Special Topics in Natural Language Processing

CS698O

Ashutosh Modi
CSE Department, IIT Kanpur

Lecture 3: Linguistic Fundamentals-2 and NLP Pipeline
Jan 8, 2020

LINGUISTIC FUNDAMENTALS

Linguistic Fundamentals

- Reviews basic structure of English in terms of word categories and phrases categories

- It applies to other languages as well

Note: Most of this section is based on Chapter 2 of [4]

Words

- “Word” appears to be the most basic unit of a language

Running *Drank*

Words

- "Word" appears to be the most basic unit of a language

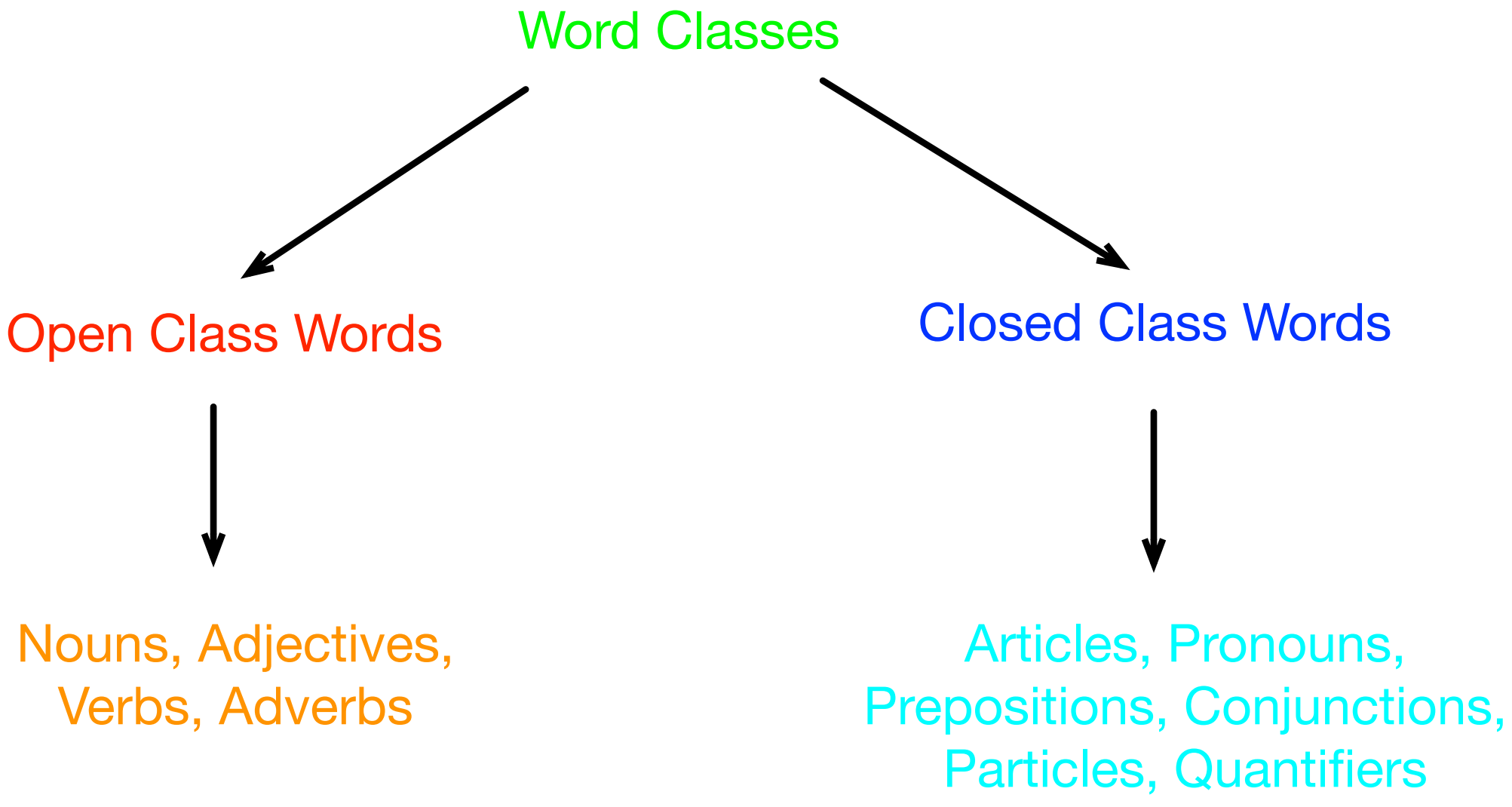
Running *Drank*

- But word are result of complex set of rules on more primitive parts.

