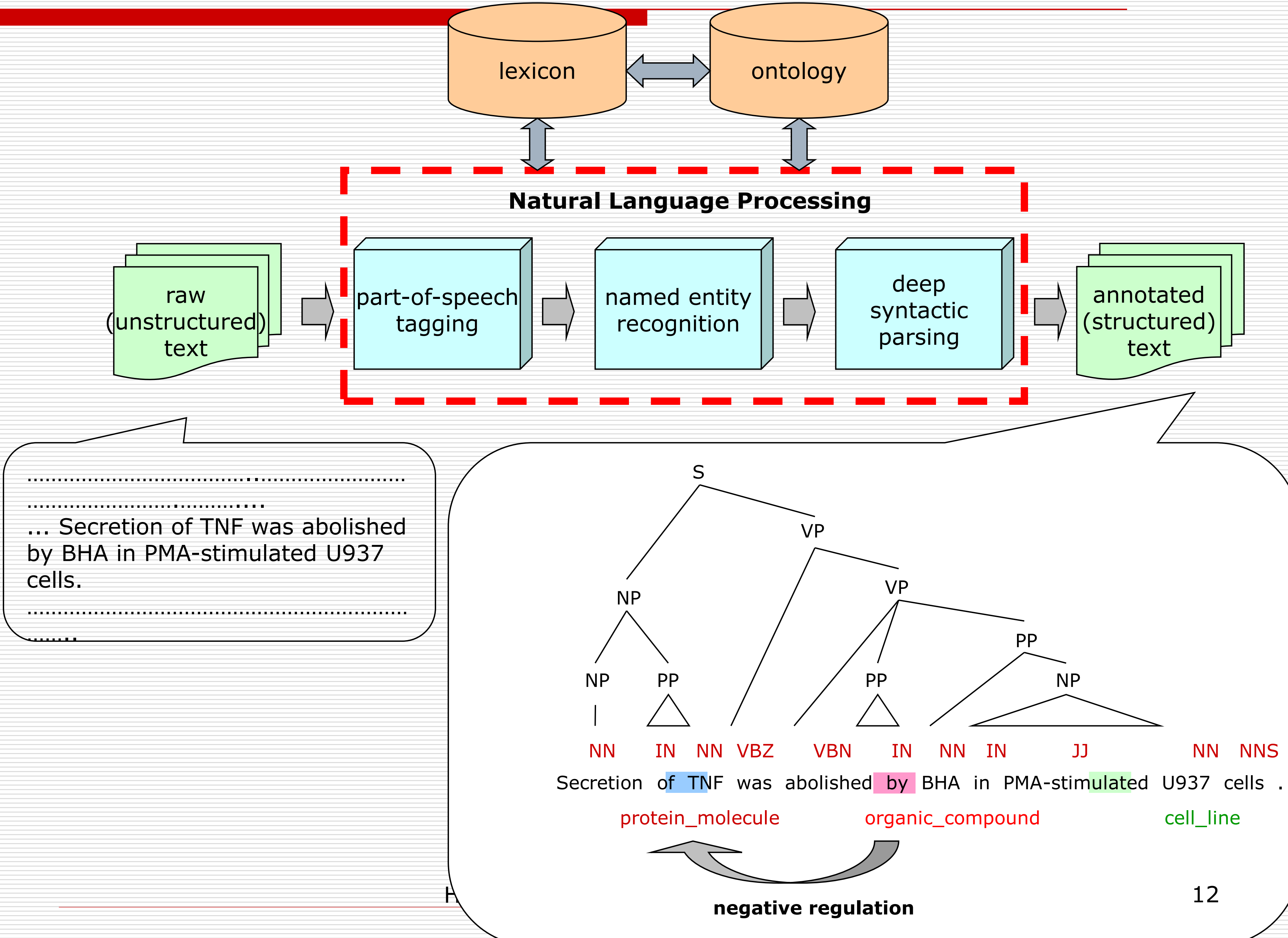
It draw widely attention currently.
Database is increasing.

