香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

CSC6052/DDA6307/MDS6002: Large Language Model

Lecture 2: Basics of Linguistics

Spring 2024
Benyou Wang
School of Data Science

Recap

A bad news

- https://chatgpt.cuhk.edu.cn/ is broken and it is also ran of out money
- We are fixing this…

What is in the previous lecture?

- **Everything about the course itself**

  Final project is all you need

  (It does not make sense to join this course if you do not want to make a nice project)

  Workload is heavy, go dropping it if you do not have enough time.

  It is waste of your time/energy even you could pass this course with minimum efforts

- **The basic review of the course**

  A story of destructure (去结构化) and scaling(模型变大), and finally application.

- **What is NLP and its applications**

  translation, QA, **embodied AI**, etc.

- **What is ChatGPT and some case study**

  Some is too much fancy and also **some limitation** existed

- **Future NLP**

  Defined by you guys (young people do this better than senior guys)

Destructure (去结构化) and scaling (模型变大),

- Structure prior harms scaling up
  - When you were a kid, some constrains (structure prior in NLP) do help you (safely grow up)
  - When you become an adult, these constrains becomes a burden

A famous joke

Every time I fire a linguist, the performance of the speech recognizer goes up

Frederick Jelinek, a renowned Czech-American researcher in natural language processing and speech recognition, famously said in 1985

He was not a pioneer of speech recognition, he was the pioneer of speech recognition.

—Steve Young (2010)

https://www.nature.com/articles/s42256-023-00703-8

A natural question arises

What is the **structure** of language?

A kind advice

So If you learn this lecture well and **only** learn this lecture, you might survive in a research institute, not in a company in coming years.

Today’s lecture

- **What is linguistics?**
- **Linguistic structure**
  - Character
  - Word
  - Sentence
  - Discourse (篇章)
- **More about desturcture and scaling**
  - Inductive bias
  - Inductive bias in NLP during many decades
  - Rethinking Empiricism vs. Rationalism
- **(Next) From linguistics to computing linguistics**

What is linguistics?

What is linguistics?

Linguistics is the scientific study of language. Linguistics is based on a theoretical as well as a descriptive study of language and is also interlinked with the applied fields of language studies and language learning, which entails the study of specific languages.

Today's course will help us understand two significant aspects in linguistics:

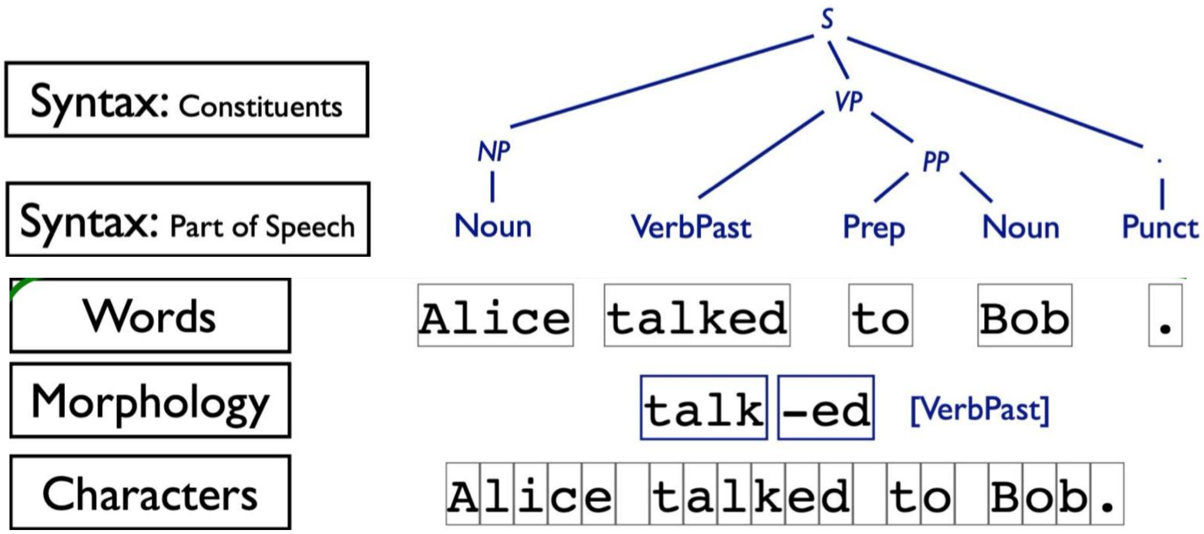
- **How do we understand language?**
- **How can computers process language?**

https://en.wikipedia.org/wiki/Linguistics

How do we understand language?

An important insight

Language is structured



- **S** - Sentence
- **NP** - Noun Phrase
- **VP** - Verb Phrase
- **Det** - Determiner
- **N** - Noun
- **V** - Verb
- **PP** - Prepositional Phrase
- **P** - Preposition
- **A** - Adjective
- **Adv** - Adverb
- **Deg** - Adverb of Degree
- **Pro** - Pronoun
- **Aux** - Auxiliary Verb
- **M** - Modal

There’s structure underlying language

Isabel broke the window
The window was broken by Isabel

The cat is batting the toy
The toy is being batted by the cat

The plid yorbed the plof
The plof was yorbed by the plid

- We have some knowledge of structure that’s separate from the words we use and the things we say.
- **Building block vs. Building strategies**

A example that does not respect linguistic structure

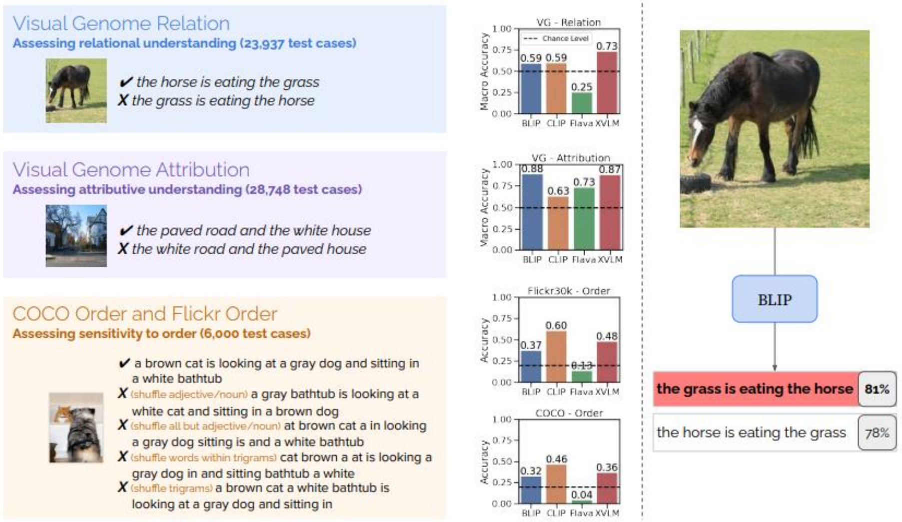
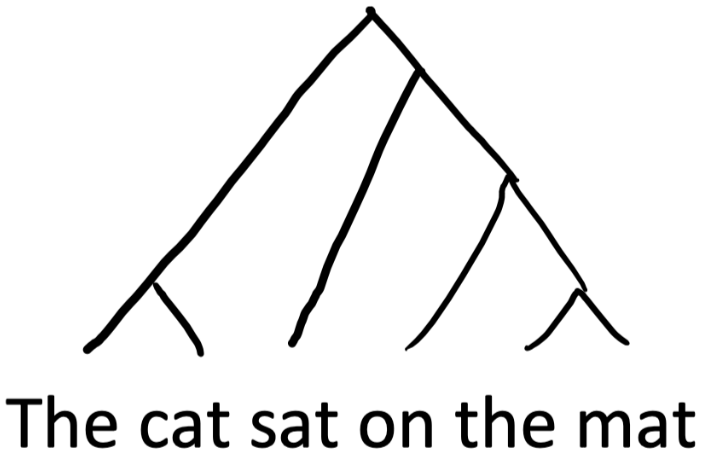


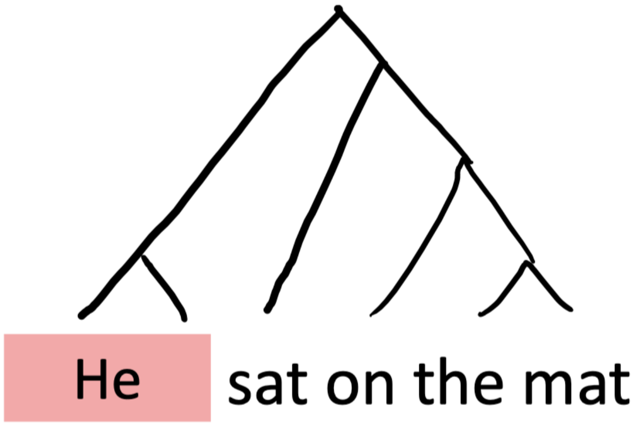
Figure 1: **ARO (Attribution, Relation and Order) a benchmark to test composition and order understanding.** We present four large-scale tasks to test the model's relational, attributive, and order understanding. These datasets probe the models' ability to pick the correct ordering of the constituents of a caption, e.g. by asking the model to pick between 'the horse is eating the grass' vs 'the grass is eating the horse'. Existing VLMs exhibit intriguing deficiencies at these simple tasks: several models remain at or below chance level. For example, BLIP chooses 'the grass is eating the horse', with 81% probability.

Mert Yuksekgonul. et.al. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It? ICLR 2023.

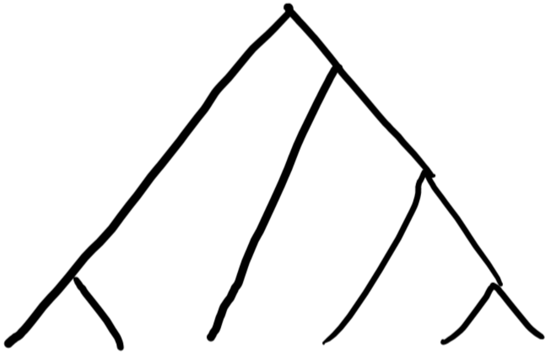
Structure dictates how we can use language



Structure dictates how we can use language



Structure dictates how we can use language

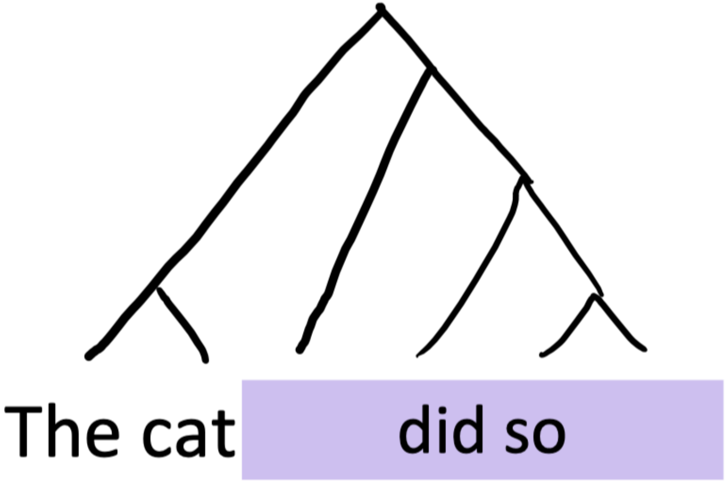


The cat sat on it

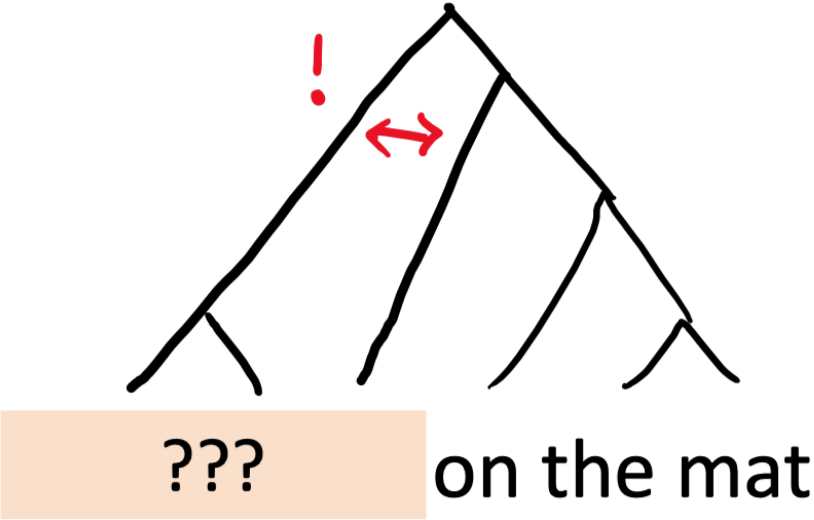
Structure dictates how we can use language