- Morphology is the study of how words are built up from more basic components (**Morphemes**) corresponding to minimal meaning units.

Run → Ran, Running

Goose → Geese *Fox → Foxes*


Word Classes
Open Class Words
Closed Class Words
Nouns, Adjectives,
Verbs, Adverbs
Articles, Pronouns,
Prepositions, Conjunctions,
Particles, Quantifiers

Phrases

- Phrase is one or more words functioning as a unit in sentence. [5]
- Every phrase has a head
- Head indicates the type of thing, action or quality that phrase describes. Head belongs to open class words

the dog

the ugly dog

the ugly dog at the ground

Phrases

- Noun Phrases
- Adjective Phrases
- Verb Phrases
- Adverbial Phrases

The desire to succeed

angry as a hippo

ate the cake

rapidly like a bird

Phrases

- Many times a phrase needs additional phrases following it to express desired meaning

Jack put

Phrases

- Many times a phrase needs additional phrases following it to express desired meaning

Jack put

- These additional phrase(s) needed to complete the meaning are called the **complement of the head**

Jack put the dog in the house

Noun Phrases (NP)

- NP refer to objects, places, concepts, events, qualities, etc.
- Head of NP is usually a common noun
- In some cases, NP can have a pronoun as the head

It hid under the rug
Once I opened the door, I regretted *it* for months

- In some cases, NP can have proper noun or count noun or mass nouns as head

John at food
Dogs are friendly
Water is necessary for life

Noun Phrases (NP)

- NP may contain ***Specifiers*** and ***Qualifiers*** preceding the head
- Qualifiers describe the general class of objects identified by the head
- Specifiers indicate how many such objects are being described

Specifier Qualifier Head

the ceiling paint can

the angry bird

Noun Phrases (NP)

- Specifiers constructed out of:
  - Ordinals (*first, second*, etc.)
  - Cardinals (*one, two*, etc.)
  - Determiners
    - Articles (*the, a*, and *an)*
    - Demonstratives (*this, that, these, those*, etc.)
    - Possessives (John*'s*, the fat man*'s*)
    - Wh-Determiners (question related word e.g. *which* and *what*)
    - Quantifying Determiners (*some, every, most, any, half*, etc.)

the first three contestants

Sentence Mood

- The way a sentence is used is called as its “mood”
- A sentence can have 4 moods:

• Declarative (or Assertive)	*The dog is running*
• Yes/No question	*Is the dog running?*
• Wh-Question	*Which dog is running?*
• Imperative (or Command)	*Feed the dog!*

Most simple and common sentence structure

- Subject-Verb-Object (SVO) (Also called as Subject-Predicate-Object)

John cooked pasta

John cooked pasta in a pan

- In other languages VSO, SOV
  - SOV (Hindi, Japanese)
  - VSO (Arabic)
- Some languages are free order (SVO, SOV, VSO are valid)
  - Sanskrit, Greek

Verb Phrase (VP)

- VP refer to some action, activity or an event
- VP consists of a verb and other constituents such as NP, PP, ADVP
- E.g. a declarative sentence can have NP (subject) followed by a VP (predicate)
- Verbs can be divided into different classes:
  - Auxiliary Verbs — *be, do, have*
  - Modal Verbs — *will, can, could*
  - Main Verbs — *eat, run, walk*

Verbs

- Verbs can have different forms: base, simple present, simple past, present participle, etc.

Form	Examples	Example Uses
base	hit, cry, go,be	*Hit* the ball! I want to *go.*
simple present	hit, cries, go, am	The dog *cries* every day. I *am* thirsty.
simple past	hit, cried, went, was	I *was* thirsty. I *went* to the movie store.
present participle	hitting, crying, going, being	I'm *going* to the store. *Being* the last in line aggravates me.
past participle	hit, cried, gone, been	I've *been* there before. The cake was *gone*.

Chapter 2 of [4]

Verbs

- Verbs can have different tenses: simple present, simple past, present perfect, simple future etc.

Tense	The Verb Sequence	Example
simple present	simple present	He walks to the store.
simple past	simple past	He walked to the store.
simple future	*will* + infinitive	He will walk to the store.
present perfect	*have* in present + past participle	He has walked to the store.
future perfect	*will* + *have* in infinitive + past participle	I will have waled to the store.
past perfect (or pluperfect)	*have* in past + past participle	I had walked to the store.

Chapter 2 of [4]

Verbs

- Verbs can be in progressive tense.

