This Lecture

- The NLP pipeline - key components of text understanding
  - Core NLP techniques: tokenization, lemmatization, stemming, chunking, Sentence splitting, part of speech tagging, syntactic parsing
  - Core NLP technologies : named entity recognition, co-reference resolution, event extraction, language modelling

- Understanding the notion of a corpus

Tokenization

- One of the more basic operations that can be applied to a text is **tokenizing**: breaking up a stream of characters into words, punctuation marks, numbers and other discrete items. So for example the character string

“Dr. Watson, Mr. Sherlock Holmes”, said Stamford, introducing us.

- can be tokenized as in the following example, where each token is separated by a space:

“ Dr. Watson , Mr. Sherlock Holmes " , said Stamford , introducing us .

Word structure - Morphology

- Words combine in different orders to form sentences and phrases; they also have internal structure. Nouns in English may have different endings according to whether they are singular (*a box*) or plural (*some boxes*) while in some languages this information may be expressed at the start of the word, for example Swahili *mtoto* (‘child’) vs *watoto* (‘children’).

Word structure - Morphology

- The study of internal structures of words and how they can be modified, parsing complex words into their components e.g. (ni)(na)(kula).

Present	Past
become	became
come	came
mistake	mistook
misunderstand	misunderstood
ring	rang
sell	sold
shake	shook
sing	sang
sink	sank
stand	stood
take	took
tell	told
travel	travelled
understand	understood
withstand	withstood

Some rules for past-tense formation

-come → -came

-take → -took

-ing → -ang

-ink → -ank

-ell → -old

-and → -ood

-el → -elled

Word structure - Morphology

- Improve the rules to account for these words:

Present	Past
bake	baked
command	commanded
bring	brought
sling	slung
smell	smelt
think	thought
wake	woke

Some rules for past-tense formation

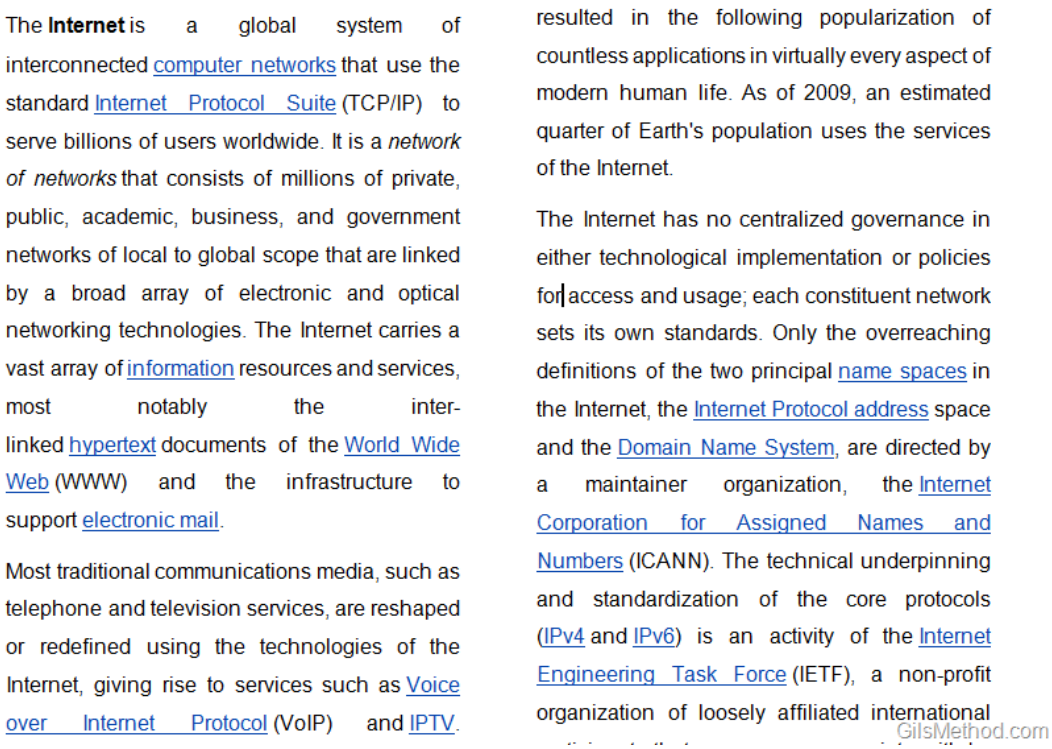
-come → -came
-take → -took
-ing → -ang
-ink → -ank
-ell → -old
-and → -ood
-el → -elled

Lemmatization vs Stemming

- The aim of both processes is the same: to reduce the inflection forms of each word into common base or root
- Stemming – The process of removing affixes from words to derive the basic form is called *stemming*.
  - cutting off the end or beginning of a word taking into account a list of common prefixes and suffixes that can be found in an inflected word. This indiscriminate cutting can be successful in some occasions but not always e.g. studies – studi, studying – study
- Lemmatization – take into consideration the morphological analysis of the words e.g. studies – study, studying – study. The lemma is the base form of all its inflectional forms, while the stem sis not.