The cat sat there

Structure dictates how we can use language

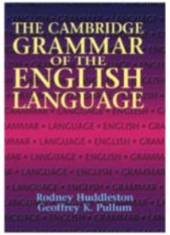


Structure dictates how we can use language



Recall the rules you learned in school

- A community of speakers (eg, Standard American English speakers) share a rough consensus of their implicit rules.
- **A grammar**: an attempt to describe all these rules
- What we are taught as "rules of grammar" often have other purposes than describing the English language



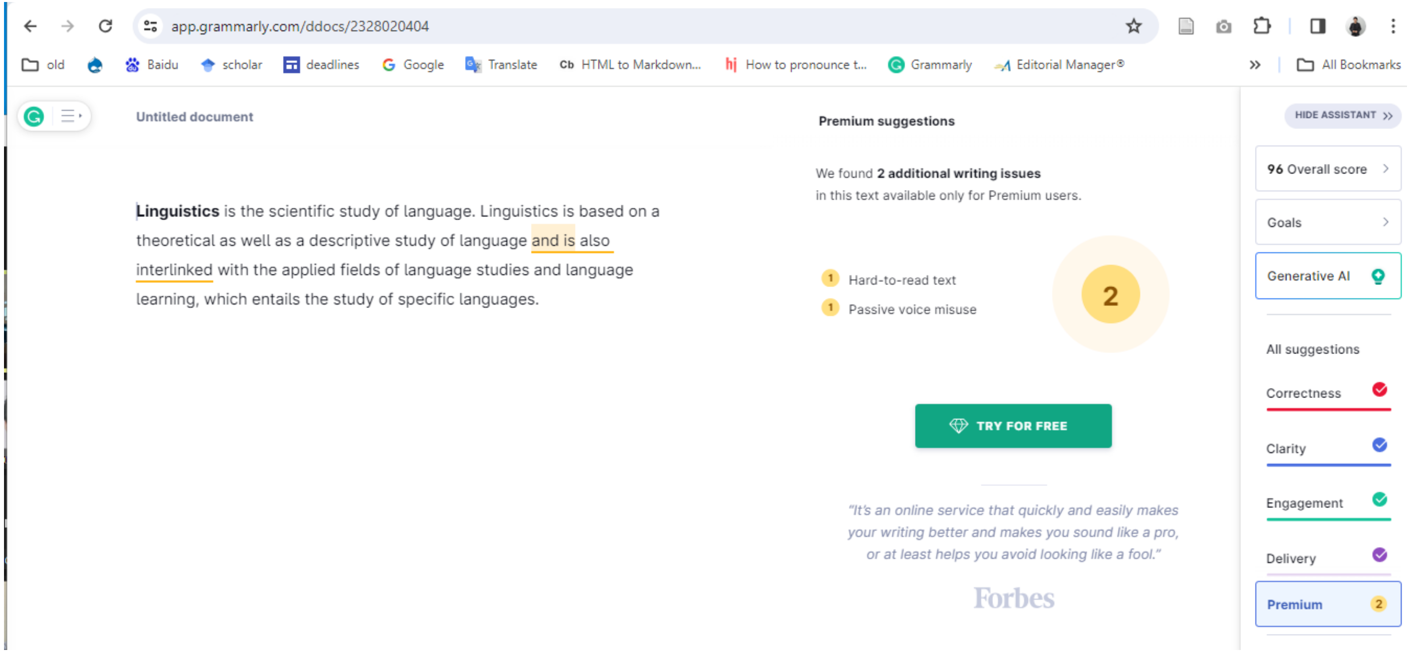
- When they say…
  - Never start a sentence with 'And'
  - It's incorrect to say "I don't want nothing"

- "Focus your thoughts and sound formal for this high school essay"
- "The dialect with the most power in the US does not do negation in this way"

Grammaticality

- A community of speakers (eg, Standard American English speakers) share a rough consensus of their implicit rules.
- All the utterances we can generate from these rules are **grammatical**.
  - If we cannot produce an utterance using these rules, it's **ungrammatical**

Grammarly



https://www.grammarly.com/

ChatGPT could also do this, sometimes much better and more intelligent.

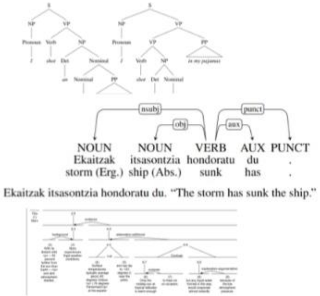
Example

- Subject, Verb, and Object appear in **SVO order**
- **Subject pronouns** (I/she/he/they) have to be subjects, object pronouns (me/her/him/them) have to be objects
    - ✔ "I love her" ✘ "Me love she"
    - ✘ "Me a cupcake ate"

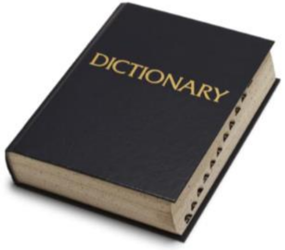
- The meaning is clear
- But our rules of grammaticality **don't seem to cut us much slack**

We rely on linguistic structure to learn language

- A set of rules that define grammaticality



- A lexicon of words that relate to the world we want to talk about



Linguistic structure is the first step in abstracting language.

Linguistic Structure in NLP

Linguistic Structure in NLP

- **Character**: Chinese Characters, Character Encoding
- **Word**: Morphemes, Lemmatization, Tokenization
- **Sentence**: Phrase Parsing, Dependency
- **Discourse**: Discourse Structure and Processing

Beginning

Character

↓

Word

↓

Sentence

↓

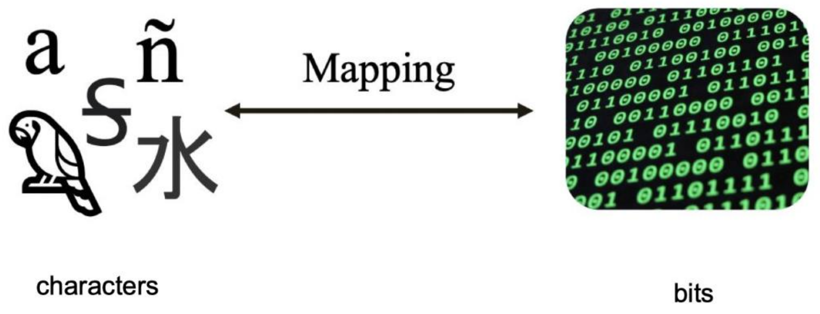
Discourse

This section will explore linguistic structure by examining the different hierarchical levels of language.

Linguistic Structure in NLP

Character

Character



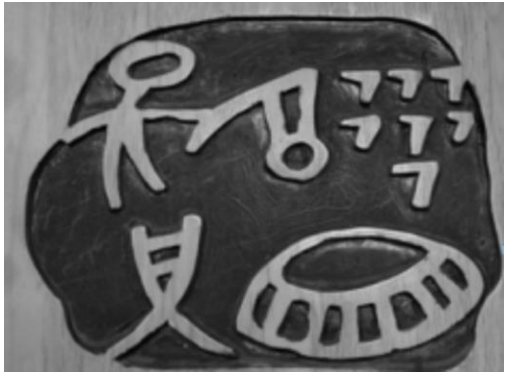
A character is the **smallest unit** of language and also the smallest unit for computer encoding. The **English language consists of only 26 characters**, which can be easily encoded using a Byte.

However, **Chinese characters are totally different** and much more complex.

Chinese Characters 汉字

- Chinese character is non-alphabetic symbol
- The live Ideographic character (表意文字)

pictograph 象形文字



意音 logograph

表意文字/形意文字
(Ideograph)

Chinese Characters 汉字

- Chinese character information

Type	Example	Amount
Pictograph 象形	日月火人	Few
Ideographic 指示	上下	Very few
Compound indicative 会意	仁、信	Very few
Semantic-phonetic compounds 形声	请、情、清、晴 钾、钠、钙、镁	> 90%

Chinese character is morphemes syllable words in ideographic writing system.

Large Number of Characters

- Much larger number of characters as compared with the number of letters
- The exact number of existent Chinese characters cannot be precisely ascertained
    - 康熙字典(Kangxi dictionary 1716) 47, 035
    - 中华字海(Zhonghua Zihai dictionary 1994) 87, 019
    - 1000 characters may cover 92% written materials, and 3000 characters cover more than 99%
- For computer processing purpose, Chinese characters are encoded in 16+ bits. (8bits: #256;16bits: #65536)

Traditional and Simplified Characters

- Traditional/simplified characters
  - Simp. Chinese: China, Singapore, Malaysia, United Nations.
  - Trad. Chinese: Taiwan, Hong Kong and Macau

- No one-one corresponding between
  - 乾‘dry’ and 幹‘to do’ -> Simplified 干
  - 遊‘travel’ and 游‘swim’ -> Simplified 游
  - Sometimes it is a many-to-one mapping

Variant Characters

- Characters that have the same meaning and sounds but different shapes 异体字
- Most of the characters in the Kangxi dictionary are variant character
  - Four variant characters of 回(囘囬囘廻)
- Often share the same components as their standard counterparts
  - 裏/裡; 膀/髈; 杯/盃; 秘/祕; 毙/毙
- Becomes hot in Internet
  - 囧(窘)

Dialect characters and Dialectal Use of Standard Characters

- The existence of dialectal characters 方言字

Cantonese	Meaning	Mandarin
而家	Now	现在
同埋	And	和
边个	Who	哪位
边度	Where	哪儿

Shanghai	Meaning	Mandarin
侬	You	你
伊	He/she	他/她
伐	Not	不
白相	Play	玩

Southern Min	Meaning	Mandarin
阮	I/We	我(们)
暗	Late	晚
郎	Person	人
呷	Eat	吃

Character Encoding Standards

- GB “National Standard” in Chinese
  - 7445 characters. It includes 6, 763 simplified characters. Class-
  - 1/Class-2 characters
  - BG2312-80, contained only one code point for each character.
  - MSB. Bit-8 of each byte, is set to 1, and therefore becomes a 8-bit character. Otherwise, the byte is interpreted as ASCII
  - Every Chinese character is represented by a two-byte code. The MSB of both the first and second bytes are set.