Tense	Structure	Example
present progressive	*be* in present + present participle	He is walking.
past progressive	*be* in past + present participle	He was walking.
future progressive	*will* + *be* in infinitive + present participle	He will be walking.
present perfect progressive	*have* in present + *be* in past participle + present participle	He has been walking.
future perfect progressive	*will* + *have* in present + *be* as past participle + present participle	He will have been walking.
past perfect progressive	*have* in past + *be* in past participle + present participle	He had been walking.

Verb Transitivity

- In a VP, depending on the verb, different complements are possible.
- Intransitive Verbs: No complement required

John has been running
John laughed

- Transitive Verbs: NP follows the verb

John found a key
John ran the machine
John ran the machine with hands

- Some verbs are only transitive, and some can be either e.g. *run*

Active vs Passive

- Transitive verbs allow *Passive Constructions*

- NP is used at the object position instead of the usual subject position

The ball was thrown by Jack

vs

Jack threw the ball

Particles

- Overlap with the class of prepositions *up, out, over, in, down*
- Assists verbs, to give a new meaning *look up, look out, look over*
- Can lead to ambiguity *Look over the newspaper*
- Verb-Particle sentence: the pronoun must precede the particle
- Verb-Prepositional sentence: pronoun follows the preposition

I looked it up

vs

I looked up it

Clausal Complements (CC)

- Clause: part of sentence that contains a subject and a verb
- Many VPs come with clausal complements
- S *that* [CC]

Sam knows that Jack ate the pizza

Sam knows that the pizza was eaten by Jack

- VP [inf] [CC] and S [inf] [CC]

Jack wishes to eat the pizza

Jack wishes for Sam to eat the pizza

- S [WH] : who, where, what, why

Sam knows why Jack ate the pizza

Sam knows who ate the pizza

Prepositional Phrase Complements (PP)

- Many VPs come with PP complements
- The structure is verb specific

Jack gave the book to the library
**Jack gave the book from the library*

Jack put the book in the box
Jack put the book inside the box
Jack put the book by the box

- *give:* VP + NP + PP[to]
- *decide:* VP + NP + PP[about]
- *blame:* VP + NP + PP[on]

Complements of a VP

Verb	Complement Structure	Example
laugh	Empty (intransitive)	Jack Laughed.
find	NP (transitive)	Jack found a key.
give	NP+NP (bitransitive)	Jack gave Sue the paper.
give	NP+PP[to]	Jack gave the book to the library.
reside	Location phrase	Jack resides in Rochester.
put	NP + Location phrase	Jack put the book inside.
speak	PP[with] + PP[about]	Jack spoke with Sue about the book.
try	VP[to]	Jack tried to apologize.
tell	NP + VP[to]	Jack told the man to go.
wish	S[to]	Jack wished for the man to go.
keep	VP[ing]	Jack keeps hoping for the best.
catch	NP + VP[ing]	Jack caught Sam looking in his desk.
watch	NP + VP[base]	Jack watched Sam eat a pizza.
regret	S[that]	Jack regretted that he'd eaten the whole thing.
tell	NP + S[that]	Jack told Sue that he was sorry.
seem	ADJP	Jack seems unhappy in his new job.
think	NP + ADJP	Jack thinks Sue is happy in her job.
know	S[WH]	Jack knows where the money is.

Chapter 2 of [4]

VPs are very important

- Every sentence needs a verb
- Lot of work on verb-argument structure
- These capture the semantics of the text

VPs are very important

- Every sentence needs a verb
- Lot of work on verb-argument structure
- These captur

https://verbs.colorado.edu/verbnet/

VPs are very important

- Every sentence needs a verb

https://verbs.colorado.edu/verbnet/

VPs are very important

- Ever

Proposition Bank

- **The original PropBank**

The original PropBank project, funded by ACE, created a corpus of text annotated with information about basic semantic propositions. Predicate-argument relations were added to the syntactic trees of the Penn Treebank. This resource is now available via LDC.