Section splitting

- Splitting a text into sections



The **Internet** is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a *network of networks* that consists of millions of private, public, academic, business, and government networks of local to global scope that are linked by a broad array of electronic and optical networking technologies. The Internet carries a vast array of information resources and services, most notably the inter-linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support electronic mail.

Most traditional communications media, such as telephone and television services, are reshaped or redefined using the technologies of the Internet, giving rise to services such as Voice over Internet Protocol (VoIP) and IPTV.

resulted in the following popularization of countless applications in virtually every aspect of modern human life. As of 2009, an estimated quarter of Earth's population uses the services of the Internet.

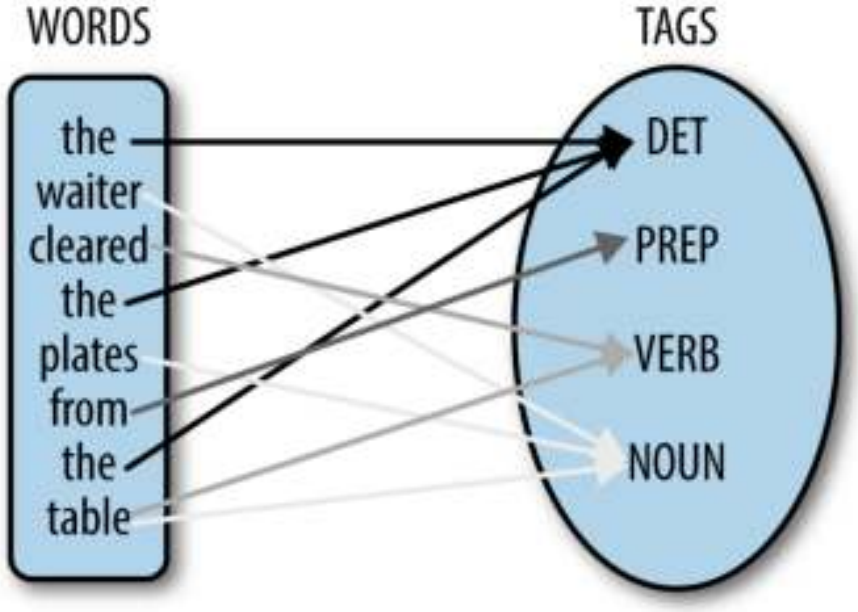
The Internet has no centralized governance in either technological implementation or policies for access and usage; each constituent network sets its own standards. Only the overreaching definitions of the two principal name spaces in the Internet, the Internet Protocol address space and the Domain Name System, are directed by a maintainer organization, the Internet Corporation for Assigned Names and Numbers (ICANN). The technical underpinning and standardization of the core protocols (IPv4 and IPv6) is an activity of the Internet Engineering Task Force (IETF), a non-profit organization of loosely affiliated international

Sentence splitting

- Splitting a text into sentences – taking care of punctuations
- Identifying the sentence boundaries
  - Different punctuations for sentence boundaries e.g. ., !, ?,
  - Identifying abbreviations .e.g Dr. Mr. U.S.A, Mt. Kenya
  - Challenges with long (nested) sentences e.g. Peter said, “Shut up!”, in front of everybody!

Part-of-speech (POS) tagging

- A further stage in analyzing text is to associate every token with a grammatical category or **part of speech** (POS).
- Assigning a syntactic tag to each word in a sentence



POS

- Challenges: Tare many words that can take different parts of speech depending on what they do in a sentence.
  - John caught a *fish*.
  - John likes to *fish* on the river bank.
- Exercise: identify the part of speech in these sentences
  - The cat sat on the mat.
  - John sat on the chair.
  - The train travelled slowly.

Parts of Speech (Penn Treebank 2014)

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
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

Constituent structure

- You will have noticed several recurring patterns in the above examples: *Det Noun*, *Prep Det Noun* and so on.
- You may also have noticed that some types of phrase can occur in similar contexts: *(John | the cat) sat*, a *Proper Noun* or a sequence *Det Noun* can come before a *Verb*.

Regular expression (RE):

(((the | a)(cat | dog))(John | Jack | Susan))(barked | slept)

- This will match any sequence which ends in a verb *barked* or *slept* preceded by **either** a Determiner *a* or *the* followed by a Noun *cat* or *dog* **or** a proper name *John*, *Jack* or *Susan*.