BioNLP = Bio-medicine info + NLP
= Bio-medicine info + CL



BioNLP in our focus as in HZAU




lexicon
ontology
Natural Language Processing
raw (unstructured) text
part-of-speech tagging
named entity recognition
deep syntactic parsing
annotated (structured) text
... Secretion of TNF was abolished by BHA in PMA-stimulated U937 cells.
S
VP
NP
VP
PP
NP
PP
PP
NP
NN IN NN VBZ VBN IN NN IN JJ NN NNS .
Secretion of TNF was abolished by BHA in PMA-stimulated U937 cells .
protein_molecule
organic_compound
cell_line
negative regulation

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- Lexicon (Part of Speech)
- Syntax (Parsing Tree)
- Semantic

Definitions formulated by some linguists/linguistican

- **Noam Chomsky(1957): "***Language is a set of finite number sentences, each finite in lingth and constructed out of a finite set of elements"*
- **Michael Halliday (2003):** *"A language is a system of meaning- a semiotic system"*

Noam Chomsky

Born	December 7, 1928 (age 87) Philadelphia, Pennsylvania,
Alma mater	•University of Pennsylvania (B.A., 1949; M.A., 1951; Ph.D., 1955) •Harvard Society of Fellows(1951–1955)
Spouse(s)	•Carol Doris Schatz (1949–2008, her death) •Valeria Wasserman (2014–present)
Website	chomsky.info

Institutions	•MIT (1955–present) •Institute for Advanced Study(1958-1959)
Main interests	•Language •Cognitive psychology •Philosophy of mind •Politics · Ethics

Chomsky’s Views

- He abandons the idea that children produce languages only by imitation (abandon behaviorism)
- He rejects the idea that direct teaching and correcting of grammar could account for children’s utterances because the rules children were unconsciously acquiring are buried in the unconscious of the adults.
- He claims that there are *generative rules* (explicit algorithms that characterize the structures of a particular language).

Chomsky's Views

Hypothesis – The inborn linguistic capacity of humans is sensitive to just those rules that occur in human languages. Language development occurs if the environment provides exposure to language. Similar to the capacity to walk.

Universal Grammar – Despite superficial differences all human languages share a fundamental structure. This structure is a universal grammar. We have an innate ability to apply this universal grammar to whatever language we are faced with at birth.

Functionalism vs. Formalism

- Functionalism or functional linguistics refers to the study of the form of language in reference to their social function in communication. It considers the individual as a social being and investigates the way in which she/he acquires language and uses it in order to communicate with others in her or his social environment.
- Representative: M. A. K. Halliday, Systemic functional grammar

Michael Halliday

M. A. K. Halliday	
Born	13 April 1925 (age 90) Leeds, Yorkshire, England
Residence	Australia
Nationality	English
Fields	Linguistics
Known for	Systemic functional linguistics
Influences	Wang Li, J.R. Firth, Benjamin Lee Whorf
Influenced	Ruqaiya Hasan, C.M.I.M. Matthiessen, J.R. Martin, Norman Fairclough
Spouse	Ruqaiya Hasan

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- **Lexicon (Part of Speech)**
- Syntax (Parsing Tree)
- Semantic

Nouns, verbs, adjectives…

One of the challenges for contemporary drug discovery and development is providing regulators, physicians, patients and payers with evidence that differentiates a new drug from the current standard-of-care treatments. This can be particularly challenging in disease areas where combination therapy is common and a wide range of drugs are already available, such as cardiovascular disease, type 2 diabetes, respiratory diseases, some infectious diseases and cancers.

(Nature Review Genetics, 2016)

- How many nouns are there in this text?

Nouns, verbs, adjectives…

One of the challenges for contemporary drug discovery and development is providing regulators, physicians, patients and payers with evidence that differentiates a new drug from the current standard-of-care treatments. This can be particularly challenging in disease areas where combination therapy is common and a wide range of drugs are already available, such as cardiovascular disease, type 2 diabetes, respiratory diseases, some infectious diseases and cancers.

