

Natural Language Processing

Introduction to NLP

Instructor: Yangqiu Song

Levels of Linguistic Analysis:

Examples of NLP Components

Pragmatics: what does it do?

implemented the
right algorithm



Semantics: what does it mean?

no implementation bugs



Syntax: what is grammatical?

no compiler errors



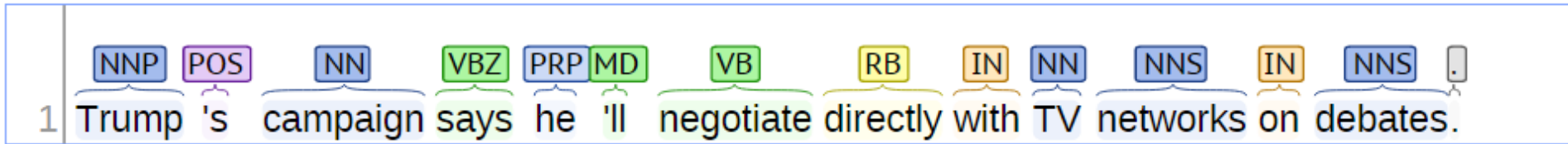
Morphology: basic unit of words

naming your world

Syntax (1)

- Part of speech

Part-of-Speech:



- Tags:
 - NN: common noun
 - NNP: proper noun
 - VB: verb, base form
 - VBZ: verb, 3rd person singular
 - ...

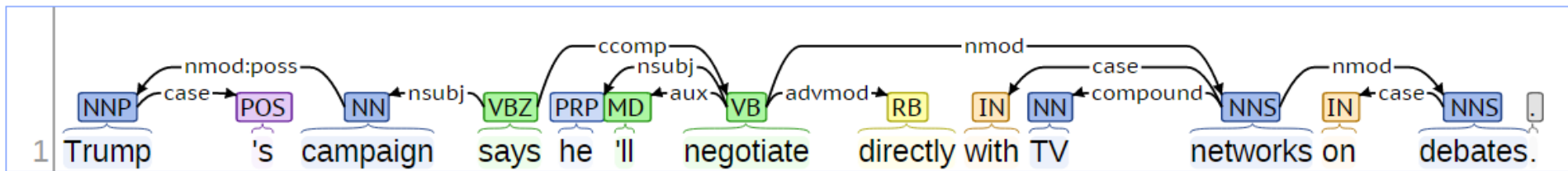
Penn Treebank part-of-speech tags (including punctuation).

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	<i>and, but, or</i>	SYM	Symbol	<i>+, %, &</i>
CD	Cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential ‘there’	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VCN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	<i>\$</i>
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	<i>#</i>
PDT	Predeterminer	<i>all, both</i>	“	Left quote	<i>(‘ or “)</i>
POS	Possessive ending	<i>’s</i>	”	Right quote	<i>(’ or ”)</i>
PRP	Personal pronoun	<i>I, you, he</i>	(Left parenthesis	<i>([, (, {, <)</i>
PRP\$	Possessive pronoun	<i>your, one’s</i>)	Right parenthesis	<i>(],), }, >)</i>
RB	Adverb	<i>quickly, never</i>	,	Comma	<i>,</i>
RBR	Adverb, comparative	<i>faster</i>	.	Sentence-final punc	<i>(. ! ?)</i>
RBS	Adverb, superlative	<i>fastest</i>	:	Mid-sentence punc	<i>(: ; ... – -)</i>
RP	Particle	<i>up, off</i>			

Syntax (2)