Constituent structure

- Regular expression:

((John|Mary|Fred) | ((the|a)(cat|dog|fish))
(barked | slept | swam)
*((and | or) (barked | slept | swam))**

matches sentences like:

1. John slept
2. The cat barked or swam
3. Mary swam and barked or slept

Constituent structure…

- A common way to represent information about constituent structure is by means of production rules of the form X -> A,B,C…

Sentence → Noun Phrase, Verb Phrase

Noun Phrase → Determiner, Noun (**Example:** the, dog)

Noun Phrase → Proper Noun (**Example:** Jack)

Noun Phrase → Noun Phrase, Conj, Noun Phrase (**Examples:** Jack and Jill, the owl and the pussycat)

Verb Phrase → Verb, Noun Phrase (**Example:** saw the rabbit)

Verb Phrase → Verb, Preposition, Noun Phrase (**Examples:** went up the hill, sat on the mat)

Constituent structure…

- Regular Grammar

$S \rightarrow$ John | Mary | Fred VP

$S \rightarrow$ the | a S1

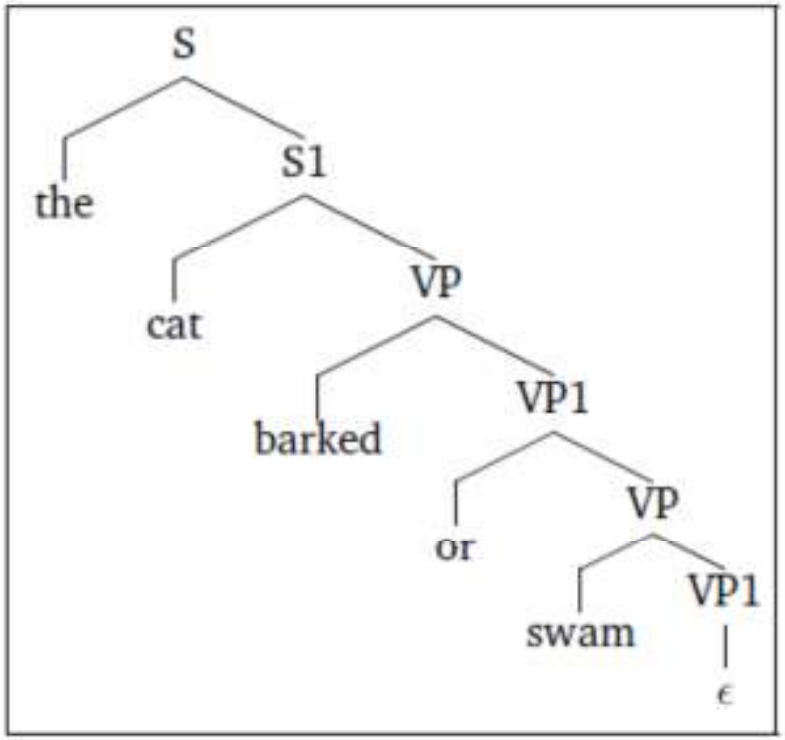
$S1 \rightarrow$ cat | dog | fish VP

$VP \rightarrow$ barked | slept | swam VP1

$VP1 \rightarrow$ and | or VP

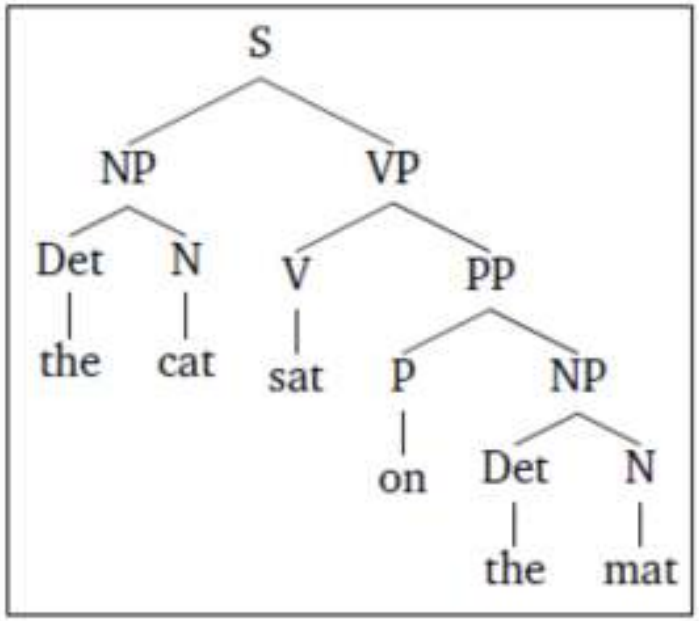
$VP1 \rightarrow \epsilon$

Syntactic structure



Syntactic (Constituency) parsing

- Syntax - The study of the structural relationships between words in a sentence
- Parsing - Building the syntactic tree of a sentence