- **PropBank today**

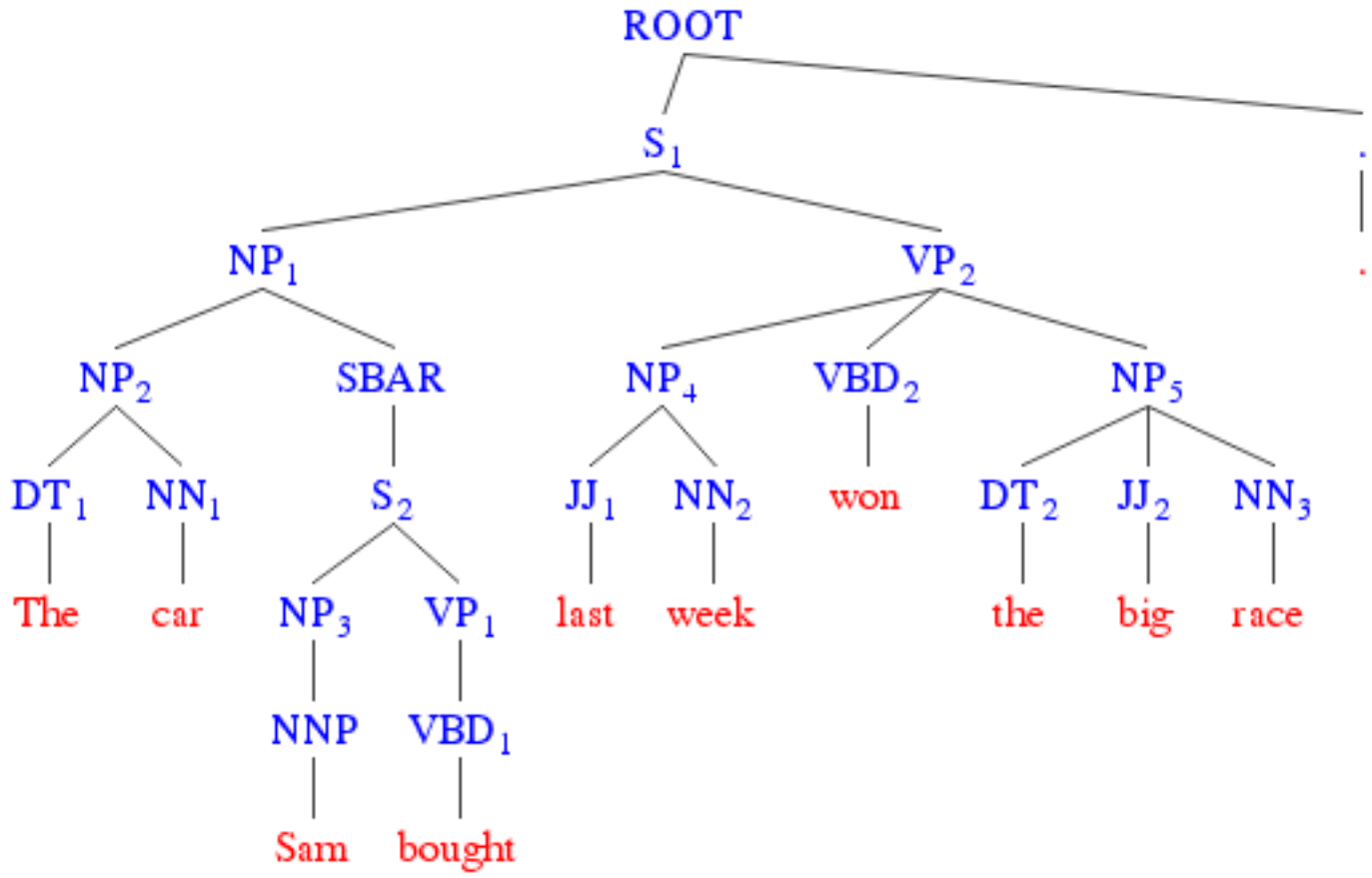
This project was continued under NSF funding and DARPA GALE and BOLT. with the aim of creating Parallel PropBanks (the English-Chinese Treebank/PropBank) and also PropBanking other genres, such as Broadcast News, Broadcast Conversation, WebText and Discussion Fora, at the University of Colorado. PropBank is also being mapped to VerbNet and FrameNet as part of SemLink: Mapping together PropBank/VerbNet/FrameNet. PropBank's coverage is also being extended to provide support for AMR annotation, which makes heavy use of PropBank frame files. This is being funded by DARPA DEFT.