- Dependency parse

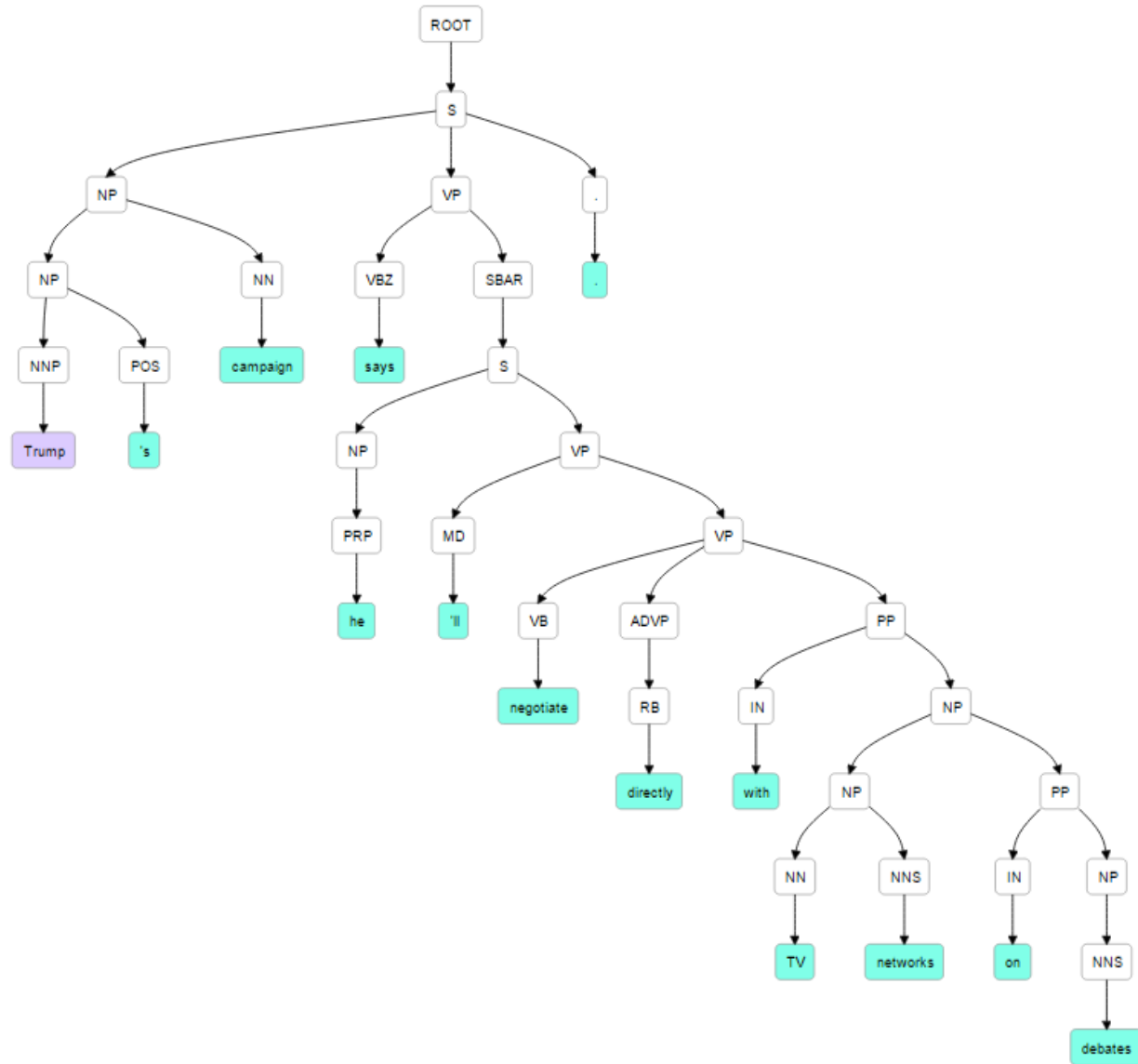
Basic Dependencies:

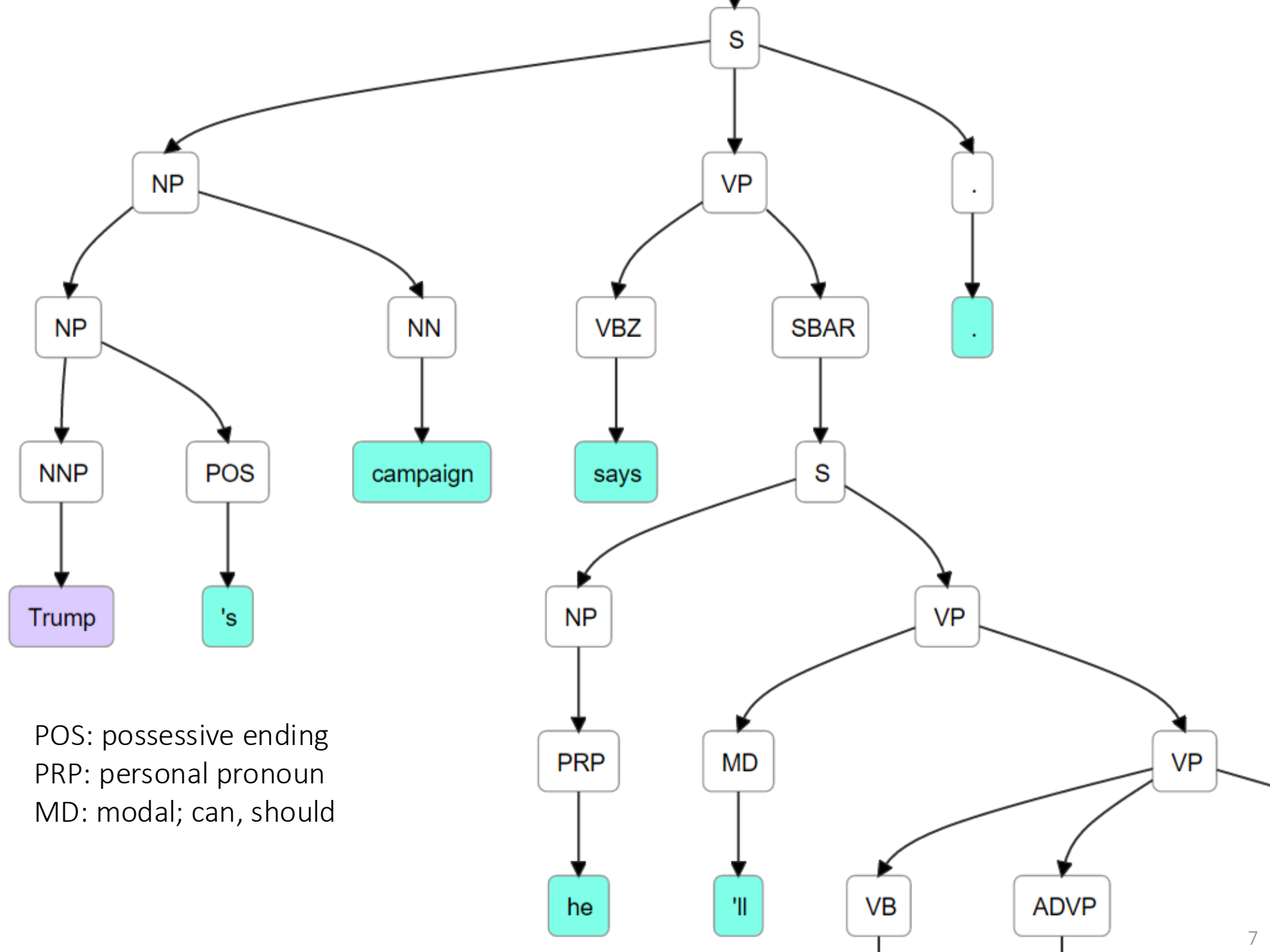


- Dependency relations:
 - nsubj: subject (nominal)
 - advmod: adverbial modifier

Syntax (3)