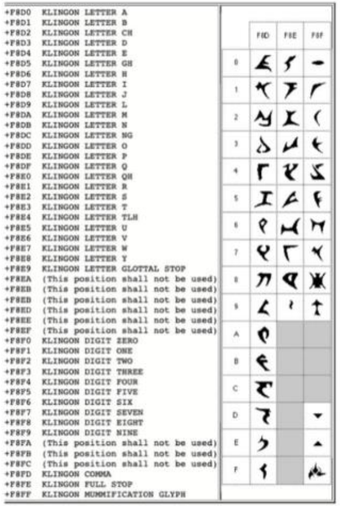
Character Encoding Standards

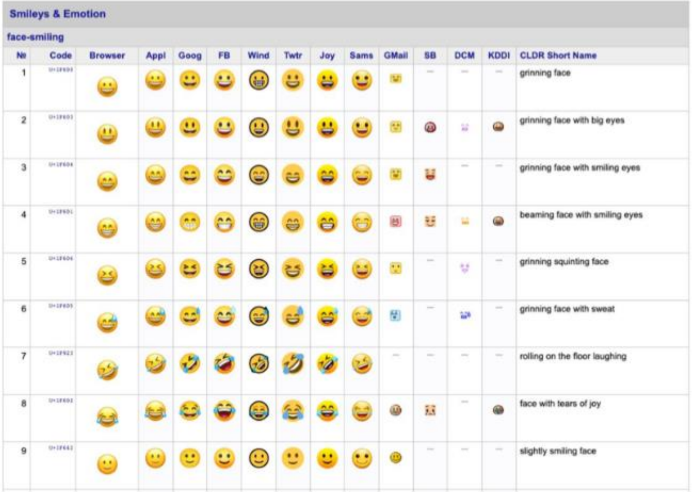
- GBK “National Standard Extension” in Chinese
  - An extension of GB2312
  - Includes 14, 240 traditional characters
  - The scheme is used by Simplified Microsoft Windows 95 and 98

- GB18030
  - Released by the China Standard Press, 2000
  - BG18030 supersedes all previous versions of GB
  - Officially mandatory for all software products sold in the PRC
  - Supports both simplified and traditional Chinese characters

Unicode

- One standard to unify all encoding standards
- ~150K characters, each mapped to a Unicode “codepoint” (e.g. U+0041)





Unicode

- Industry standard, Universal Character Set
  - More than 100, 000 characters
  - Originates from East Asia
- Implemented by different character encodings
  - **UTF-8:** uses 1 byte for all ASCII characters, up to 4 bytes for other characters
  - **UCS-2:** uses 2 bytes for all characters, but does not include every character in the Unicode
  - **UTF-16:** using 4 bytes to encode characters missing from UCS-2
- Simplified and traditional characters as part of the project of Han unification

Example

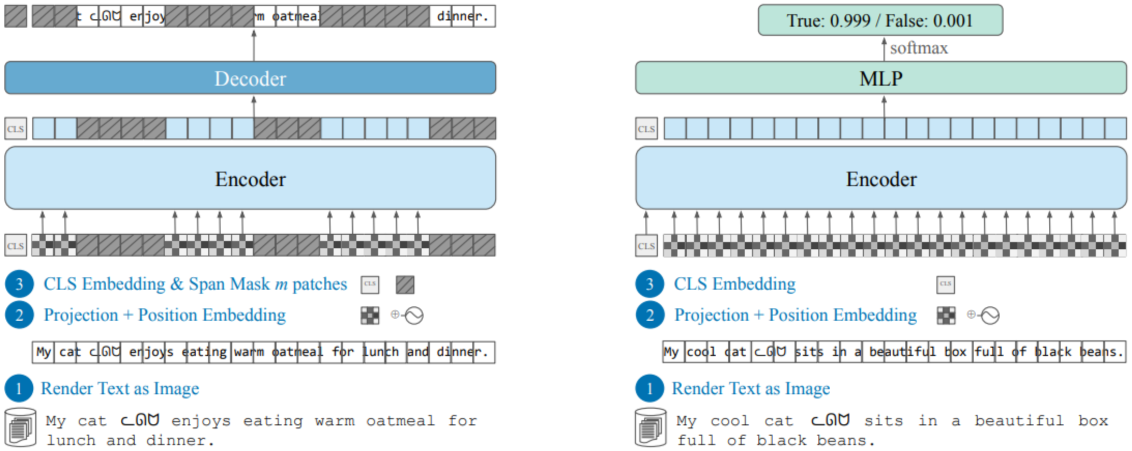
	自	然	语	言	处	理
	B0–F7 lowerA1–EF					
GB2312	D7 D4	C8 BB	D3 EF	D1 D4	B4 A6	C0 ED
	三位 1110（E） 两位 110 一位0 四位1111					
UTF-8	E8 87 AA	E7 84 B6	E8 AF AD	E8 A8 80	E5 A4 84	E7 90 86
	According to UCS					
UTF-16	EA 81	36 71	ED 8B	00 8A	04 59	06 74

Chinese letters vs. Latin letters

- **English – “Reduced Instruction Set Architecture (RISC)”**
- **Chinese – “Complex Instruction Set Architecture (CISC)”**

This becomes interesting when adapting LMs to perceive image input

Language model in pixels



(a) PIXEL pretraining

(b) PIXEL finetuning

Figure 1: Overview of PIXEL's architecture. Following He et al. (2022), we use a masked autoencoder with a ViT architecture and a lightweight decoder for pretraining (left). At finetuning time (right), the decoder is replaced by a task-specific classification head that sits on top of the encoder.

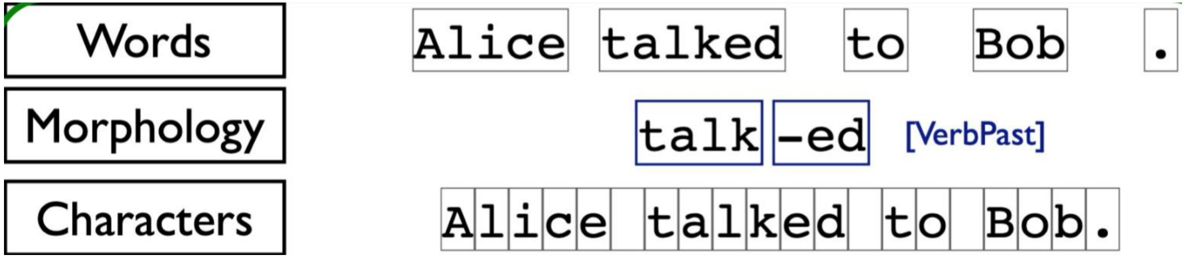
Phillip Rust et. Al. Language modeling with pixels. ICLR 2023

Can we create new words?

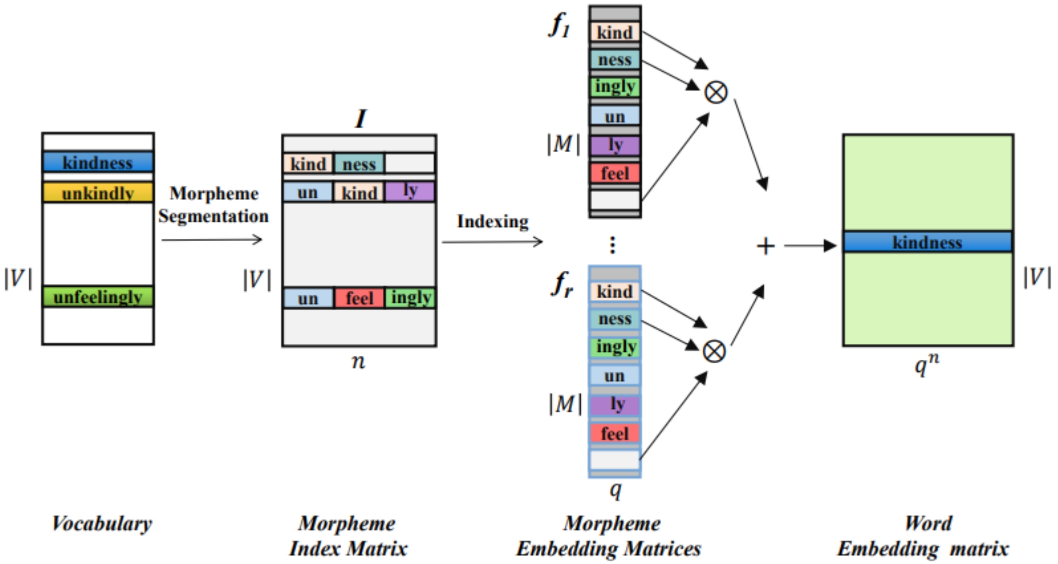
Linguistic Structure in NLP

Word

Character to Word



Our work to inject morphology in word embedding



Guobing Gan, Peng Zhang, Sunzhu Li, Xiuqing Lu, and Benyou Wang. MorphTE: Injecting Morphology in Tensorized Embeddings. NeurIPS 2022

How do we identify the words in a text?

For a language like English, this seems like a really easy problem:

A word is any sequence of alphabetical characters between whitespaces that's not a punctuation mark?

That works to a first approximation, but…

… what about abbreviations like *D.C.*?
… what about complex names like *New York*?
… what about contractions like *doesn't* or *couldn't've*?
… what about *New York-based* ?
… what about names like *SARS-Cov-2*, or R2-D2?
… what about languages like Chinese that have no whitespace, or languages like Turkish where one such "word" may express as much information as an entire English sentence?

Words aren't just defined by blanks

Problem 1: Compounding

"ice cream", "website", "web site", "New York-based"

Problem 2: Other writing systems have no blanks

Chinese: 我开始写小说	=	我	开始	写	小说
		I	*start(ed)*	*writing*	*novel(s)*

Problem 3: Contractions and Clitics

English: "doesn't" , "I'm" ,

Italian: "dirglielo" = dir + gli(e) + lo

tell + *him* + *it*

Tokenization

Any actual NLP system will assume a particular tokenization standard.

- Because so much NLP is based on systems that are trained on particular corpora (text datasets) that everybody uses, these corpora often define a de facto standard.

Penn Treebank 3 standard:

Input:

"The San Francisco-based restaurant,"
they said, "doesn’t charge $10".

Output:

“_ The _ San _ Francisco-based _ restaurant _ , _” _
they_ said_ ,_ "_ does _ n’t _ charge_ $_ 10 _ " _ . _

Spelling variants, typos, etc.

The same word can be written in different ways:

— with different **capitalizations**:
lowercase "cat" (in standard running text)
capitalized "Cat" (as first word in a sentence, or in titles/headlines),
all-caps "CAT" (e.g. in headlines)

— with different **abbreviation** or **hyphenation** styles:
US-based, US based, U.S.-based, U.S. based
US-EU relations, U.S./E.U. relations, ...

— with **spelling variants** (e.g. regional variants of English):
labor vs labour, materialize vs materialise,

— with **typos** (teh)

Good practice: Be aware of (and/or document) any normalization (lowercasing, spell-checking, …) your system uses!

Examples

- The lecture is super looooooooong!
- bert2BERT …

How many different words are there in English?

How large is the **vocabulary** of English
(or any other language)?

Vocabulary size = the number of distinct word types

Google N-gram corpus: 1 trillion tokens,
13 million word types that appear 40+ times

If you count words in text, you will find that…

…a few words (mostly closed-class) are very frequent
(the, be, to, of, and, a, in, that,...)

… most words (all open class) are very rare.

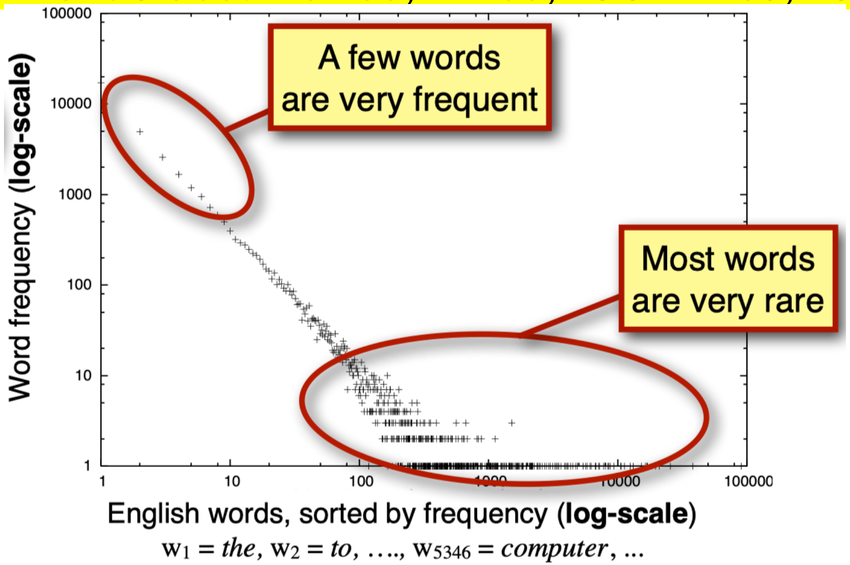
… even if you've read a lot of text,
you will keep finding words you haven't seen before.

Word frequency: the number of occurrences of a word type in a text (or in a collection of texts)

Zipf's law: the long tail

the r-th most common word w_r has $P(w_r) \propto 1/r$

How many words occur once, twice, 100 times, 1000 times?



In natural language:

- A small number of events (e.g. words) occur with **high frequency**
- A large number of events occur with **very low frequency**

Implications of Zipf's Law for NLP

The good:

Any text will contain a number of words that are very **common**.