Resources - The Resources below are being transitioned to a New PropBank Github Resource Page

- Martha Palmer, Dan Gildea, Paul Kingsbury, The Proposition Bank: A Corpus Annotated with Semantic Roles *Computational Linguistics Journal, 31:1, 2005.*
- Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.
- The Necessity of Parsing for Predicate Argument Recognition. Daniel Gildea and Martha Palmer. 2002. In Proceedings of ACL 2002, Philadelphia, PA.
- OLD 2005 Annotation guidelines for PropBank
- OLD 2012 Annotation guidelines for English PropBank

VerbNet (VN) (Kipper-Schuler 2006) is the larges
syntactic and semantic patterns. It is a hierarchical, domain-indepe
lexicon with mappings to other lexical resources, such as WordNet (Miller, 1990; Fellbaum,
PropBank (Kingsbury and Palmer, 2002), and FrameNet (Baker et al., 1998). VerbNet is organized into verb classes extending Levin (1993) classes through refinement and addition of subclasses to achieve syntactic and semantic coherence among members of a class. Each verb class in VN is completely described by thematic roles, selectional preferences of the arguments, and frames consisting of a syntactic description and a semantic representation with subevent structure

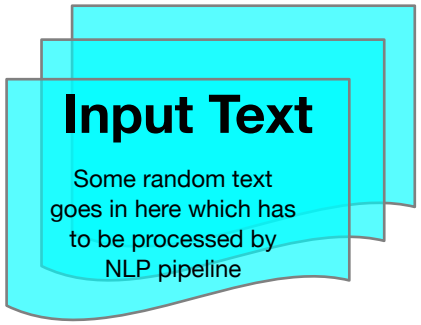
Phrase Structure Tree (Constituent Tree)



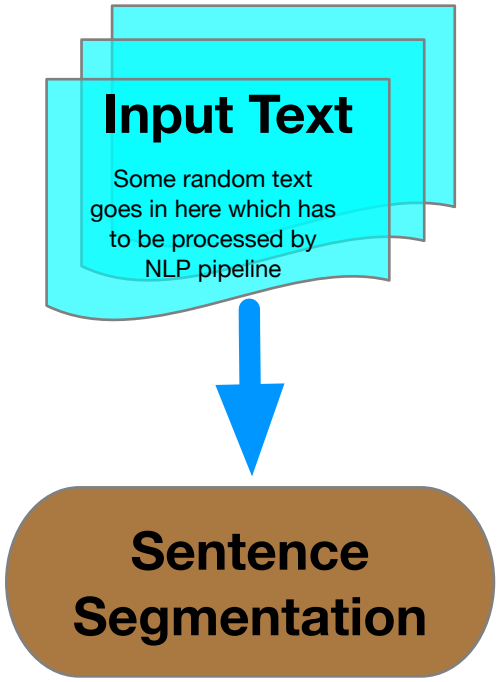
References

1. Chapter 2, Natural Language Understanding, James Allen
2. Chapter 12, Speech and Language Processing, Dan Jurafsky and James Martin

NLP PIPELINE



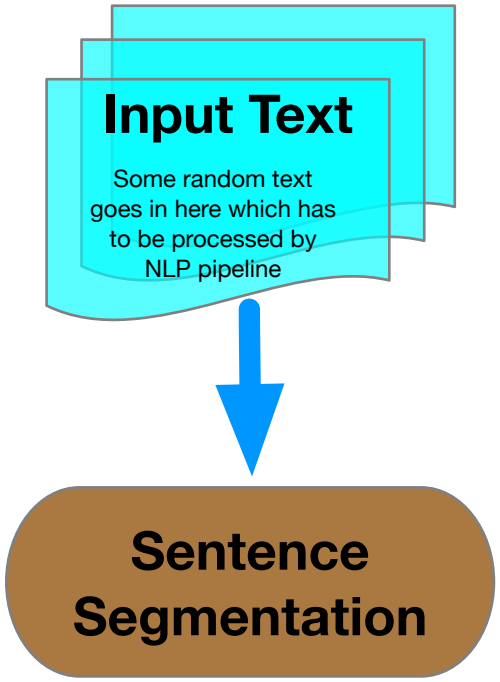
Nepal is a landlocked country in South Asia. It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain. It has a diverse geography, including fertile plains, subalpine forested hills, and eight of the world's ten tallest mountains.



Nepal is a landlocked country in South Asia.

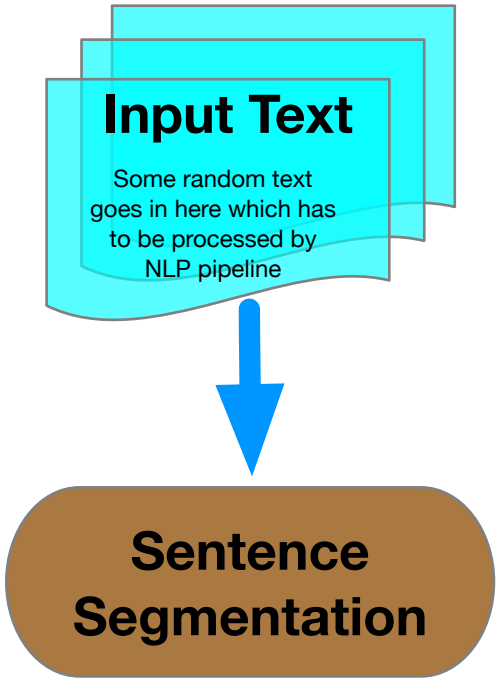
It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain.