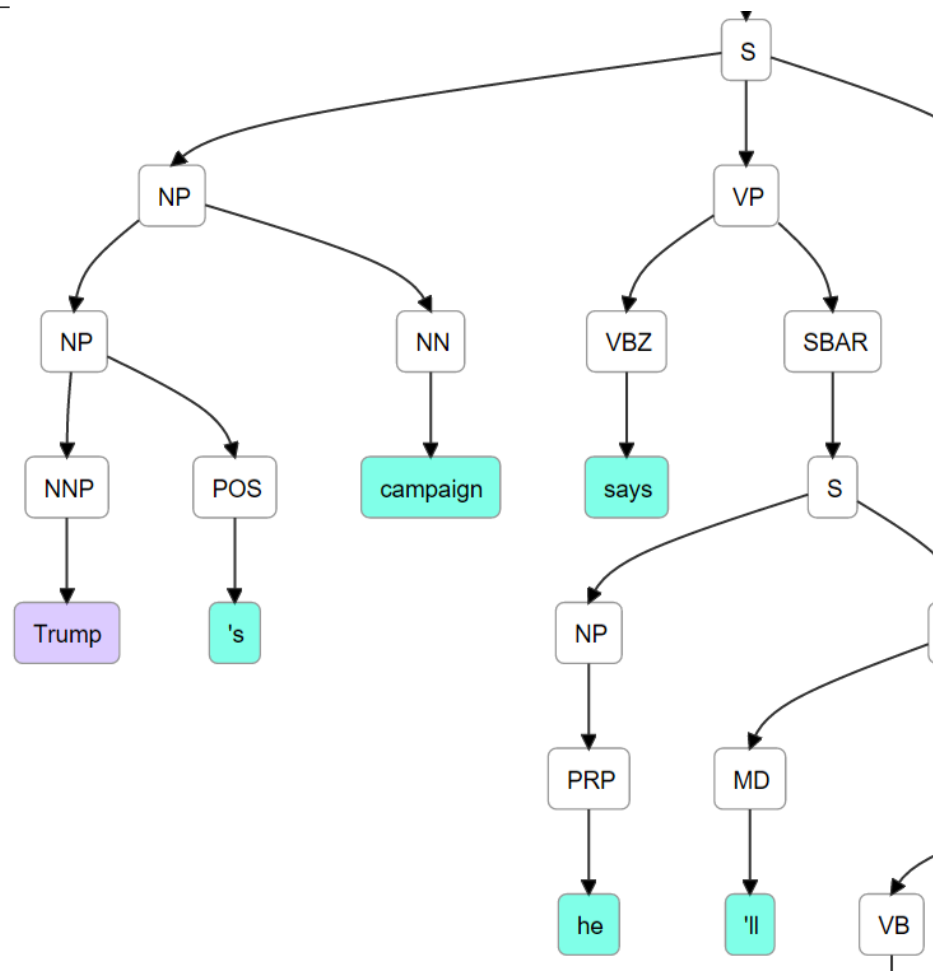
- Constituency parsing





Penn Treebank constituent (or nonterminal) labels.

Label	Description
ADJP	Adjective Phrase
ADVP	Adverb Phrase
CONJP	Conjunction Phrase
FRAG	Fragment
INTJ	Interjection
NAC	Not a constituent
NP	Noun Phrase
NX	Head subphrase of complex noun phrase
PP	Prepositional Phrase
QP	Quantifier Phrase
RRC	Reduced Relative Clause
S	Simple declarative clause (sentence)
SBAR	Clause introduced by complementizer
SBARQ	Question introduced by <i>wh</i> -word
SINV	Inverted declarative sentence
SQ	Inverted yes/no question
UCP	Unlike Co-ordinated Phrase
VP	Verb Phrase
WHADJP	<i>Wh</i> -adjective Phrase
WHADVP	<i>Wh</i> -adverb Phrase
WHNP	<i>Wh</i> -noun Phrase
WHPP	<i>Wh</i> -prepositional Phrase



Semantics (1)

Named Entity Recognition:

- 1 Person Trump's campaign says he'll negotiate directly with TV networks on debates.
- 2 The move by Person Trump, coming just Dur hours after his and other campaigns huddled in a Location Washington suburb to craft a three-page letter of possible demands, thwarts an effort to find consensus after what most candidates agreed was a debate hosted by Org CNBC Date last week.