(Nature Review Genetics, 2016)

What defines a Part of Speech?

- Noun
  - *a word (other than a pronoun) used to identify any of a class of people, places, or things (common noun), or to name a particular one of these (proper noun)*

    Semantic definition
  - *any member of a class of words that typically can be combined with determiners to serve as the subject of a verb, can be interpreted as singular or plural, can be replaced with a pronoun, and refer to an entity, quality, state, action, or concept*

    Syntactic and semantic definition

What Parts of Speech are there?

Open word classes	Closed word classes
Nouns (*table*, *time*, *Wiebke*)	Determiners (*the*, *some*, *what*)
Verbs (*go*, *use*, *sleep*)	Auxiliary verbs (*be*, *have*, *must*)
Adjectives (*nice*, *white*, *absent*)	Pronouns (*I*, *ourselves*, *his*)
Adverbs (*quickly*, *clockwise*, *yesterday*)	Prepositions (*on*, *by*, *after*)
Interjections (*wow*, *ouch*, *er*)	Conjunctions (*and*, *while*, *either … or …*)

- More (closed) word classes in English
- More (or less, or different) word classes in other languages
- Different word classes in different linguistic models

Part-of-speech tags

- The Penn Treebank tagset
  - http://www.cis.upenn.edu/~treebank/
  - 45 tags

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
:	:
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBZ	Verb, 3rd person singular present
:	:

JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
:	:
DT	Determiner
CD	Cardinal number
CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
FW	Foreign word
:	:

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun

Number	Tag	Description
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Lexicon: Part-of-Speech Tagging

- Linguistic background
  - What are parts of speech?
  - How do we recognize them?
- Practical usage
  - What are POS taggers good for?
  - What should they do?
- Implementation
  - What are the possible problems?
  - What are the possible solutions?

Why do we need POS tags?

- Main aim: disambiguation
- Useful for most advanced computational linguistic applications
  - Machine translation
  - Named Entity Recognition/Extraction
  - etc.

Part-of-speech tagging (Example 1)

The peri-kappa B site mediates human immunodeficiency
DT NN NN NN VBZ JJ NN
virus type 2 enhancer activation in monocytes …
NN NN CD NN NN IN NNS

- Assign a part-of-speech tag to each token in a sentence.

Part-of-Speech Tagging (Example 2)

- Not surprisingly, an application for determining parts of speech in a text

- Not$_{ADV}$ surprisingly$_{ADV}$, an$_{DET}$ application$_{N}$ for$_{PREP}$ determining$_{V}$ parts$_{N}$ of$_{PREP}$ speech$_{N}$ in$_{PREP}$ a$_{DET}$ text$_{N}$

Part-of-speech tagging is not easy

- Parts-of-speech are often ambiguous

I have to go to school.

verb

I had a go at skiing.

noun

- We need to look at the context
- But how?

Part-of-Speech Tagging – find rules? Example 1.

I have to go to school. I had a go at skiing.

verb noun

- If the previous word is "to", then it's a verb.
- If the previous word is "a", then it's a noun.
- If the next word is …

:

Writing rules manually is impossible

Part-of-Speech Tagging – find rules? Example 2

- Rule-based POS Tagging?
  - Possible rules (simplified):
    - If ends in „est", then it's an adjective (superlative form)
      - *Pest? Rest?*
    - If ends in „ed", it's a verb (past or participle form)
      - *Bed? Sled?*
  - Rules of this kind are few and unreliable
  - Largest problem: they don't help with the ambiguous words!

Part-of-Speech Tagging –From rules to HMM.

Part-of-Speech Tagging – started from rules?