We have seen these words often enough that we know (almost) everything about them. These words will help us get at the structure (and possibly meaning) of this text.

The bad:

Any text will contain a number of words that are **rare**.

We know *something* about these words, but haven't seen them often enough to know everything about them. They may occur with a meaning or a part of speech we haven't seen before.

The ugly:

Any text will contain a number of words that are **unknown** to us.

We have *never* seen them before, but we still need to get at the structure (and meaning) of these texts.

How many different words are there in English?

How large is the **vocabulary** of English
(or any other language)?

Vocabulary size = the number of distinct word types

Google N-gram corpus: 1 trillion tokens,
13 million word types that appear 40+ times

[here, we're treating inflected forms (took, taking) as distinct]

You may have heard statements such as
"adults know about 30,000 words"
"you need to know at least 5,000 words to be fluent"

Such statements do not refer to inflected word forms (take/takes/taking/take/takes/took) but to lemmas or dictionary forms (take), and assume if you know a lemma, you know all its inflected forms too.

How many different words are there in English?

Actual text doesn't consist of dictionary entries:

wants is a form of want
took is a form of take
courses is a form of course

Linguists distinguish between

— the **(surface) forms** that occur in text:

want, wants, beginners', took,...

— and the **lemmas** that are the uninflected forms of these words:

want, beginner, take, ...

In NLP, we sometimes map words to lemmas (or simpler "stems"), but the raw data always consists of surface forms

Word Mapping

Add some generalization by **mapping** different forms of a word to the same symbol:

- **Normalization:** map all variants of the same word (form) to the same canonical variant (e.g. lowercase everything, normalize spellings, perhaps spell-check)
- **Lemmatization**: map each word to its lemma (esp. in English, the lemma is still a word in the language, but lemmatized text is no longer grammatical)
- **Stemming**: remove endings that differ among word forms (no guarantee that the resulting symbol is an actual word)

Motivating challenge: out-of-vocabulary (OOV) words

Many NLP systems assume a fixed vocabulary, but still have to handle **out-of-vocabulary (OOV)** words.

Example

So if our test sentence was

the dogs love the cats

The input would look like:

The **<OOV> <OOV>** the **<OOV>**

Where **<OOV>** is the “out of vocabulary” symbol

Morphemes: stems, affixes

dis-grace-ful-ly
prefix-stem-suffix-suffix

Many word forms consist of a *stem*
plus a number of *affixes* (*prefixes* or *suffixes*)

Exceptions: *Infixes* are inserted inside the stem

Circumfixes (German *ge*seh*en*) surround the stem

Morphemes: the smallest (meaningful/grammatical) parts of words.

Stems (grace) are often free morphemes.

Free morphemes can occur by themselves as words.

Affixes (dis-, -ful, -ly) are usually bound morphemes.

Bound morphemes *have* to combine with others to form words.

Let's see how Chinese words are different.

Chinese Words

- Distinct from both morphemes at a lower level and from phrases at a higher level

- Distributional restriction
  - Can occur freely by itself vs. bound morphemes
  - Grammatical morphemes 的、地、得
  - Content morphemes 饮 冷饮 饮食

- Integrity of word meanings
  - Have meanings that are not predictable from the meanings of their component morphemes
  - 大人 vs. 小人 打人 vs. 打手

Chinese Word Formation

- Disyllabic compounds
  - 小人 热心 报告 声音
- Tri-syllabic compounds
  - 口香糖 大学生 吹牛皮
- Quad-syllabic compounds
  - 花言巧语 口是心非
- Affixation
  - Prefix: 第一/第二 Suffix: 儿子/ 化学
- Reduplication
  - 商量商量 高高兴兴
- Ionization
  - 理发 ‘cut-hair’ 理短发 ‘have short haircut’ 发理了吗 ‘hair has been cut?’

Challenges in Chinese - Few formal morphological markings

- No verbal inflections:
  - No tense: a verb will have the same form
    - 我过去 是 学生。 I **was** a student.
    - 我现在 是 学生。 I **am** a student.
    - 我将来 是 学生。 I **will be** student.
  - No personal and number agreements
    - 我 去。 I **go**.
    - 她 去。 She **goes**.
- No nominal endings
  - No number marking
    - 我的书 My book(s)
  - No gender marking
  - No case marking
    - **I** love **her**. Vs. **She** loves **me**.
    - (I=subject case; me = object case; she=subject case; her=object case)
    - 我爱她 她爱我。
    - (我=both subject and object case; 她=both subject and object case)

Challenges in Chinese - Ambiguities in Words

- Lexical Ambiguities:
  - 他很好吃
  - 炸鸡很好吃
- Structural Ambiguities
  - Overlapping (crossing) ambiguity 交集型歧义
    - 网球场 美国会
  - Combinatorial ambiguity 组合型歧义
    - 才能 学生会
  - Mixed type 混合型歧义
    - 太平淡 (too dull), 太平 (peaceful), 平淡

A famous example

- 南京市长江大桥
- 南京市长江大桥

How do state-of-the-art models represent word?

Tokenization for GPT-3, GPT-2... most likely ChatGPT

Tokens 97

Characters 248

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

一个有用的经验法则是，对于通常英文文本而言，一个标记对应约4个字符的文本。这大约相当于一个字的¾（因此100个标记~=75 个）。

TEXT TOKEN IDS

https://platform.openai.com/tokenizer

Tokenization for GPT-3, GPT-2… most likely ChatGPT

Byte Pair Encoding (BPE) (Sennrich et al., 2015) is a practical middle ground between character and word level language modeling which effectively interpolates between word level inputs for frequent symbol sequences and character level inputs for infrequent symbol sequences. Despite

Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

Byte-Pair Encoding (BPE) for Tokenization

- Use our *data* to automatically tell us what tokens should be
- **Aim:** induce tokens that are **subwords** that
  - Are smaller than words
  - Can be morphemes (e.g., *-est* or *-er)*
  - Or just arbitrary substrings
- Useful for dealing with words never been seen before at training time (out-of-vocabulary words)

Token learner
Input: Raw training corpus
Output: Induces vocabulary via BPE

Token segmenter
Input: Learned vocabulary, test sentence
Output: Segmented token list

Byte-Pair Encoding (BPE) deals OOV words

1. Convert each character into **unicode** bytes

2. Use these **bytes as the base vocabulary** for the BPE algorithm.

3. Tokenize via BPE.

4. Big picture: <u>Now we can deal with any word or character!</u>

Parts of Speech

DT	VBZ	DT	JJ	NN	**PART OF SPEECH**
This	is	a	simple	sentence	**WORDS**
	be 3sg present		SIMPLE1	SENTENCE1	**MORPHOLOGY**

"Parts of speech" are categories of words based on their function within a sentence. Understanding them is fundamental to grasping the rules of grammar and syntax in a language.

Examples from 文心一言

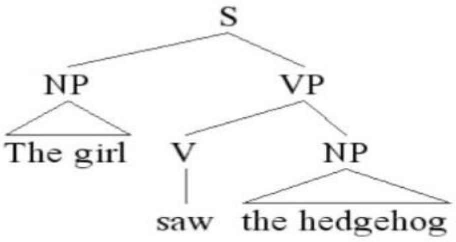
Mixture of characters and words

Linguistic Structure in NLP

Sentence

Phrase

- Words are organized into phrases, then comes to sentence



- Syntax studies the regularities and constraints of word order and phrase structure

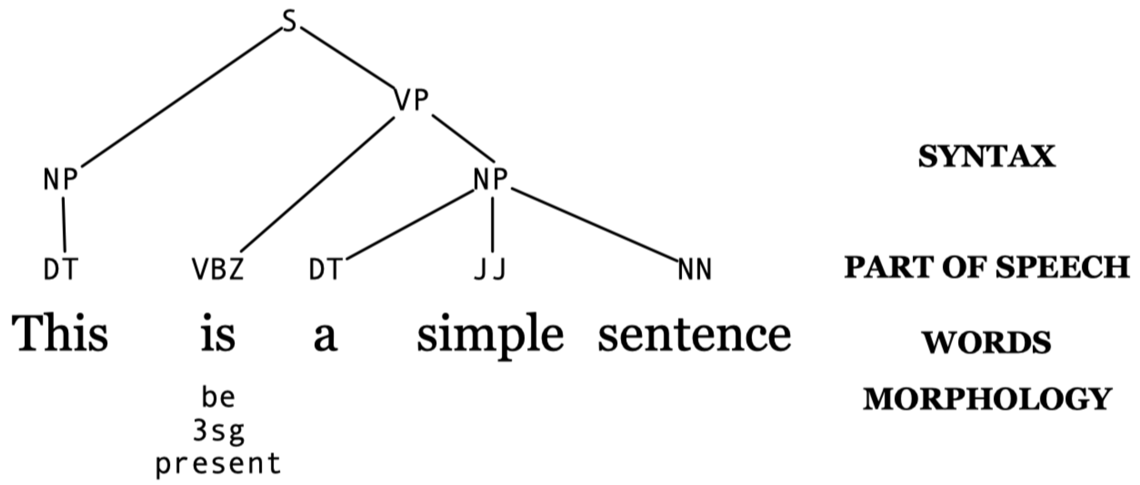
- Major phrase categories in English
  - Noun phrase. Prepositional phrases
  - Verb phrases. Adjective phrases

English Phrase

Category	Description	Examples
Noun Phrase (NP)	A noun and all its modifiers	**The bewildered tourist** was lost.
Verb Phrase (VP)	A verb and all its modifiers	He **was waiting for the rain to stop**.
Gerund Phrase (GP)	A noun phrase that starts with a gerund.	**Taking my dog for a walk** is fun.
Infinitive Phrase (IP)	A noun phrase that begins with an infinitive verb.	**To make lemonade**, you have **to start with lemons**.
Appositive Phrase (AP)	It restates and defines a noun. It consists of one or more words.	**My favorite pastime**, needlepoint, surprises some people.
Participial Phrase (PP)	Begins with a past or present participle.	**Washed with my clothes**, my cell phone no longer worked.
Absolute Phrase	It modifies the whole sentence, not just a noun.	**His tail between his legs**, the dog walked out the door.

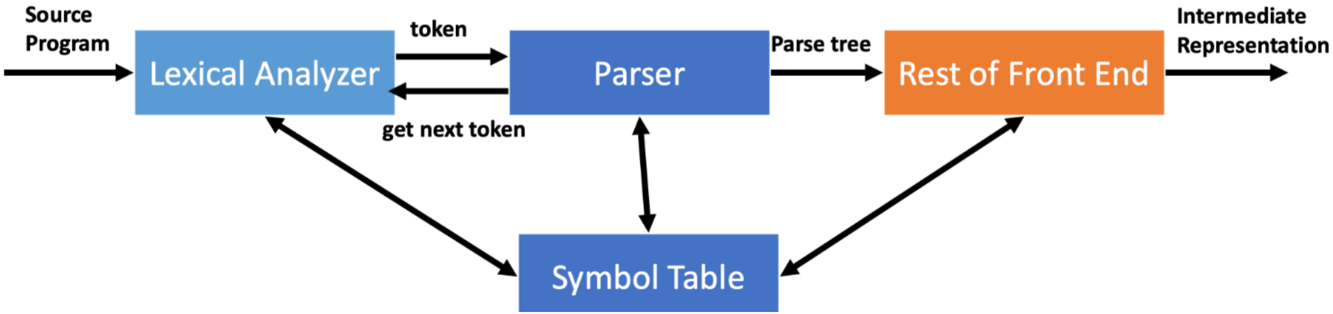
Phrase Structure Grammar

Tells us how to determine the meaning of the sentence from the meaning of the words.



Parsing (Syntax Analysis)

- Syntax
  - Is one of the major components of **grammar**.
  - Is the proper order of words in a phrase or sentence.
  - Is a tool used in writing proper grammatical sentences
  - Native speakers of a language learn correct syntax without realizing it.
  - The complexity of a writer's or speaker's sentences creates a formal or informal level of diction that is presented to its audience.