- Person
- Location
- Org: Organization
- Date/time
- ...

Semantics (2)

- Frame based semantics

Cynthia sold the bike to Bob for \$200

SELLER PREDICATE GOODS BUYER PRICE

- Semantic role labeling (SRL)

	<div><div></div>SRL</div>	<div><div></div>SRL</div>	<div><div></div></div>	<div><div></div>Preposition</div>	<div><div></div>+</div>
The	Logical subject, patient, thing declining [A1]				
stocks					
declined		V: decline.01		Governor	
on	temporal [AM-TMP]		Temporal (on)		
Tuesday			Object		
.					
John		entity turning down [A0]			
declined		V: decline.02			
the		thing turned down [A1]			
cake					

Semantic (3)

- Topics

Trump's campaign says he'll negotiate directly with TV networks on debates. The move by Trump, coming just hours after his and other campaigns huddled in a Washington suburb to craft a three-page letter of possible demands, thwarts an effort to find consensus after what most candidates agreed was a debacle hosted by CNBC last week.

Category 1 politics

Category 2 entertainment

- Categorization/classification
- Clustering
- Topic modeling

Semantics (4) Lexical/Compositional Semantics

- Word

light

- Multi-word expressions: meaning unit beyond a word

light bulb

- Morphology: meaning unit within a word

light lighten lightening relight

- Polysemy: one word has multiple meanings (word senses)

- The light was filtered through a soft glass window. ((physics) electromagnetic radiation that can produce a visual sensation)
- He stepped into the light. (an illuminated area)

- Synonymy/paraphrasing

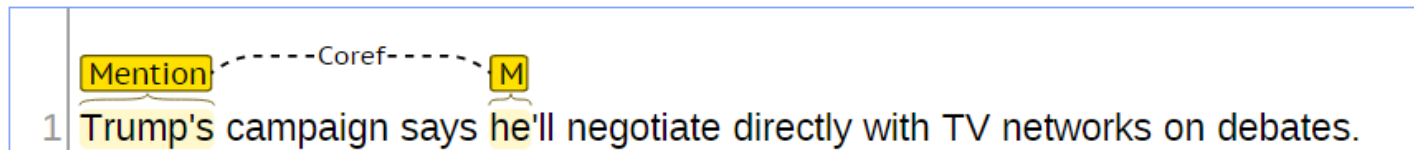
confusing and unclear

- I have fond memories of my childhood.
- I reflect on my childhood with a certain fondness.
- I enjoy thinking back to when I was a kid.

Discourse

- Coreference

Coreference:



- The dog chased the **cat**, which ran up a tree. **It** waited at the **top**.
 - The **dog** chased the cat, which ran up a tree. **It** waited at the **bottom**.
 - **Paul** tried to call George on the phone, but **he** wasn't **successful**.
 - Paul tried to call **George** on the phone, but **he** wasn't **available**.
- Discourse: sentence relation
 - S1: Senator calls this “the first gift of democracy”. **But**
 - S2: The Poles might do better to view this as a Trojan Horse.

Contrast relationship

Pragmatics

- Semantics: what does it mean literally?
- Pragmatics: what is the speaker really conveying?
 - Conversational implicature
 - A: What on earth has happened to the roast beef?
 - B: The dog is looking very happy.
 - Implicature: The dog ate the roast beef.
 - Presupposition: background assumption independent of truth of sentence
 - I have stopped eating meat.
 - Presupposition: I once was eating meat.