- The wind is blowing.
  - How do we know *wind* is a noun and not a verb?
  - Because it appears after an article and before a verb
    - ART ____ VERB → ART NOUN VERB
- We need rules about inter-word relations
- Though hard to say what the rules are

- *Wind*: 76% noun usage, 24% verb usage
- *ART ____ VERB:* 72% noun, 1% adverb

The wind blows:

- Verb probability: 24% x 0% = 0%
- Adverb probability: 0% x 1% = 0%
- Noun probability: 76% x 72% = 55%

Careful!

The numbers **are invented**, and the calculation is more complex than that.

We need…

- A tokenizer to split the text into tokens
- Tag probabilities for the tokens
  - E.g. *left*: 46% adjective, 31% noun, 23% verb
- Tag sequence probabilities
  - E.g. ADJ ____ NOUN: 57% noun, 43% adjective
  - How long should the sequences be?
- Methods for estimating unknown words
  - E.g. 80% proper noun probability if capitalized

Tag probabilities

The wind blows.

- The: 98% article, 2% adverb
- Wind: 76% noun, 24% verb
- Blows: 53% verb, 47% noun

- Article → Noun: 72%, Article → Verb 1%
- Adverb → Noun 0%, Adverb → Verb 6%
- Noun → Verb 61%, Noun → Noun 4%
- Verb → Verb 3%, Verb → Noun 59%.

Tag probability calculation

The wind blows.

- Article – noun – verb: 98% x 72% x 76% x 61% x 53% = 17%
- Article – noun – noun: 98% x 72% x 76% x 4% x 47% = 10%
- Article – verb – noun: 98% x 1% x 24% x 39% x 47% = 0.04%
- Article – verb – verb: 98% x 1% x 24% x 3% x 53% = 0.0004%
- …

- The complexity of calculations explodes when the length of the sentences and the number of tags increase.

Part-of-speech tagging with Hidden Markov Models

$$P(t_1 \ldots t_n \mid w_1 \ldots w_n) = \frac{P(w_1 \ldots w_n \mid t_1 \ldots t_n) P(t_1 \ldots t_n)}{P(w_1 \ldots w_n)}$$