It has a diverse geography, including fertile plains, subalpine forested hills, and eight of the world's ten tallest mountains.



Nepal has an estimated population of 26.4 million, it is 48th largest country by population and 93rd largest country by area. *Nepal is a landlocked country in South Asia. It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain. It has a diverse geography, including fertile plains, subalpine forested hills, and eight of the world's ten tallest mountains.*

What about this?

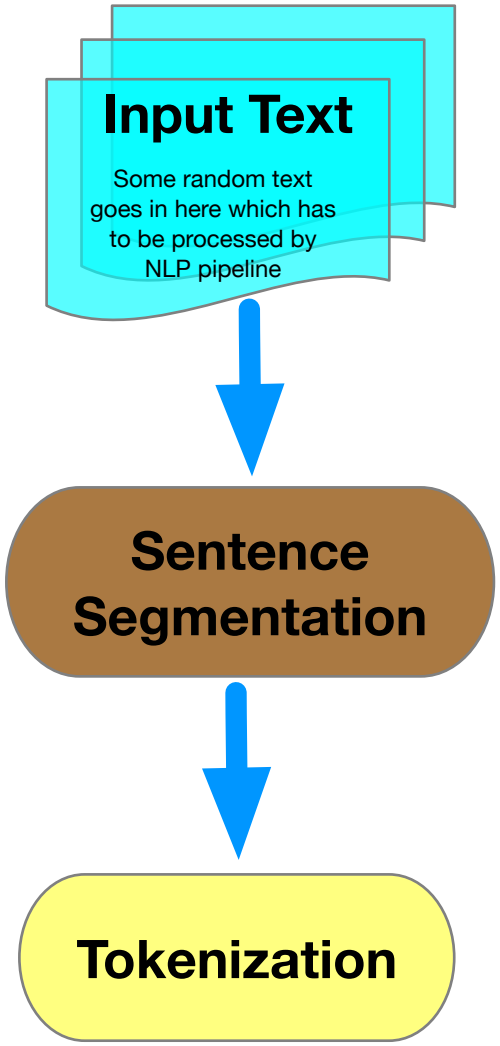


Nepal has an estimated population of 26.4 million, it is 48th largest country by population and 93rd largest country by area.

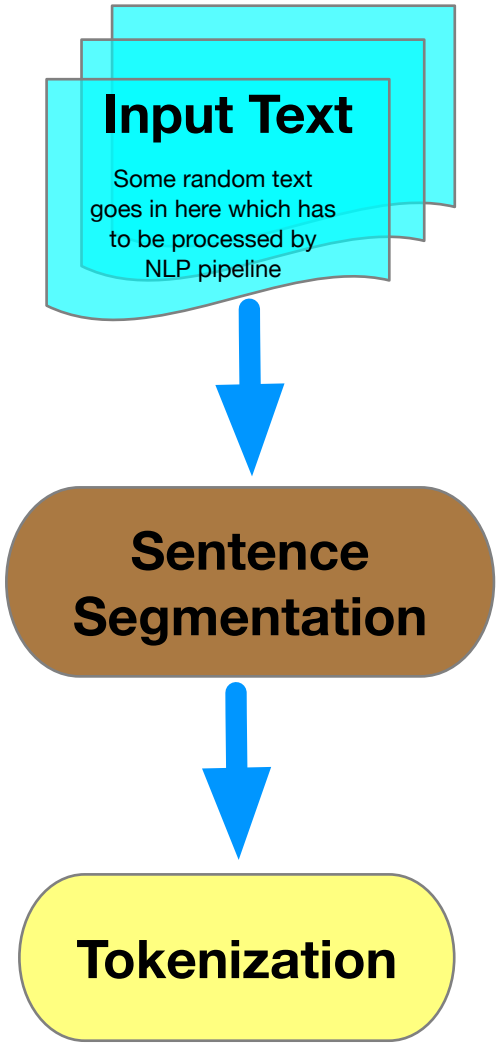
Nepal is a landlocked country in South Asia.