Theory of Mind

- ToM
 - i.e., the development of **knowledge** that **others have beliefs, desires, and intentions** that are different from one's own
 - Possessing a functional theory of mind is crucial for success in everyday human social interactions
- The BDI (Belief-Desire-Intent) Model
 - **Beliefs** and **desires** are mediated by **intentions** which in turn controls human's **actions (or speech)** (Kashima et al., 1998)

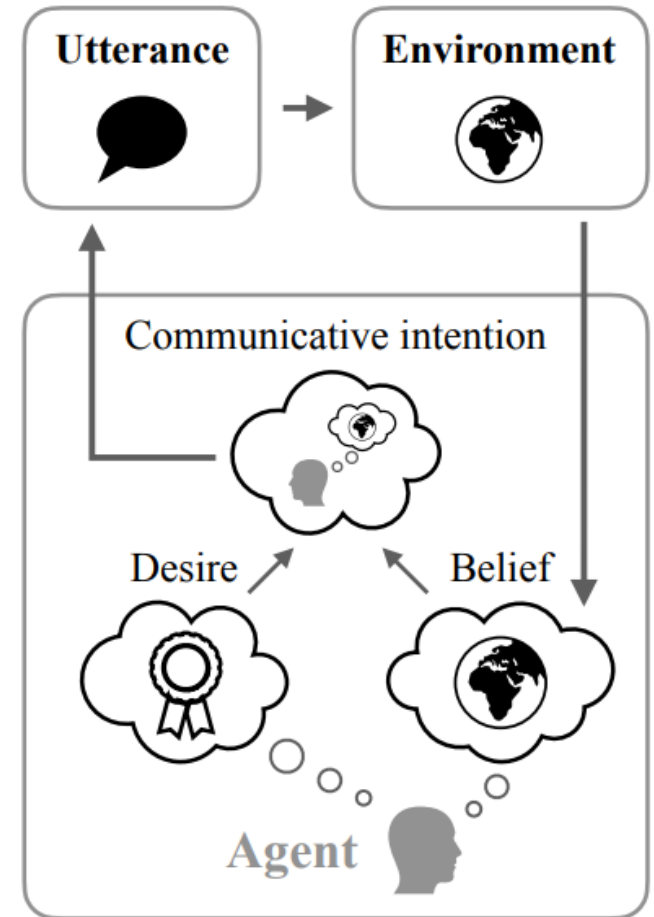


Figure taken from
Andreas (2022)

“Language Models as Agent Models”

- “Current LMs sometimes infer approximate, partial representations of the beliefs, desires and intentions possessed by the agent that produced the context”

Pat watches a demonstration of a bowling ball and a leaf being dropped at the same time in a vacuum chamber. Pat, **who is a physicist**, predicts that the bowling ball and the leaf will **fall at the same rate**.

. . . Pat, **who has never seen this demonstration before**, predicts that the **bowling ball will fall to the ground first**. This is incorrect. In a vacuum chamber, there is no air resistance. Therefore, both the bowling ball and the leaf will fall at the same rate.

The beliefs of Pat has been told in prompts, so the communicative intent of the language model agent has been changed.

To Summarize: The Big Picture of NLP

Arabic text كلب هو مطاردة صبي في الملعب.

How can a computer make **sense** out of this **string**?

- Morphology** - What are the basic units of meaning (words)?
 - What is the meaning of each word?
- Syntax** - How are words related with each other?
- Semantics** - What is the “combined meaning” of words?
- Pragmatics** - What is the “meta-meaning”? (speech act)
- Discourse** - Handling a large chunk of text
- Inference** - Making sense of everything

A Running Example

A dog is chasing a boy on the playground.

Det Noun Aux Verb Det Noun Prep Det Noun

Noun Phrase

Complex Verb

Noun Phrase

Noun Phrase

Prep Phrase

Verb Phrase

Verb Phrase

Sentence

Lexical analysis
(part-of-speech
tagging)

Syntactic analysis
(Parsing)

A person saying this may
be reminding another
person to get the dog back...

Pragmatic analysis
(speech act)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

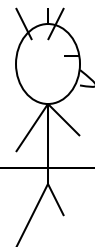
+

Scared(x) if Chasing(_,x,_).