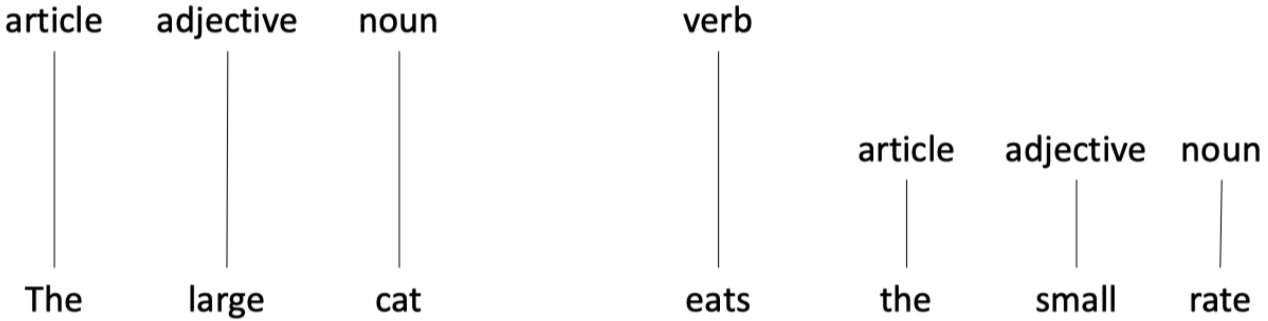
Parse tree (Syntactic Tree)

"The large cat eats the small rat"

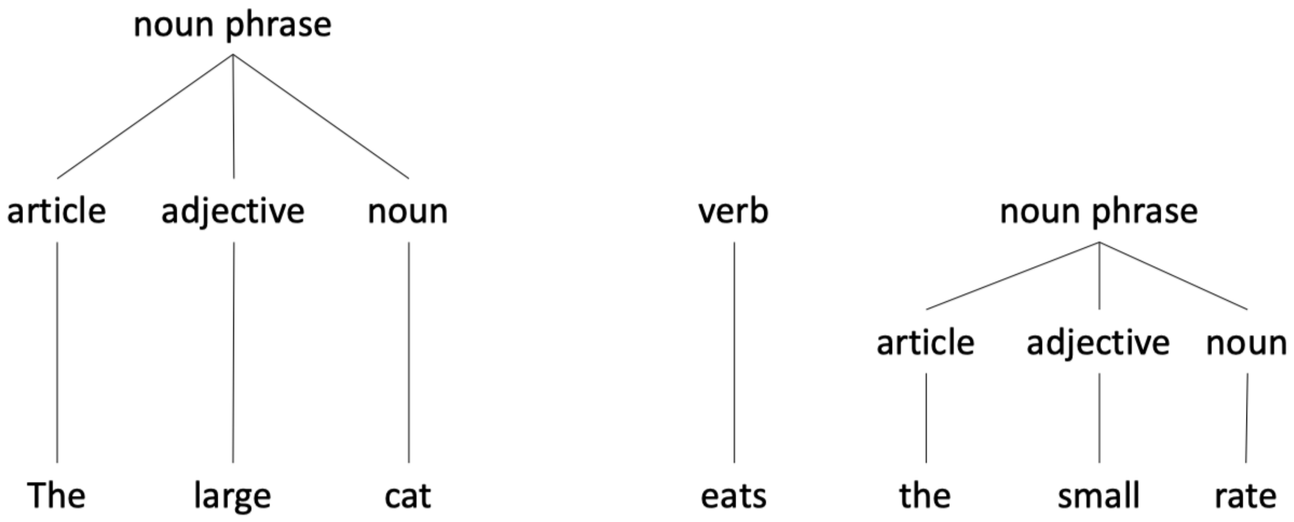
Parse tree (Syntactic Tree)

The large cat eats the small rate

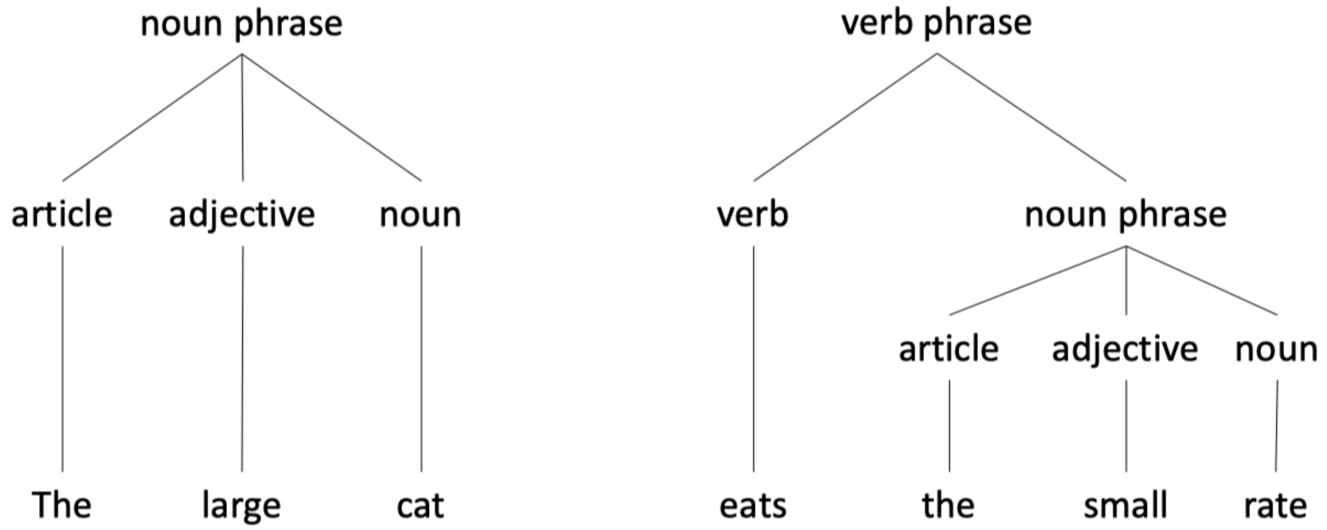
Parse tree (Syntactic Tree)



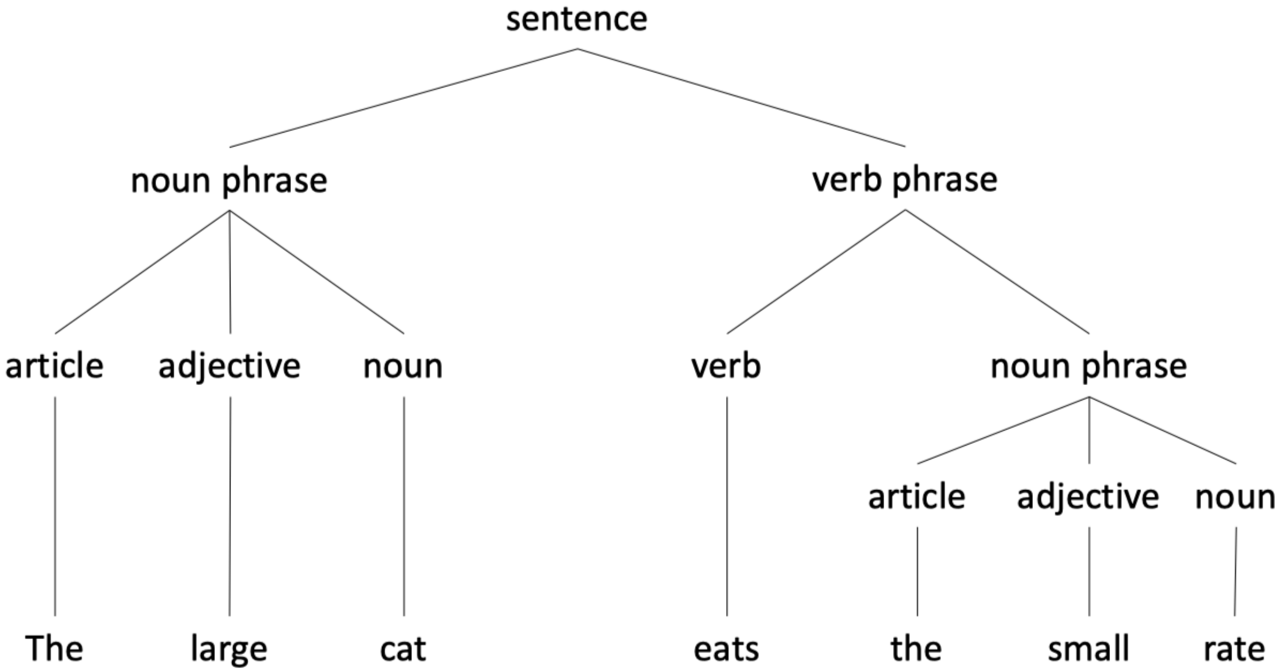
Parse tree (Syntactic Tree)



Parse tree (Syntactic Tree)

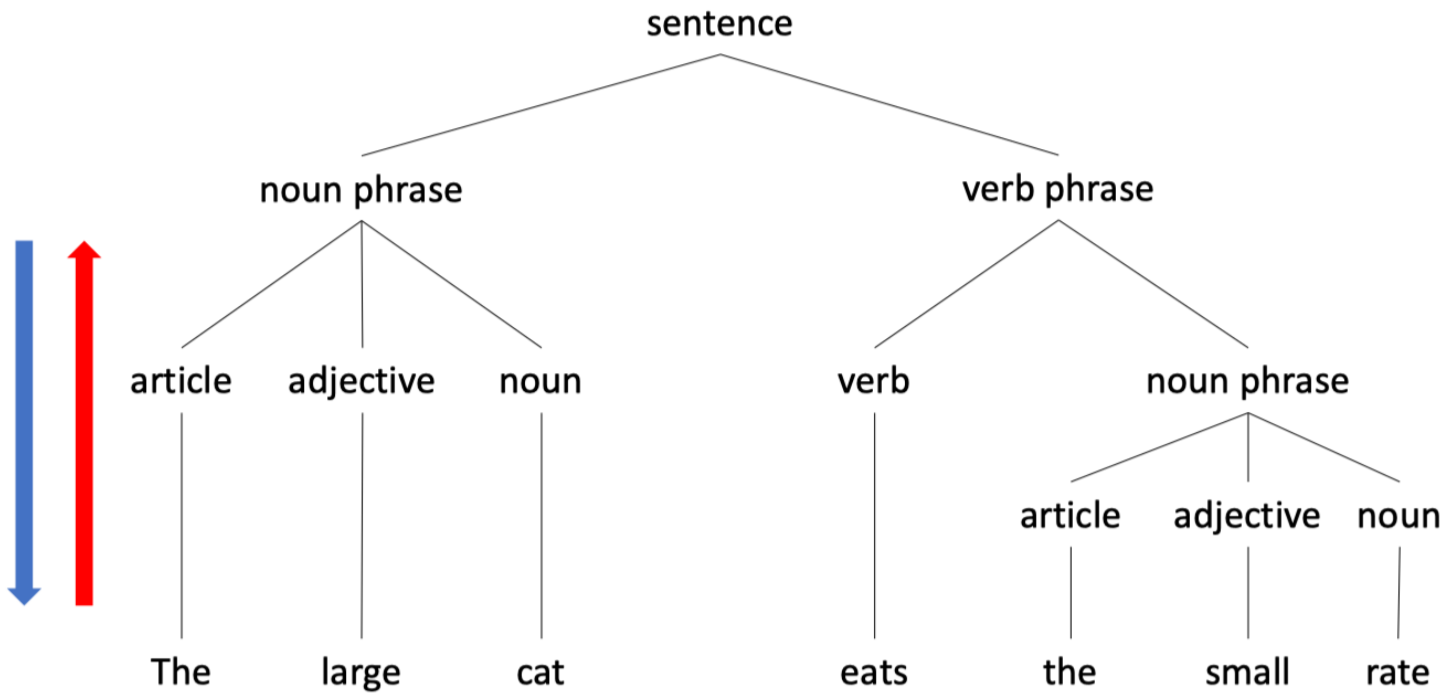


Parse tree (Syntactic Tree)



Parse tree (Syntactic Tree)