It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain.

Nepal has a diverse geography, including fertile plains, subalpine forested hills, and eight of the world's ten tallest mountains.



Split the sentence into tokens (roughly same as words).

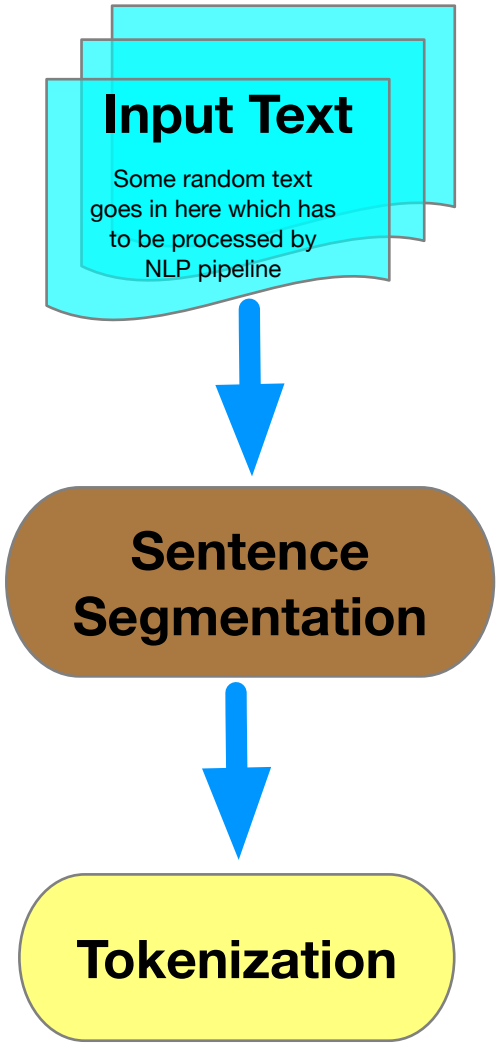


[Nepal] [has] [an] [estimated] [population] [of] [26.4] [million] [,] [it] [is] [48th] [largest] [country] [by] [population] [and] [93rd] [largest] [country] [by] [area] [.]

[Nepal] [is] [a] [landlocked] [country] [in] [South] [Asia] [.]

[It] [is] [located] [mainly] [in] [the] [Himalayas] [,] [but] [also] [includes] [parts] [of] [the] [Indo] [-] [Gangetic] [Plain] [.]

[Nepal] [has] [a] [diverse] [geography] [,] [including] [fertile] [plains] [,] [subalpine] [forested] [hills] [,] [and] [eight] [of] [the] [world] [’s] [ten] [tallest] [mountains] [.]

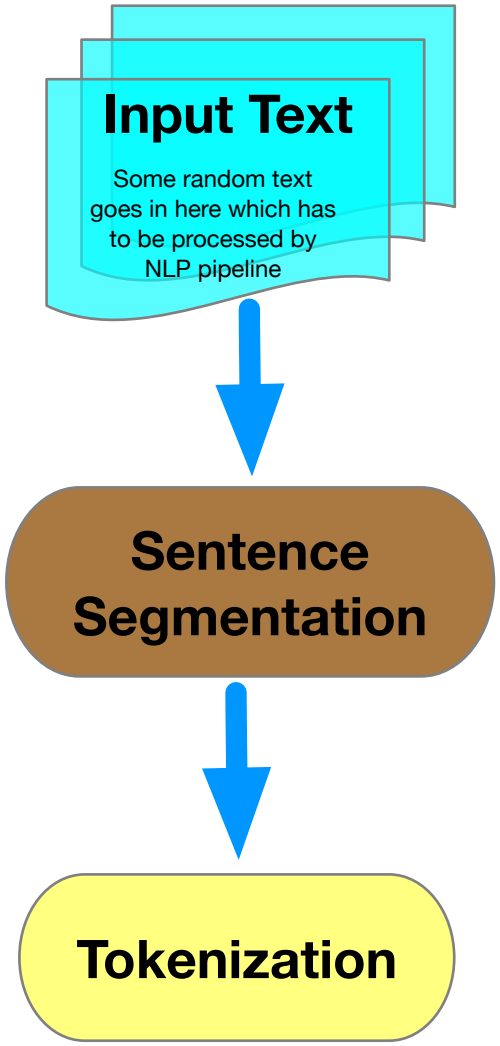


[Nepal] [has] [an] [estimated] [population] [of] [26.4] [million] [,] [it] [is] [48th] [largest] [country] [by] [population] [and] [93rd] [largest] [country] [by] [area] [.]