Syntactic parsing

- Context Free Grammar

$S \rightarrow$ Either S or S

$S \rightarrow$ If S then S

$S \rightarrow NP\ VP$

$NP \rightarrow Det\ N$

$Det \rightarrow a \mid the \mid \epsilon$

$N \rightarrow girl \mid boy \mid dog \mid cat \mid burgers \mid candy \mid cream \mid cake$

$VP \rightarrow V\ NP$

$VP \rightarrow V\ PP$

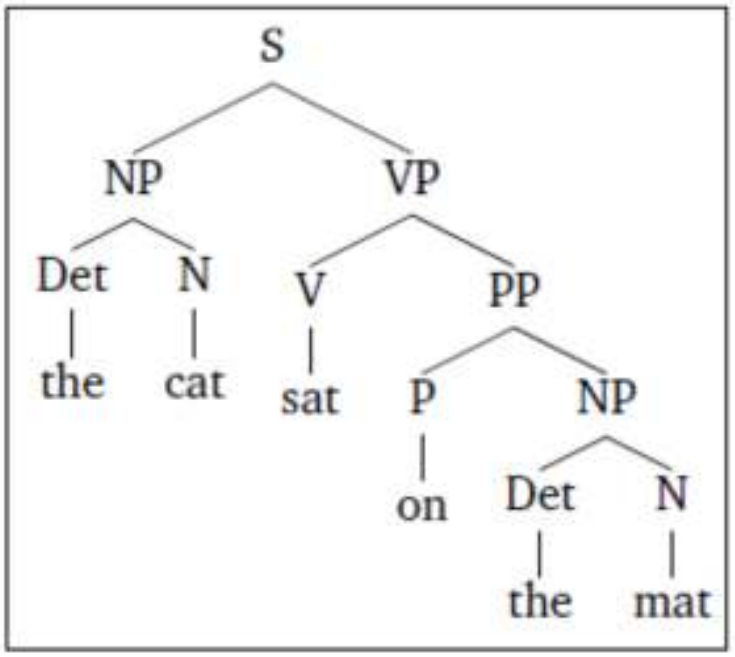
$PP \rightarrow P\ NP$

$V \rightarrow eats \mid likes \mid sat$

$P \rightarrow on$

Syntactic parsing