- Bottom-up or Top-down



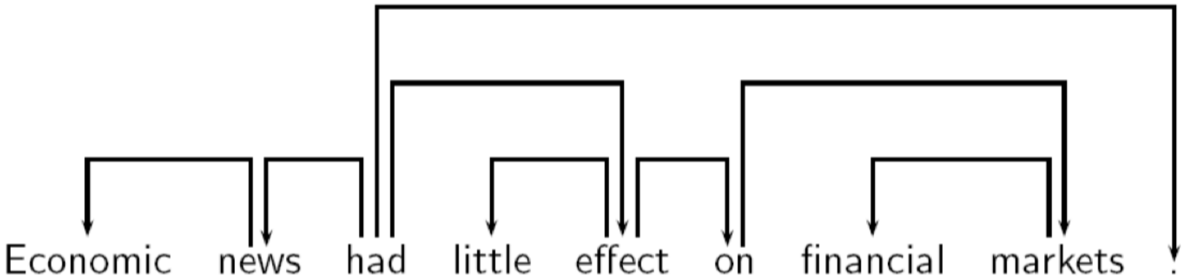
Dependency

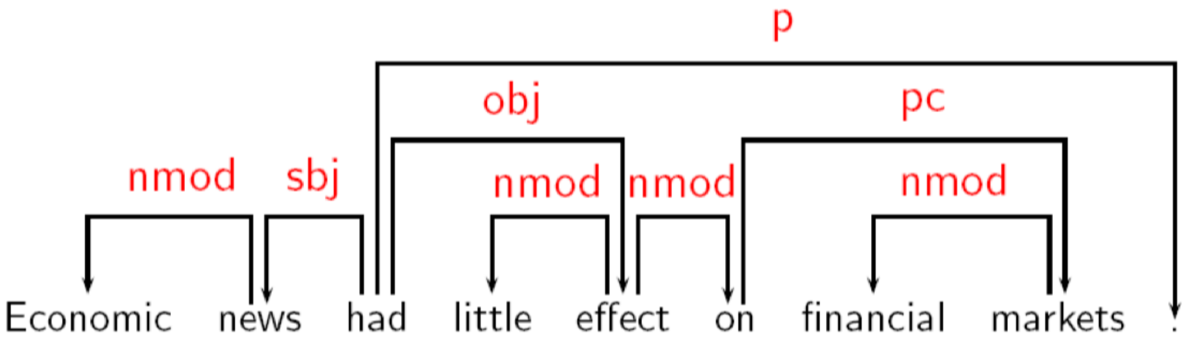
Syntactic structure consists of **lexical items**, linked by binary asymmetric relations called **dependencies**.

The sentence is an *organized whole*, the constituent elements of which are *words*. Every word that belongs to a sentence ceases by itself to be isolated as in the dictionary. Between the word and its neighbors, the mind perceives *connections*, the totality of which forms the structure of the sentence. The structural connections establish *dependency* relations between the words. Each connection in principle unites a *superior* term and an *inferior* term. The superior term receives the name *governor*. The inferior term receives the name *subordinate*.

Tesnière, L. (1959). Eléments de Syntaxe Structurale.

Dependency Structure

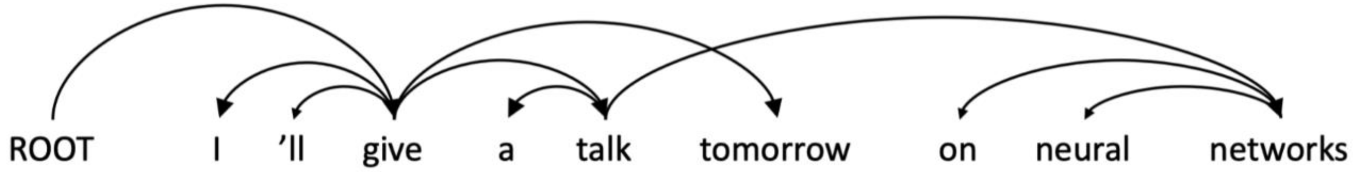




Dependency Parsing

- A sentence is parsed by choosing for each word what other word (including ROOT) it is a dependent of

- Usually some constraints:
  - Only one word is a dependent of ROOT
  - Don't want cycles A → B, B → A
- This makes the dependencies a tree
- Final issue is whether arrows can cross (be non-projective) or not



Phrase Structure vs. Dependency Structure

	Phrase structure	Dependency structure
word relation	phrasal constitution	head-dependent
categories	syntactic functional	syntactic structural
new node (Y/N)	multiple nodes per word	one node per word
operation	waiting for complete phrase	word-at-a-time

Problems in Parsing - Ambiguity

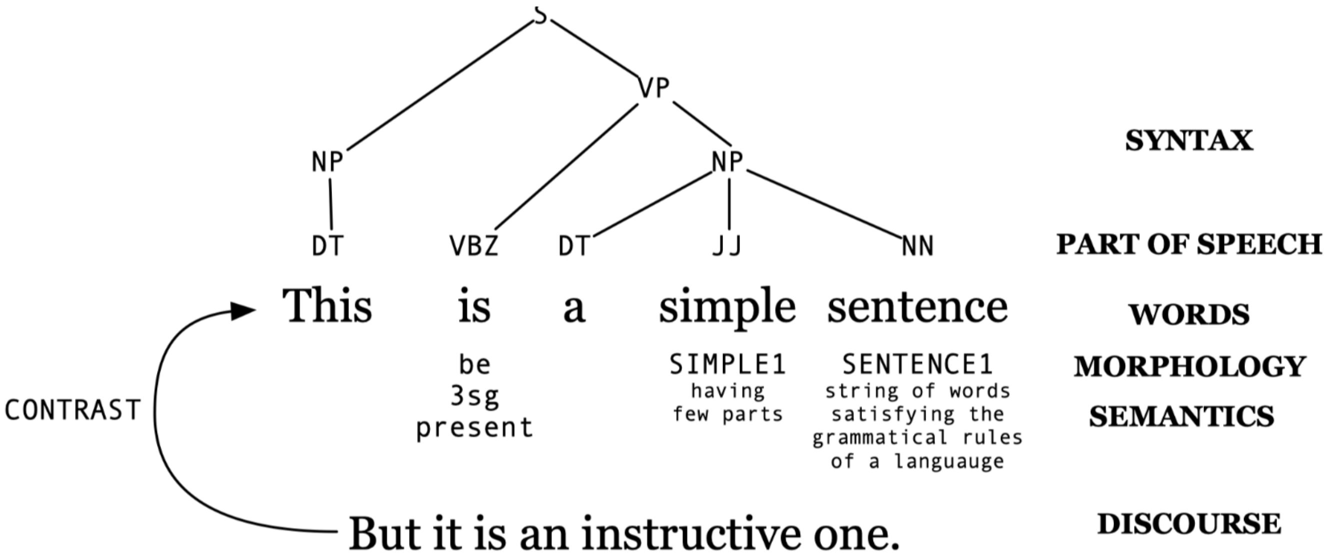
- "One morning, I shot an elephant in my pajamas. How he got into my pajamas, I don't know."
- Syntactical/Structural ambiguity
  - Several parse trees are possible e.g. above sentence
- Semantic/Lexical ambiguity
  - Several word meanings. e.g. bank (银行) or bank (河流)
- Different word categories
  - "He **books** the flight" vs. "The **books** are here."
  - "Fruit flies from the balcony" vs. "Fruit flies are on the balcony"
- Attachment
  - In particular PP (prepositional phrase) binding; often referred to as "binding problem"
  - "One morning, I shot an elephant **in my pajamas**."

Linguistic Structure in NLP

Discourse

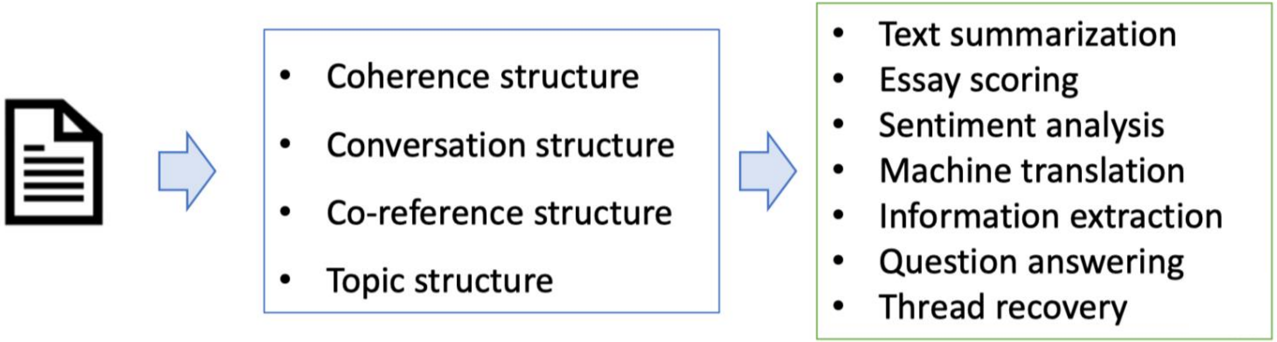
(篇章/语篇)

Discourse Structure



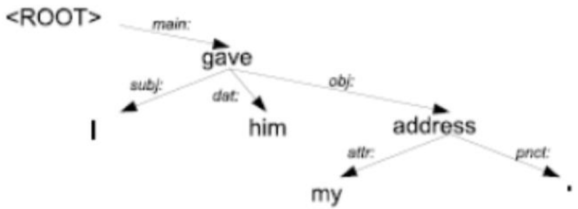
Discourse Processing

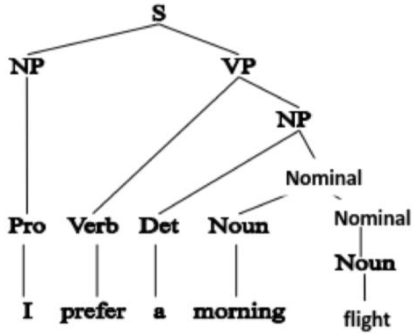
- **Discourse** is a group of collocated and coherent sentences
- **Discourse theory** deals with language phenomena that operate beyond the single sentence
- **Discourse analysis/processing** is a suite of **Natural Language Processing (NLP) tasks** to uncover linguistic structures from multi-sentential texts at several levels, which can support many "downstream" NLP applications.



Modeling Coherence Structure

- **Grammaticality:** distinguishes well-structured sentences from random sequences of words. Models are specified as groupings and relations between words





- **Coherence** plays the same role at the multi-sentence level. Models are also specified as groupings and relations between …?

Discourse/Coherence Relations

- Specify the relations between **sentences** or **clauses**.
- Due to the relations, two adjacent sentences can look coherent.

What is the discourse relation between the following two sentences?

"Explanation" relation

John hid Bill's car keys. He was drunk

vs.

John hid Bill's car keys. He likes spinach

More Discourse Relations

- Elaboration 细化
  - Dorothy was from Kansas. She lived on the Kansas prairies.
- Result 因果
  - The tin woodman was caught in the rain. His joints rusted.
- Parallel 并列
  - The scarecrow wanted some brains. The tin woodsman wanted a heart.

- How many relations? Which ones?
- What kind of structures? Flat? Trees? Graphs?

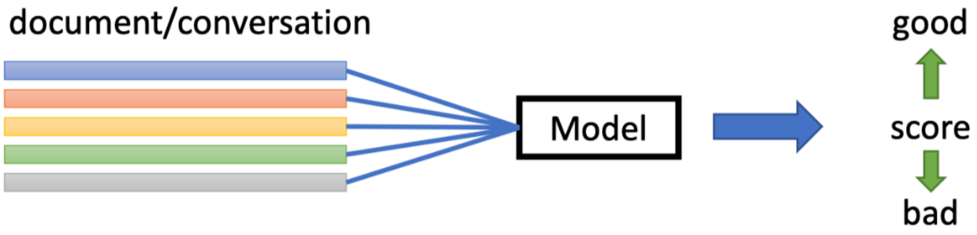
Discourse Parse

- Either two clauses / sentences connected by an explicit <u>connective</u>
  - The federal government suspended sales of U.S. savings bonds <u>because</u> Congress hasn’t lifted the ceiling on government debt.
    cause-reason
  - The subject will be written into the plots of prime-time shows, <u>and</u> viewers will be given a 900 number to call. Conjunction

- Or two adjacent sentences connected by an implicit connective
  - Some have raised their cash positions to record levels. High cash positions help buffer a fund when the market falls.
    Implicit=because(cause-reason)

Webber B L, Joshi A K. Anchoring a lexicalized tree-adjoining grammar for discourse[J]. arXiv preprint cmplg/9806017, 1998.