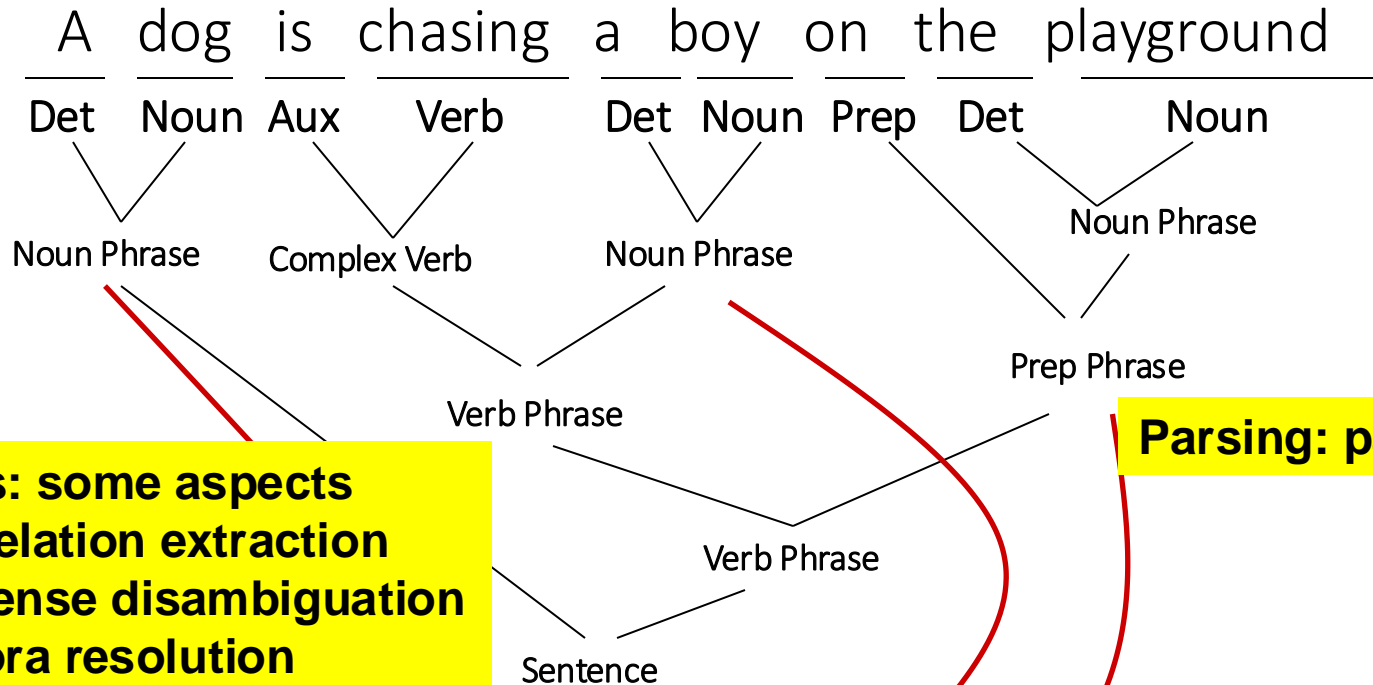


Scared(b1)

Inference



The state of the art



**POS
Tagging:
97%**

Parsing: partial >90%

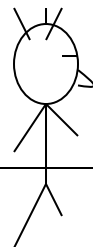
Semantics: some aspects

- Entity/relation extraction
- Word sense disambiguation
- Anaphora resolution

Inference: ???



Speech act analysis: ???





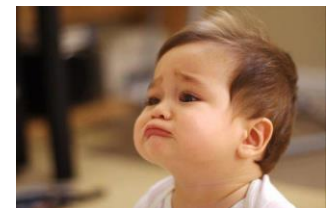
- *Automatically answer our emails*
- *Translate languages accurately*
- *Help us manage, summarize, and aggregate information*
- *Use speech as a UI (when needed)*
- *Talk to us / listen to us*

If we can do this for all the sentences in all languages, then ...