Syntactic structure



Context Free Grammar

S → Either S or S

S → If S then S

S → NP VP

NP → Det N

Det → a | the | ϵ

N → girl | boy | dog | cat | burgers | candy | cream | cake

VP → V NP

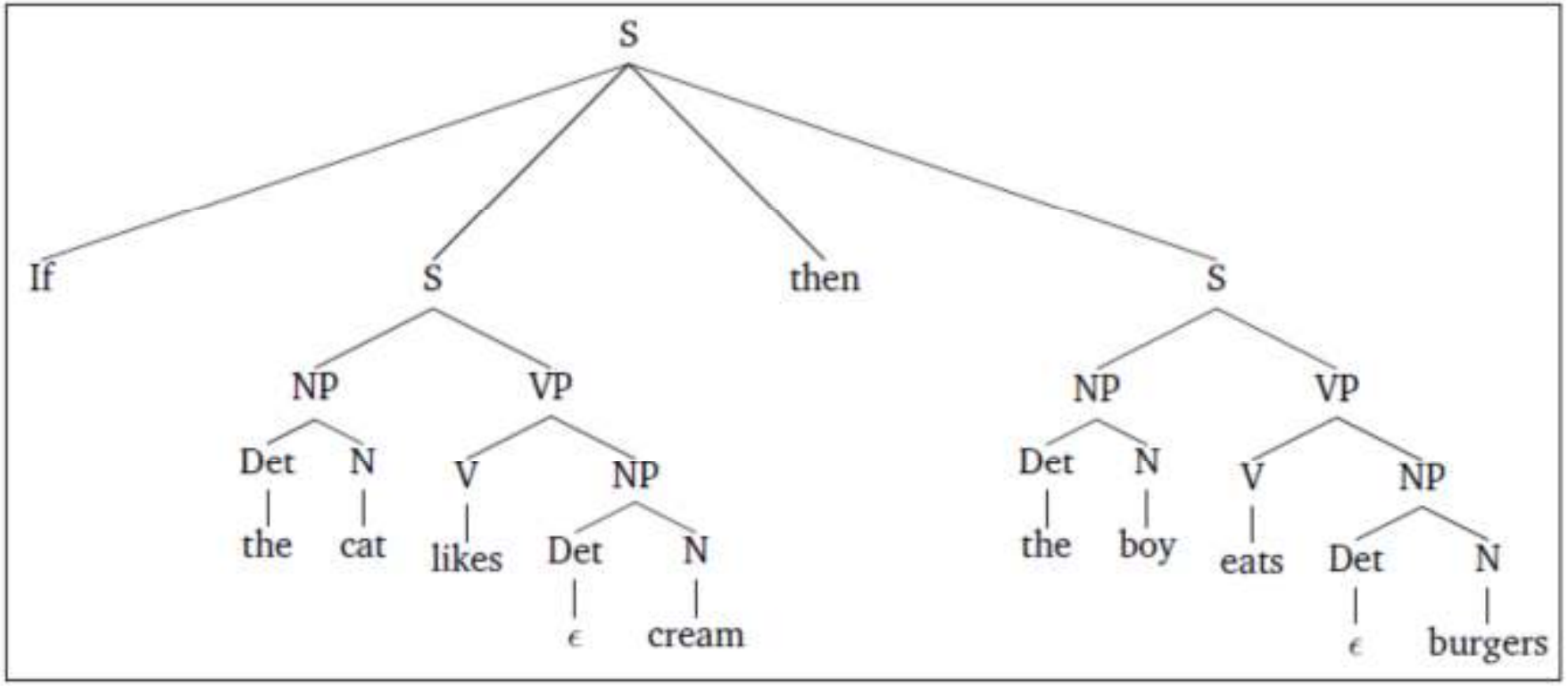
VP → V PP

PP → P NP

V → eats | likes | sat

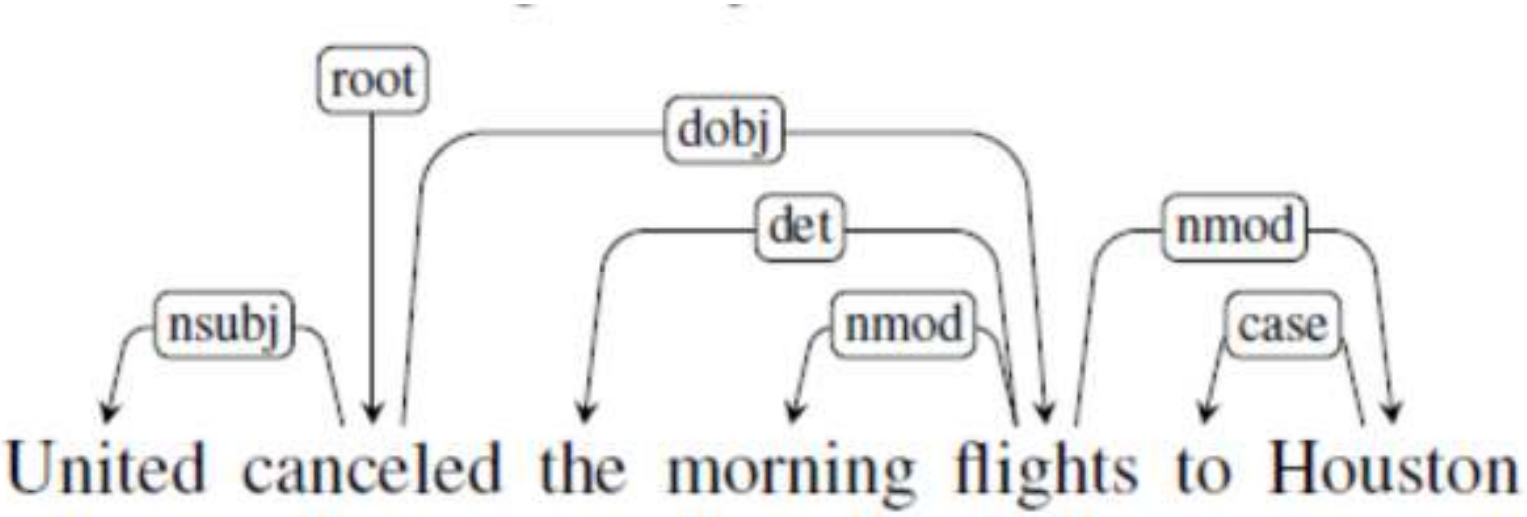
P → on

Syntactic parsing…



Dependency parsing

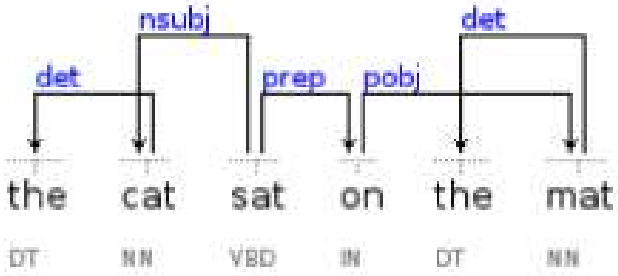
- Focus on grammatical relations – head and depedant