Coherence Models



- Helps predict which sentences are pragmatically appropriate
- Tells us which sentences are closely related

- Applications
  - Essay scoring
  - Summarization (sentence selection and ordering)
  - Generation (including MT)
  - …

An interesting fact

Lexicon and Grammar could be changing

- A case in Lexicon
  - Chinglish: open your computer vs. turn on you computer
- A case in Grammar
  - You okay?

Linguistic Structure for today's NLP

Linguistic Structure for today's NLP

Implicit Mastery: Evolving Beyond Explicit Linguistic Rules in Modern LLMs

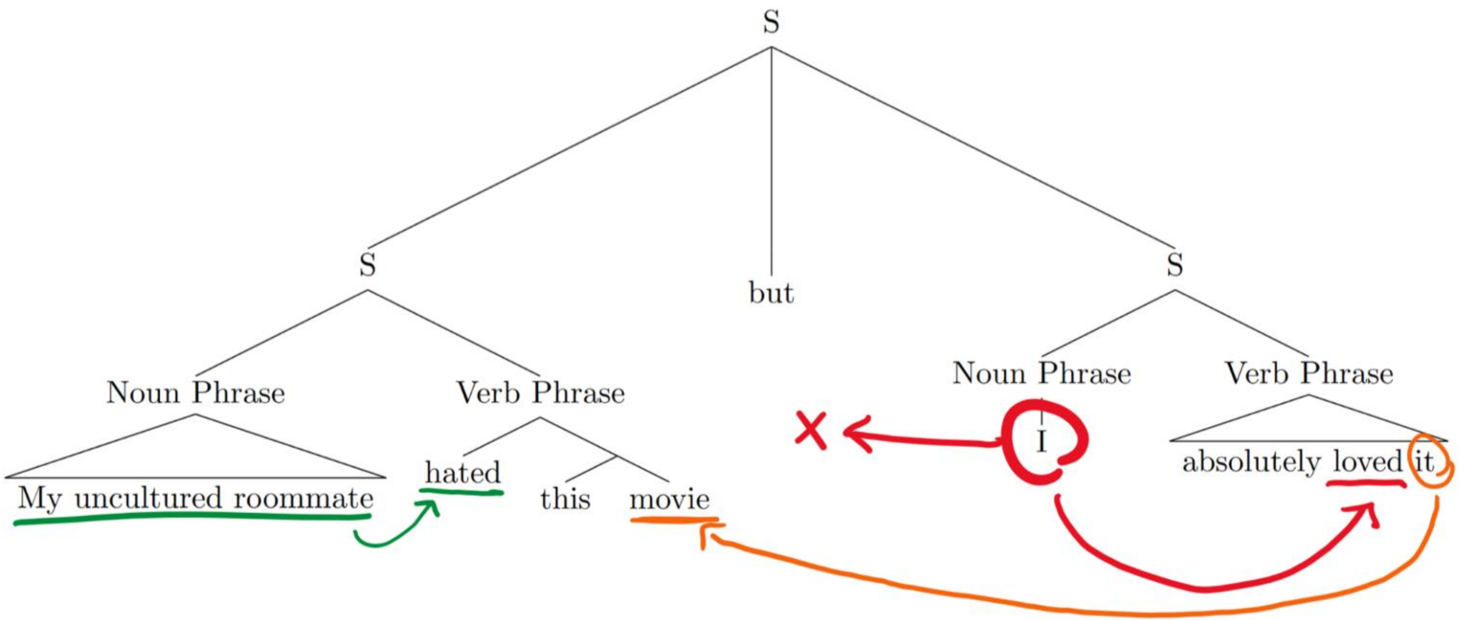
Before Self-supervised Learning

Before self-supervised learning Like GPT, the way to approach doing NLP was through understanding the human language system, and trying to imitate it.

- **Example:** Parsing
  - I want my sentiment analysis system to classify this movie review correctly
    - "My uncultured roommate hated this movie, but I absolutely loved it"
  - How would we do this?
  - We might have some semantic representation of some key words like "hate" and "uncultured", but how does everything relate?

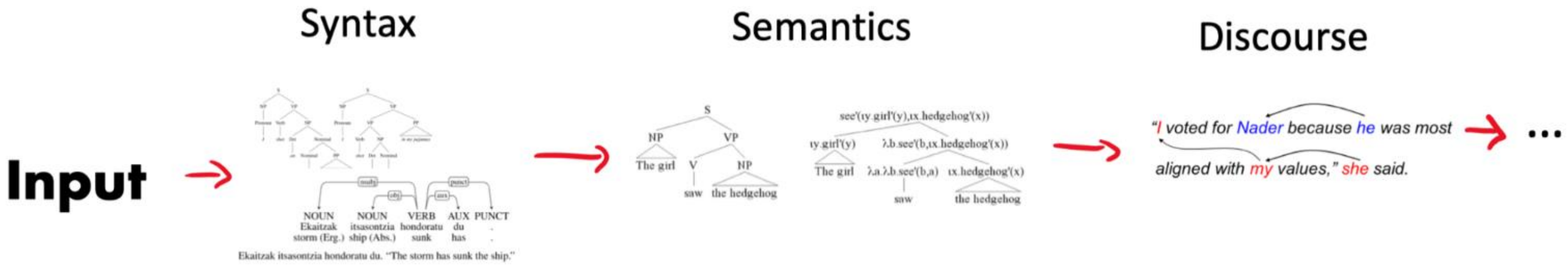
Before Self-supervised Learning

Many linguists might tell you something like this:



Before Self-supervised Learning

They built **Pipelines** for language processing based on language structures.



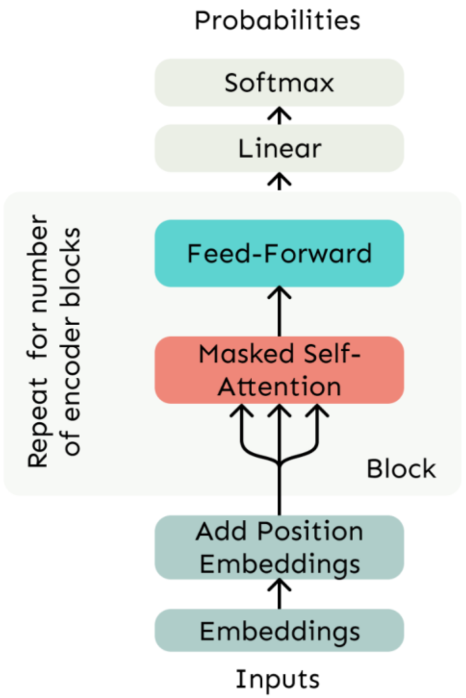
Now, language models just seem to catch on to a lot of these things!

IS

Hi! Whaddya think of this movie review? "My uncultured roommate hated this movie, but I absolutely loved it"

As an AI language model, I do not have the ability to form opinions or beliefs. However, based on the given movie review, it seems that the person who wrote it had a positive experience watching the movie, while their roommate did not enjoy it. It's important to remember that everyone has their own tastes and preferences when it comes to movies and other forms of entertainment, and it's okay to have differing opinions. It's always good to approach a movie with an open mind and form your own opinion rather than solely relying on someone else's viewpoint.

How we train our models these days!



Directly learn from unstructured corpora through self-supervised learning, without concerning with Linguistic Structure.

Structured Learning to non-structured Learning

Contemporary Large Language Models like GPT-4 employ a non-structured learning approach, directly learning from raw corpora. This contrasts with earlier methods that depended heavily on explicit, rule-based learning of linguistic structure.

This approach is favored for several reasons:

- It allows for scalability, facilitating the learning of vast, unstructured natural language corpora without the need for explicit programming of linguistic rules.
- Language structures are inherently complex and vary depending on context and word usage, making them difficult to effectively encode and learn.

This is akin to us as humans

When we read and speak, we do not deliberately analyze the grammatical structure of each sentence.

But we implicitly know complex rules about structure

- What can we pull out to make a question?

Leon is a doctor	What is Leon?
My cat likes tuna	What does my cat like?
Leon is a doctor and an activist	✕ What is Leon a doctor and?

But we implicitly know complex rules about structure

- When can we move the object to the end?

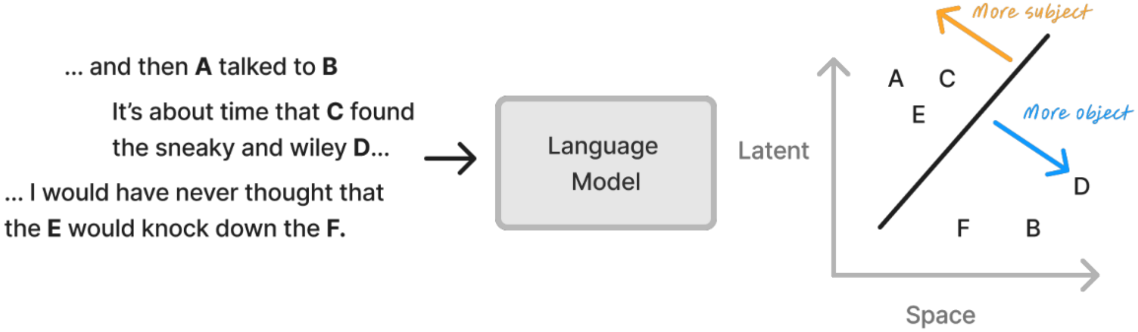
I dictated the letter to my secretary

I dictated the letter that I had been procrastinating writing for weeks and weeks to my secretary

I dictated to my secretary the letter that I had been procrastinating writing for weeks and weeks

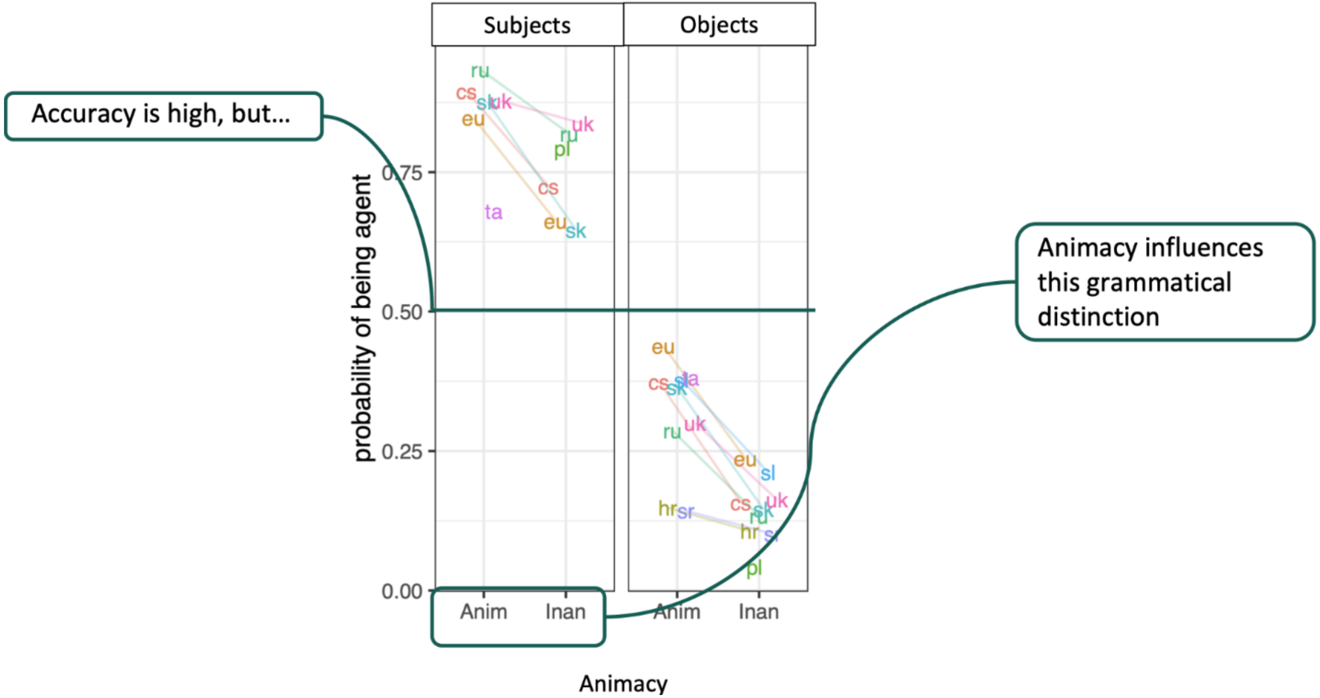
✕ I dictated to my secretary the letter

Language models are also aware of linguistic structure



Papadimitriou, Isabel, et al. "Deep subjecthood: Higher-order grammatical features in multilingual BERT." *arXiv preprint arXiv:2101.11043* (2021).