tags words

$$\propto P(w_1 \ldots w_n \mid t_1 \ldots t_n) P(t_1 \ldots t_n)$$

$$\approx \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

output probability transition probability

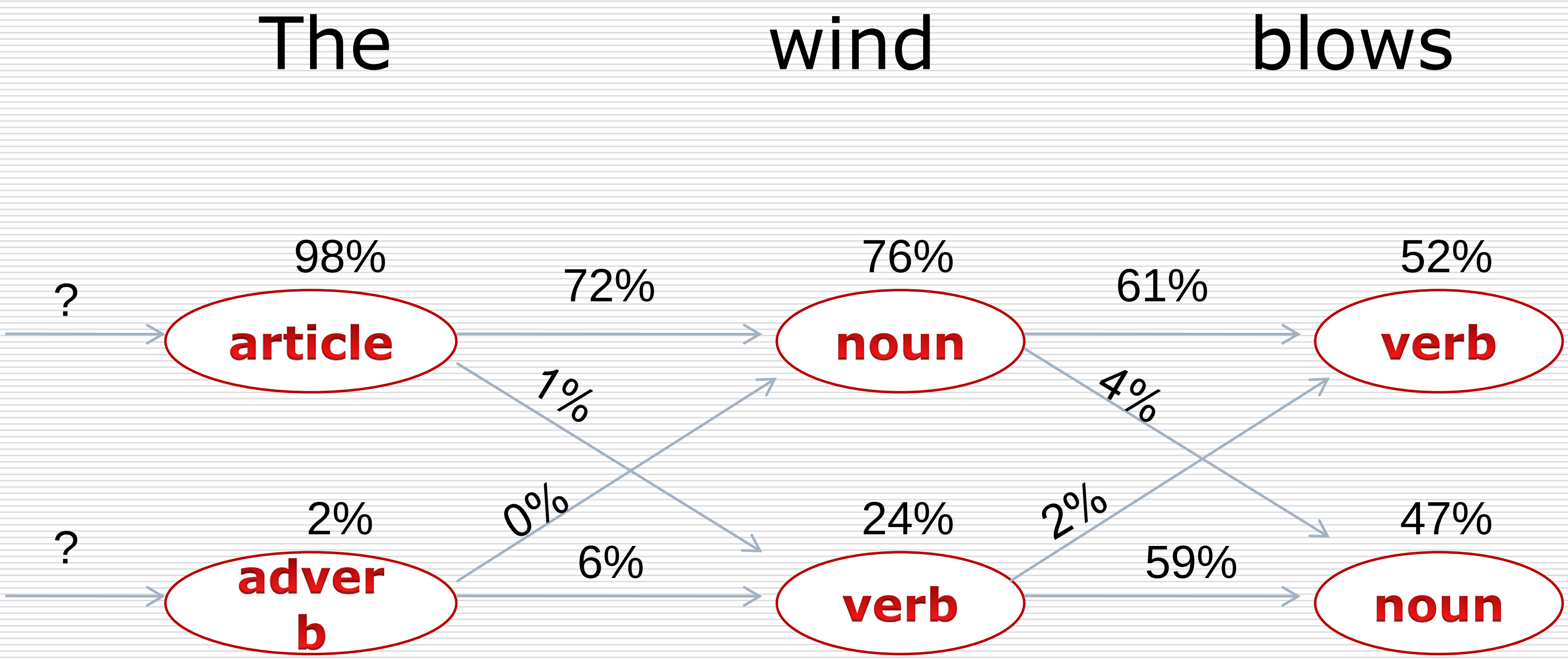
First-order Hidden Markov Models

- Training
  - Estimate
    $$\begin{cases} P(word_j \mid tag_x) \\ P(tag_y \mid tag_z) \end{cases}$$
  - Counting (+ smoothing)