Dependencies

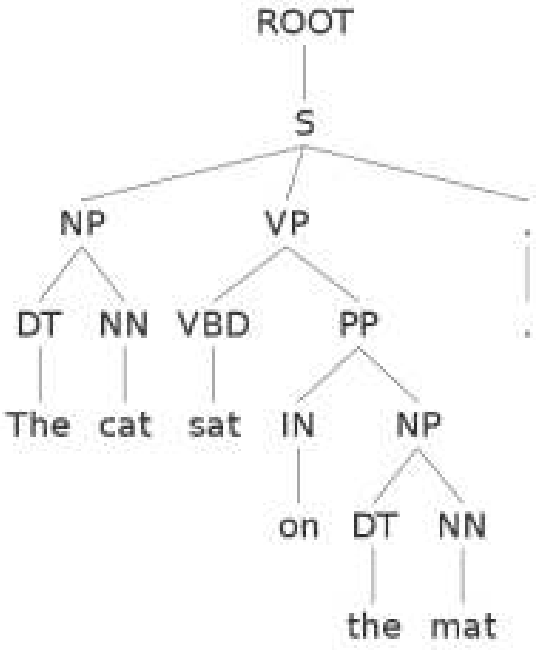
"The cat sat on the mat"

dependency tree

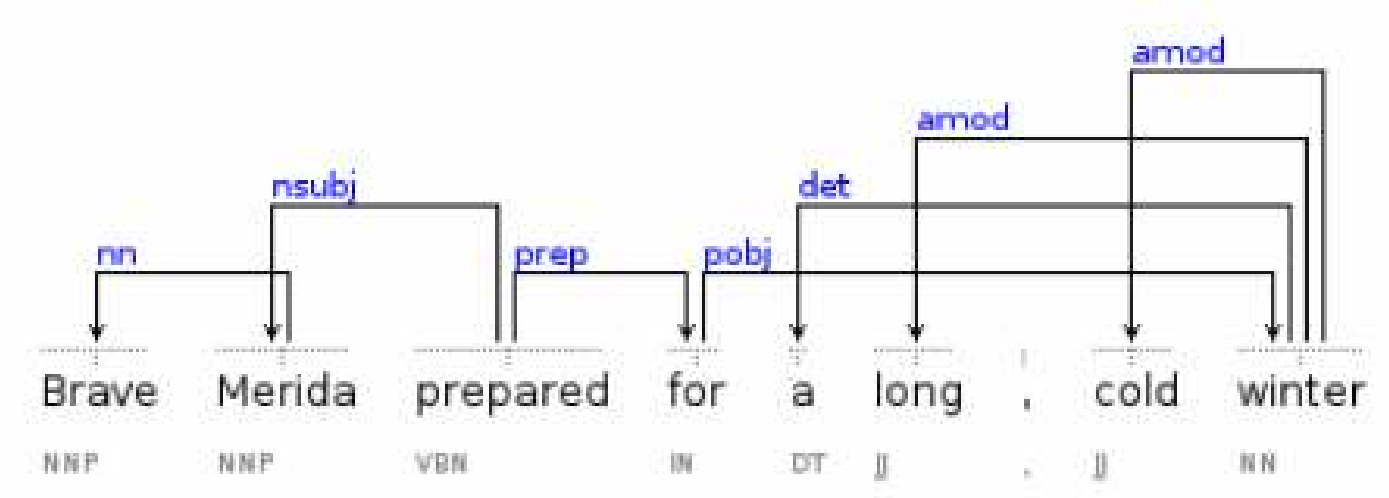


constituency labels of leaf nodes

parse tree



"Brave Merida prepared for a long, cold winter"