- BAD
- Unfo
- Gen



nt now.



NLP is difficult!!!!!!

- Natural language is designed to make human communication efficient. Therefore,
 - We omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
 - We keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve
- This makes EVERY step in NLP hard
 - Ambiguity is a “killer”!
 - Common sense reasoning is pre-required

More about “Commonsense Knowledge”

- When we communicate,
 - we omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
 - we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve

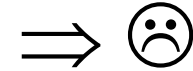
What is Commonsense Knowledge?

Knowledge about the everyday world that is possessed by all people

- A lemon is sour.
- A coat is used for keeping warm.
- To open a door, you must usually first turn the doorknob.
- People want to be respected.
- If you forget someone's birthday, they may be unhappy with you.

Detecting Moods (“affect”) in Text

“My wife left me; she took the kids and the dog”



- Approach:
 - Mood keyword (e.g., “sad”) → mine a “small society of linguistic models of effect” from the KB (=?)
- Applications:
 - Empathy Buddy: (purpose=?)
 - Summarizing a collection of reviews about a topic



Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

- The classic acid test for natural language processing.
- Requires capabilities in both interpretation and generation.
- About \$33 billion spent annually on human translation!

Statistical Solution

- Parallel Texts
 - Instruction Manuals
 - Hong Kong/Macao Legislation
 - United Nations Reports
 - Official Journal of the European Communities
 - Translated news

A Practice for You: Translation Exercise: Learning to translate using parallel text

- Centauri/Arcturan [Knight, 1997]
 - Your assignment, translate this to Arcturan:
 - farok crrrok hihok yorok klok kantok ok-yurp

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anak plok sprok .	8a. lalok brok anak plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok .	10a. lalok mok nok yorok ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

It's Really Spanish/English

Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Statistical Learning to MT?

- Traditional SMT did a lot of alignment like this
 - It's actually a learning+inference process
 - Learning: learning the parameters to fit the data
 - Inference: make decisions online for the alignment with large likelihood and constraints (rules defined by human)
 - Deep learning
 - Much simplified the process
 - Sequence to sequence model, which we will see
 - Can we borrow the ideas for deep learning?
 - Still an active research area

A brief history of NLP

- Early enthusiasm (1950's): Machine Translation
 - Too ambitious
 - Bar-Hillel report (1960) concluded that fully-automatic high-quality translation could not be accomplished without knowledge (Dictionary + Encyclopedia)
- Less ambitious applications (late 1960's & early 1970's): Limited success, failed to scale up
 - Speech recognition
 - Dialogue (Eliza) **Shallow understanding**
 - Inference and domain knowledge (SHRDLU="block world")

Deep understanding in limited domain
- Real world evaluation (late 1970's – 2020)
 - Story understanding (late 1970's & early 1980's) **Knowledge representation**
 - Large scale evaluation of speech recognition, text retrieval, information extraction (1980 – now) **Robust component techniques**
 - Statistical approaches enjoy more success (first in speech recognition & retrieval, later others) **Statistical language models**
- Current trend:
 - One large language model for all tasks
 - Boundary between statistical and symbolic approaches is disappearing.
 - We need to use all the available knowledge **A lot of applications nowadays**
 - Application-driven NLP research (bioinformatics, Web, Question answering...)

Goal of this Lecture

- Get familiar with some of the NLP terminologies and tasks
 - We won't be able to touch all of the tasks
 - But we will use some of the terminologies (will also re-introduce when needed)
- Modern NLP solutions are based on large scale data
 - Then the problem becomes machine learning
 - how to (over-)fit the data
 - A detour: there is still a lot of problem
- We will focus on machine learning algorithms that can handle text data and work on more practical tasks
 - Not just syntax/semantics, but more pragmatics