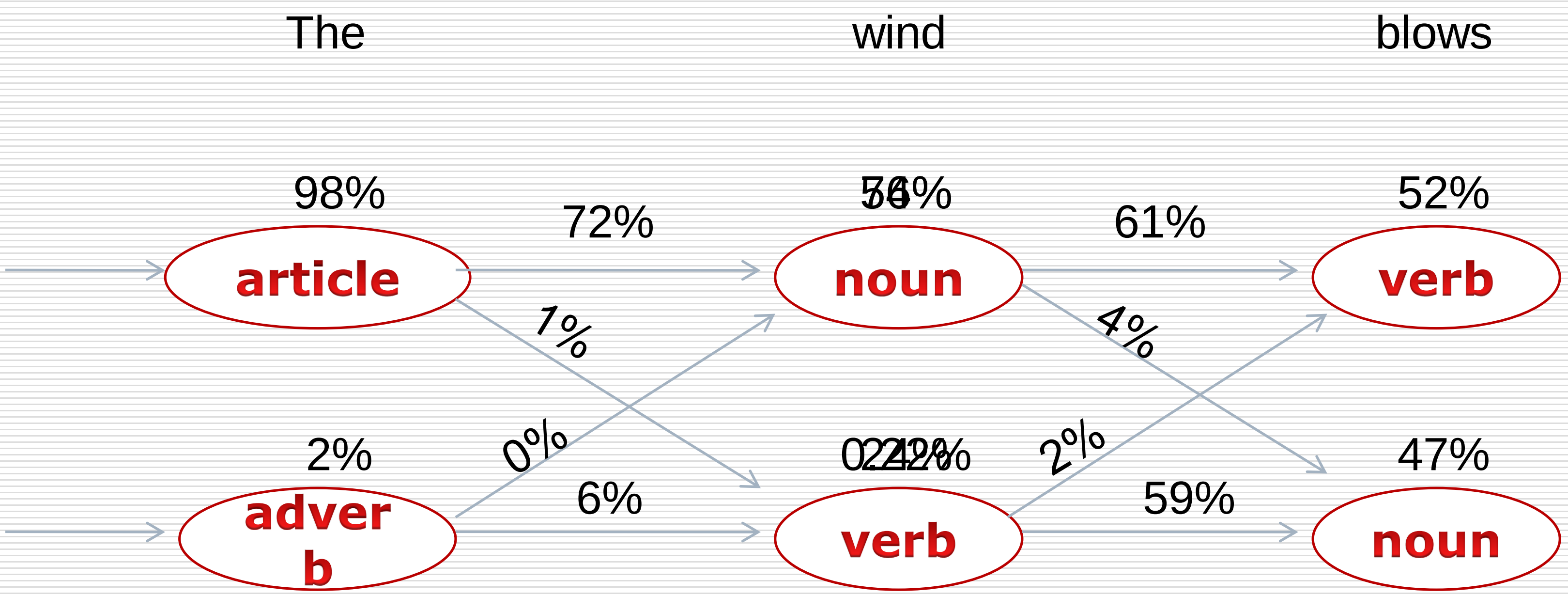
- Using the tagger

$$\arg\max \prod_{i=1}^{n} P(w_i \mid t_i) P(t_i \mid t_{i-1})$$

Hidden Markov Models



Viterbi Algorithm



article: 98%

adverb: 2%

article – noun: 54%

article – verb: 0.22%

adverb – noun: 0%

adverb – verb: 0.02%

article – noun – verb: 17%

article – noun – noun: 1%

article – verb – verb: 0.02%

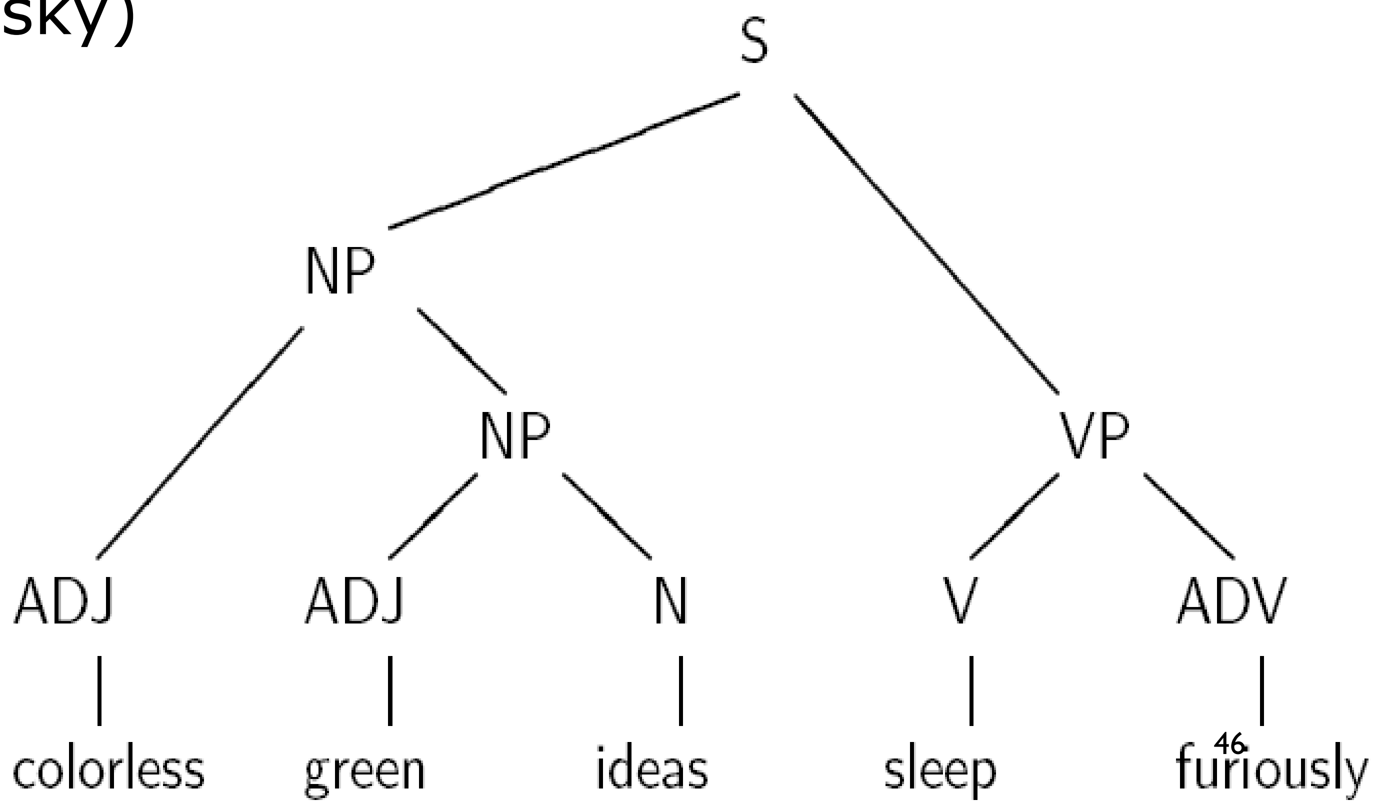
article – verb – noun: 0.05%

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- Lexicon (Part of Speech)
- **Syntax (Parsing Tree)**
- Semantic

Syntax

- Syntax studies the structure of sentences
  - How can we put words together to get sentences?
    - Colourless green ideas sleep furiously. (N. Chomsky)



Syntax

- How do we understand the meaning of a sentence given the meanings of its words?
- What syntactic theory is right?

Syntax

- Syntactic problems:
  - Ambiguity
    - The woman saw the man with the binoculars
    - I made her duck
  - Control
    - I asked her to call Marta.
    - I promised her to call Marta.
  - Coordination
    - John and Alex and Chris and Alice are married.

Outline

- Why Computational Linguistics?
- Two Main Branches of Linguistics
- Lexicon (Part of Speech)
- Syntax (Parsing Tree)