Language models are also aware of linguistic structure



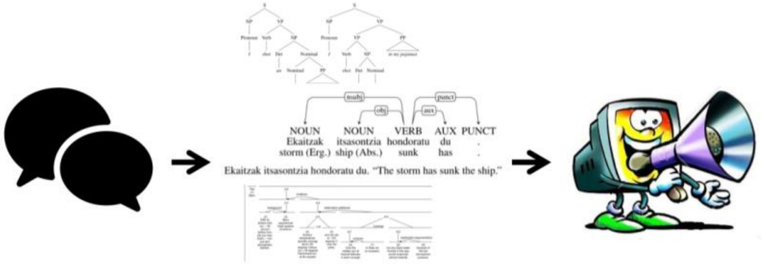
Papadimitriou, Isabel, et al. "Deep subjecthood: Higher-order grammatical features in multilingual BERT." *arXiv preprint arXiv:2101.11043* (2021).

Structure in language models

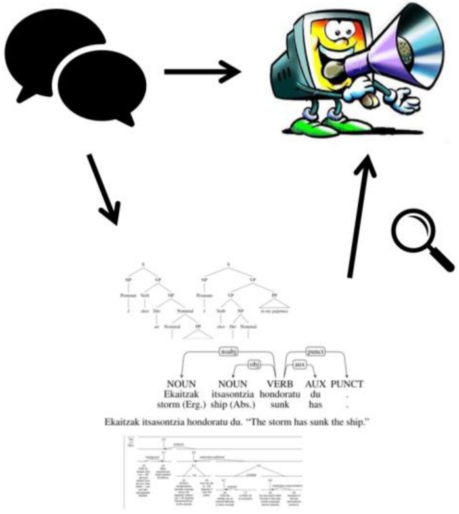
- Language models aren't engineered around discrete, linguistic rules
- But the pretraining process isn't just a bunch of surface-level memorization
- How much do we get a discrete, rule-based system from large scale pretraining?
    - There's syntactic knowledge, but it's complicated
- But remember – there's no ground truth for how language works!
    - If we knew how to fully describe English with a bunch of discrete rules, we would just make an old-school NLP system and it would be amazing.

Large language models: a paradigm to implicitly learn linguistics

Before:



Now:



(hidden comprehension)

Linguistic Structure for today's NLP

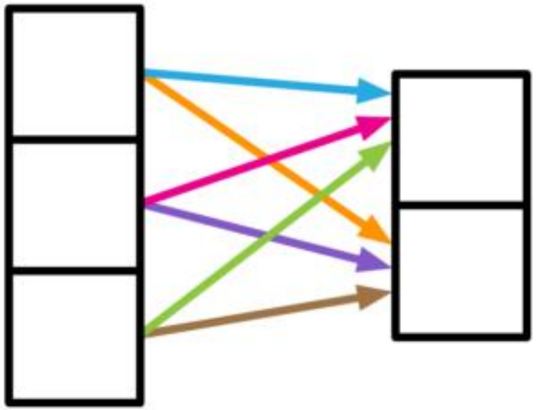
How does it work with linguistic structure?

Inductive bias

- Every machine learning model requires some type of architecture design and possibly some initial assumptions about the data we want to analyze. **Generally, every building block and every belief that we make about the data is a form of inductive bias.**

Some typical inductive biases in neural networks

- **Weak Relation**

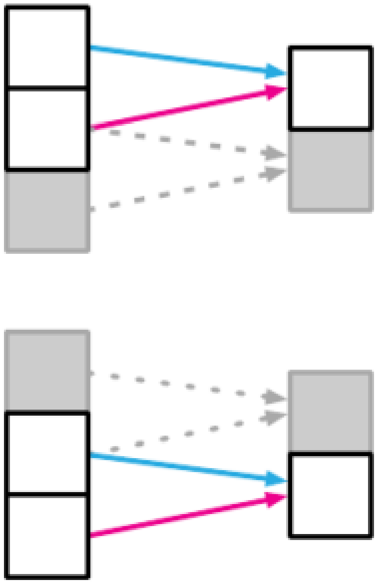


So we need **fully-connected** layers

https://www.baeldung.com/cs/ml-inductive-bias

Some typical inductive biases in neural networks

- **Locality**

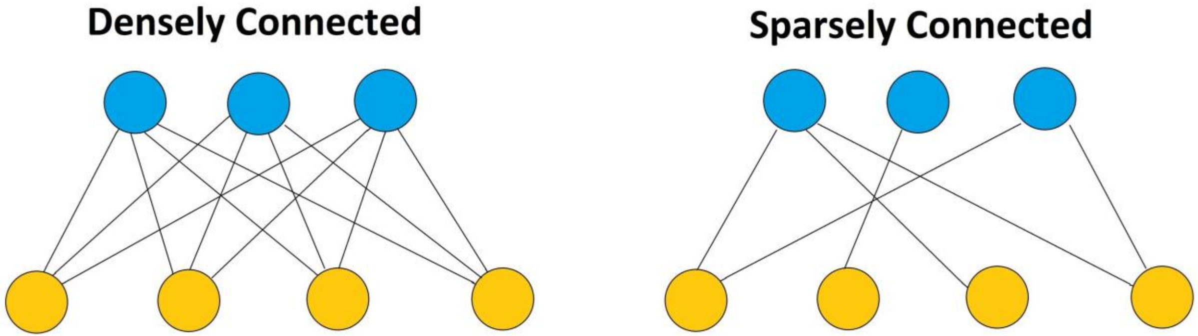


So we could use **local connections**

https://www.baeldung.com/cs/ml-inductive-bias

Some typical inductive biases in neural networks

- **Sparsity**

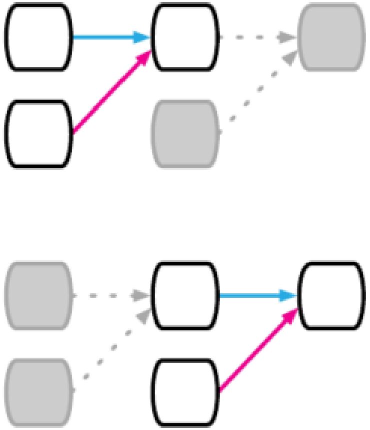


So we could do pruning and MOE

https://www.baeldung.com/cs/ml-inductive-bias

Some typical inductive biases in neural networks

- **Sequential Relation**

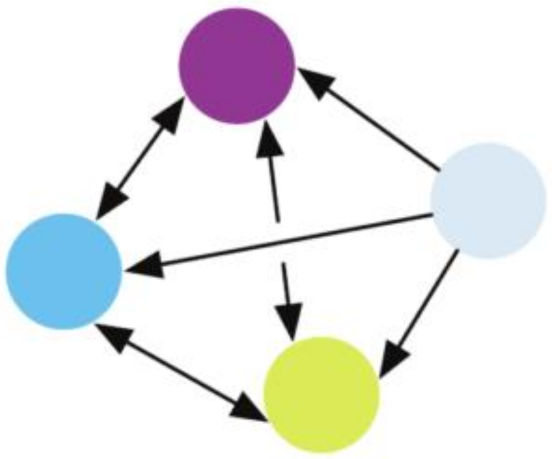


we could model input **sequence**

https://www.baeldung.com/cs/ml-inductive-bias

Some typical inductive biases in neural networks

- **Arbitrary Relation**



we could model input **graphs**

https://www.baeldung.com/cs/ml-inductive-bias

Inductive bias in NLP

Natural language is
- Bag of words?
- A sequence?
- A (syntax or dependency) tree?

How modern NN perceives structure

- Bag of words
- Word sequence
- Injected structure
  - syntax or dependency tree (Recursive NN)
  - with local connections (Convolution NN)
  - with a recurrent bias (Recurrent NN)

Transformer： bag-of-words models with **position embeddings**

Structure is learned in a data-driven way thanks to free attention.

Benyou Wang, Lifeng Shang, Christina Lioma, Xin Jiang, Qun Liu, Jakob Grue Simonsen. On position embeddings in BERT. ICLR 2021,
Benyou Wang*, Donghao Zhao*, Christina Lioma, Qiuchi Li, Peng Zhang, Jakob Grue Simonsen. Encoding word order in complex embeddings. ICLR 2020, Spotlight paper (acceptance rate: 6%)

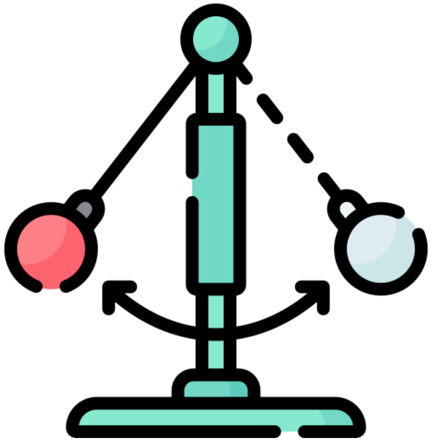
Linguistic Structure for today's NLP

Pendulum Swing: The Future Convergence of Linguistics and Deep Learning

Pendulum Swing

- **Empiricism**: Empiricism in NLP, exemplified by deep learning, relies on vast amounts of data and statistical models to understand and generate human language.
- **Rationalism**: Emphasizes linguistic structure, advocates for a more rule-based approach, rooted in the understanding of grammatical and syntactic principles of language.

Empiricism
(e.g., deep learning in NLP)



Rationalism
(e.g., Linguistic Structure)

The Empirical Era: Deep Learning's Triumph in NLP

- we are currently witnessing a pronounced **swing towards empiricism**, primarily fueled by the advancements in deep learning.
- This era is marked by the emergence of sophisticated language models like GPT-4 and BERT, which have revolutionized our ability to process and generate human language.
- These models, built on the backbone of massive datasets, excel in a variety of tasks such as text generation, translation, and semantic understanding. The reliance on extensive data has not only improved accuracy but also expanded the scope of applications, making NLP more versatile and accessible than ever before.

As the pendulum inevitably swings

However, the dominance of empirical methods in NLP is not without its challenges. While these models excel in processing vast amounts of data, they often grapple with issues like hallucinations, lack of control, and opacity in their decision-making processes.These problems, seemingly intractable within the current framework of deep learning, point to inherent limitations.

As the pendulum inevitably swings, the future of Large Language Models (LLMs) might see a greater integration of rationalism. Such a combination could lead to more robust, controllable, and interpretable AI systems. As LLMs gradually encounter developmental bottlenecks, the incorporation of linguistic structure and theory promises to unlock new avenues and breakthroughs in the field.

Acknowledgement

- CS224N/Ling284: Natural Language Processing with Deep Learning, Stanford University
- COMP5021: Natural Language Processing, Xu Ruifeng ,Harbin Institute of Technology, Shenzhen
- CS447: Natural Language Processing, University of Illinois Urbana-Champaign
- CSCI 375: Natural Language Processing(Spring 2023), Williams College
- INFR10078, Foundations of Natural Language Processing, The University of Edinburgh

Thanks

Tips for scaling

- Structure-free
- Enough complexity (or smaller models solved it and it turns to be saturated)
- Fully-interacted between features internally
- End-to-end training without pipelines (Pipelines leads to gradient stopping)
- Multi-tasking for generalization (also improve and complexity)
- Fully-trainable parameters