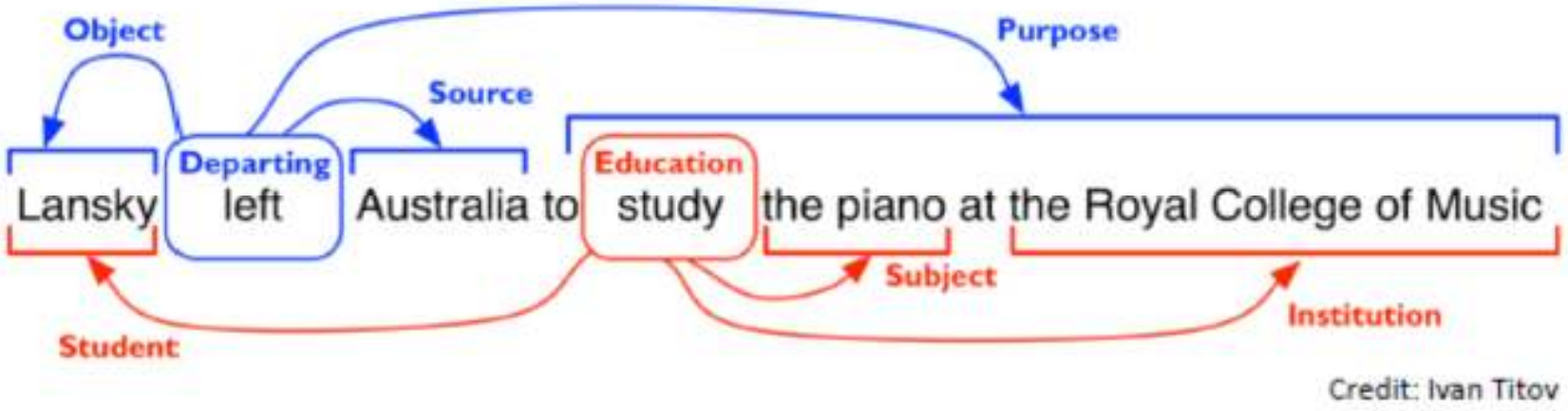


Semantic analysis

- Semantics - The study of the meaning of words, and how these combine to form the meanings of sentences
  - Synonymy: fall & autumn
  - Hypernymy & hyponymy (is a): animal & dog
  - Meronymy (part of): finger & hand
  - Homonymy: fall (verb & season)
  - Antonymy: big & small

Semantic analysis

- Semantic analysis - The process of relating syntactic structures, from the level of phrases of levels, clauses, sentences and paragraphs to the level of the writing as a whole, to the language independent meaning.



Semantic role labeling

- Extracting subject-predicate-object triples from a sentence
- Semantic roles, express the role that arguments of a predicate take in the event
- semantic role labeling - the task of assigning roles to spans in sentences, and **selectional restrictions:** the preferences that predicates express about their arguments, such as the fact that the theme of eat is generally something edible.
- Look at FRAMENET – project for semantic role labeling for English

Semantic Role labeling

Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

John *broke the window.*
AGENT THEME

John *broke the window with a rock.*
AGENT THEME INSTRUMENT

The rock *broke the window.*
INSTRUMENT THEME

The window broke.
THEME

The window was broken by John.
THEME AGENT

Chunking

- Chunking.- the process of extracting phrase from unstructured texts. Instead of just simple tokens which may not represent the actual meaning of the text, it is advisable to use phrases e.g. “South Africa” as a single word instead of “South” and “Africa” as separate words.
- Chunking works on top f POS tagging, it uses POS as input and provides chunks as output e.g NP, VP
- Chunking is important for information extraction , e.g. Named entity extraction- Locations, Person names,

Co-reference resolution

- The task of finding all expressions that refer to the same entity in text.

Christopher Robin is alive and well. **He** is the same person that you read about in the book, **Winnie the Pooh**. As a **boy**, **Chris** lived in a pretty home called **Cotchfield Farm**. When **Chris** was three years old, **his father** wrote a poem about **him**. The poem was printed in a magazine for others to read. **Mr. Robin** then wrote a book

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Named entity recognition

- Identifying pre-defined entity types in a sentence
- A named entity is, roughly speaking, anything that can be referred to with a proper name: a person, a location, an organization.
- The term is commonly extended to include things that aren't entities parse, including dates, times, and other kinds of temporal expressions, and even numerical expressions like prices.

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Named entity recognition…

Challenge: Ambiguity in named entities

[PER Washington] was born into slavery on the farm of James Burroughs.
[ORG Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [LOC Washington] for what may well be his last state visit.
In June, [GPE Washington] passed a primary seatbelt law.
The [VEH Washington] had proved to be a leaky ship, every passage I made...

Word sense disambiguation

- Words are ambiguous: the same word can be used to mean different things.
- word sense disambiguation (WSD), the task of determining disambiguation which sense of a word is being used in a particular context
- A sense (or word sense) is a discrete representation of one aspect of the meaning of a word. Loosely following lexicographic tradition, we represent each sense with a superscript: bank1 and bank2, mouse1 and mouse2.
- In context, it's easy to see the different meanings:
  - mouse1 : .... a mouse controlling a computer system in 1968.
  - mouse2 : .... a quiet animal like a mouse
  - bank1 : ...a bank can hold the investments in a custodial account ...
  - bank2 : ...as agriculture burgeons on the east bank, the river ...
- Check WORDNET and VERBNET, they try to define all possible senses of English words

Relation extraction

- Finding and classifying semantic extraction relations among the text entities.
- These are often binary relations like child-of, employment, part-whole, and geospatial relations.