- Semantic

Lambda Calculus (Church and Kleene 1930's)

A unified language to manipulate and reason about functions.

Given $\boldsymbol{f(x) = x^2}$,

$$\lambda x.\ x^2$$

represents the same ***f*** function, except it is *anonymous*.

To represent the function evaluation $\boldsymbol{f(2) = 4}$,
we use the following λ-calculus syntax:

$$(\lambda x.\ x^2\ 2) \Rightarrow 2^2 \Rightarrow 4$$

More on the Lambda Calculus

- **Lambda Calculus Semantic Model**
- Example: *transitive predicate*:
  - **Phrase** — **Lambda Calculus**
  - *likes* — λy.[λx.x likes y]
  - *likes Mary* — [λy.[λx.x likes y]](Mary)
  - λx.x likes Mary
  - *John likes Mary* — [λx.x likes Mary](John)
  - John likes Mary

How to do variable substitution

Official Name: Beta (β)-reduction

Example	**Expression**
likes	[λy.[λx.x likes y]]
likes Mary	[λy.[λx.x likes y]](Mary)

<u>means</u> (basically):

(1) delete the outer layer, i.e. remove [λy. □](Mary) part, and

(2) keep the □ part, and

(3) replace all occurrences of the deleted lambda variable y in □ with Mary

[λy.[λx.x likes y]](Mary)

⇩

[λx.x likes **y**] [λy.](**Mary**)

⇩

[λx.x likes **Mary**]

Note:

nesting order of λy and λx matters

why:

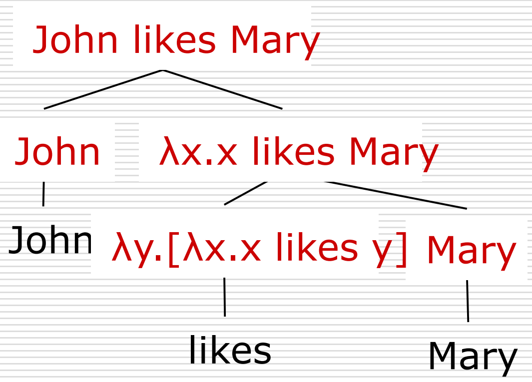
λy.[λx.x likes y]

λx.[λy.x likes y]

here: lambda expression quantifier for the object must be outside because of phrase structure hierarchy

Example:

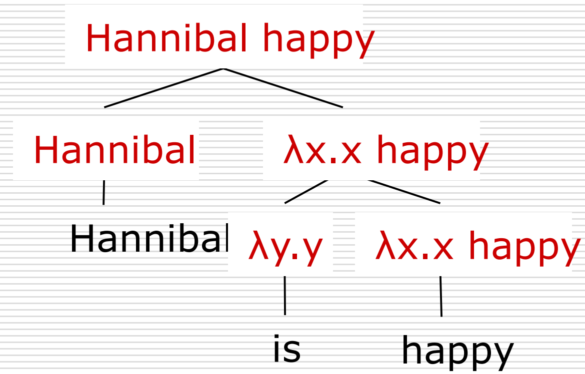
Phrase	**Lambda Calculus**
likes	λy.[λx.x likes y]
likes Mary	[λy.[λx.x likes y]](Mary)
	λx.x likes Mary
John likes Mary	[λx.x likes Mary](John)
	John likes Mary



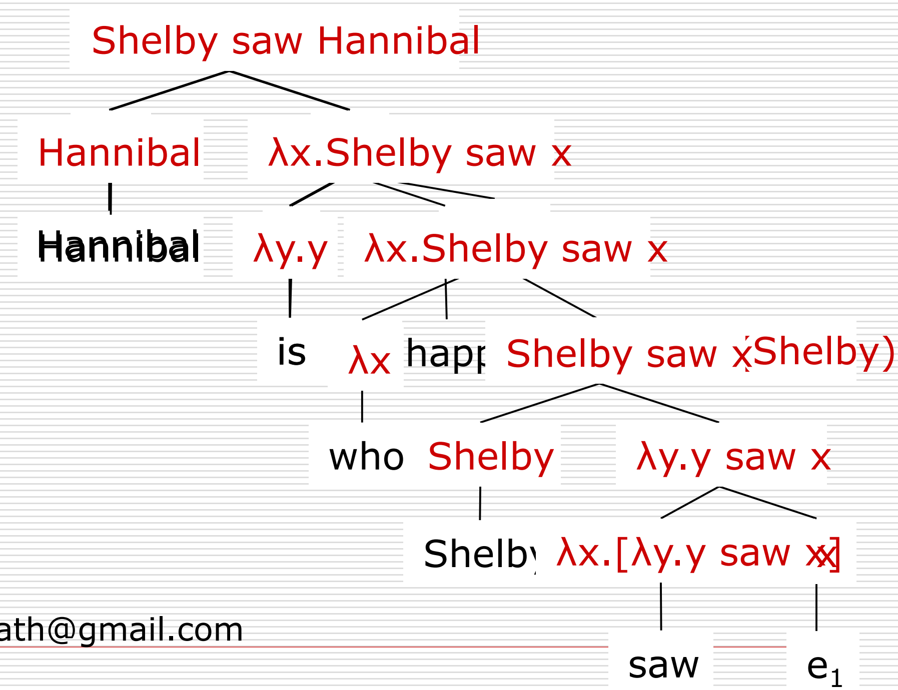
Hannibal is happy

- In the lambda calculus, the semantics of copula *be* is the **identity function**, e.g. λy.y
- Example Derivation:

Phrase	**Lambda Calculus**
is	λy.y
happy	λx.x happy
is happy	[λy.y](λx.x happy)
	λx.x happy



Hannibal is [who Shelby saw]



Reference

- LING 364: Introduction to Formal Semantics