Relation extraction

- The text tells us, for example, that **Tim Wagner is a spokesman for American Airlines**, that **United is a unit of UAL Corp.**, and that **American is a unit of AMR**.

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Relation extraction…

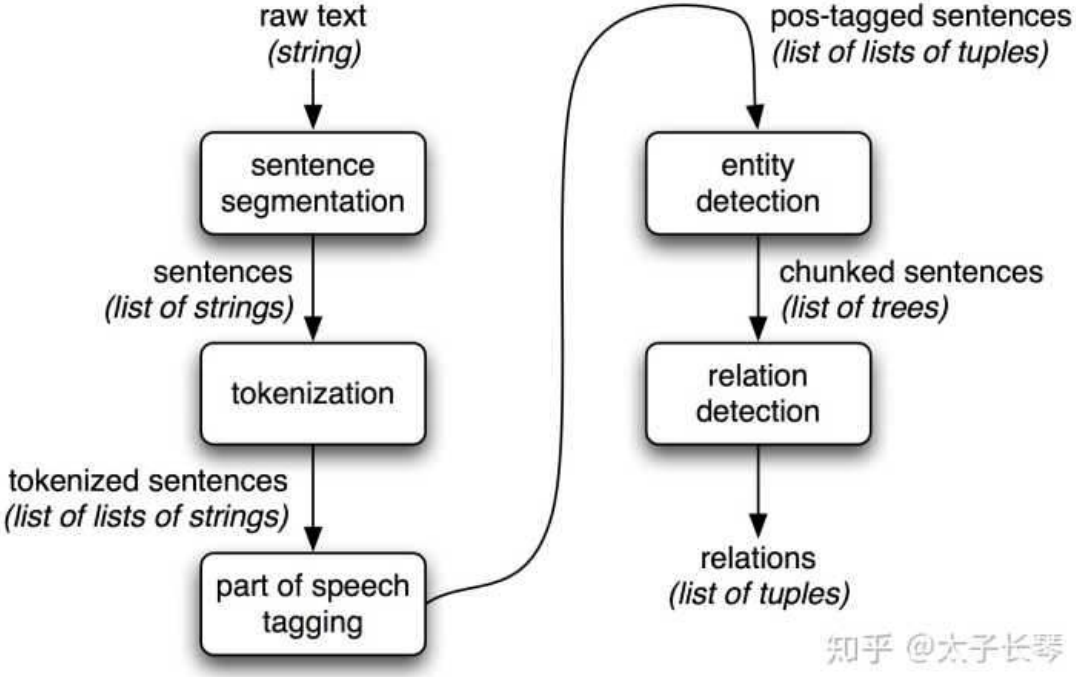
- suppose we are trying to learn the place-of-birth relationship between people and their birth cities.
  - ...Hubble was born in Marshfield...
  - ...Einstein, born (1879), Ulm...
  - ...Hubble's birthplace in Marshfield...
- Seed-based approach

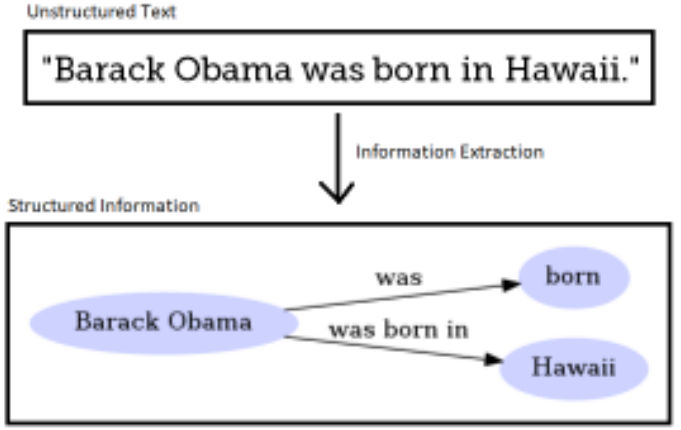
PER was born in LOC
PER, born (XXXX), LOC
PER's birthplace in LOC

<born-in, Edwin Hubble, Marshfield>
<born-in, Albert Einstein, Ulm>
<born-year, Albert Einstein, 1879>

Information extraction

- Turns the unstructured information embedded in texts into structured data, for example for populating a relational database to enable further processing.





language modelling

- Statistical Language Modeling, or Language Modeling and LM for short, is the development of probabilistic models that are able to predict the next word in the sequence given the words that precede it
- Language modeling is the task of assigning a probability to sentences in a language.
- Besides assigning a probability to each sequence of words, the language models also assigns a probability for the likelihood of a given word (or a sequence of words) to follow a sequence of words