[Nepal] [is] [a] [landlocked] [country] [in] [South] [Asia] [.]

*[It] [is] [located] [mainly] [in] [the] [Himalayas] [,] [but] [also] [includes] [parts] [of] [the] [**Indo**] [-] [**Gangetic**] [Plain] [.]*

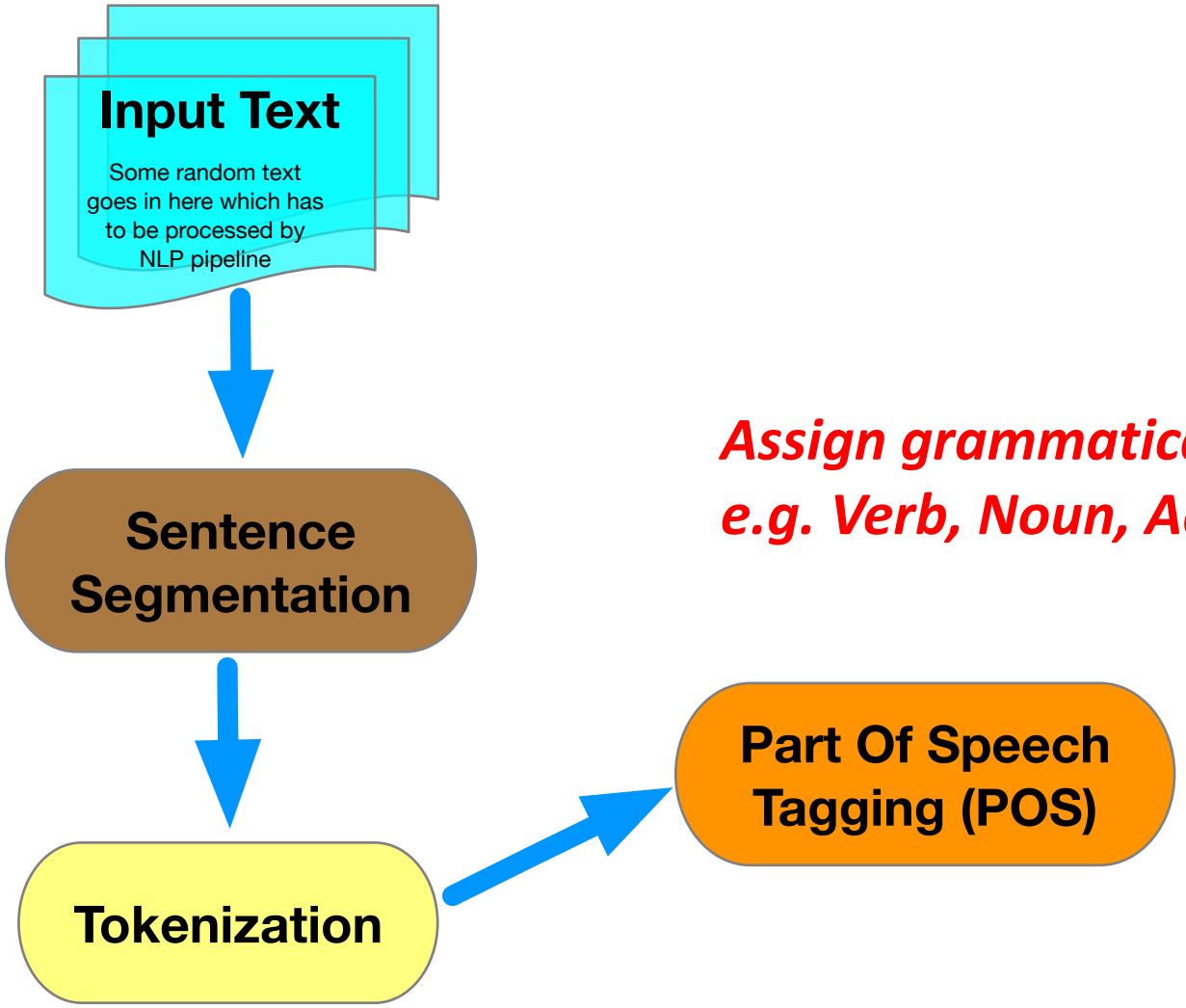
*[Nepal] [has] [a] [diverse] [geography] [,] [including] [fertile] [plains] [,] [subalpine] [forested] [hills] [,] [and] [eight] [of] [the] [world] [**'s**] [ten] [tallest] [mountains] [.]*



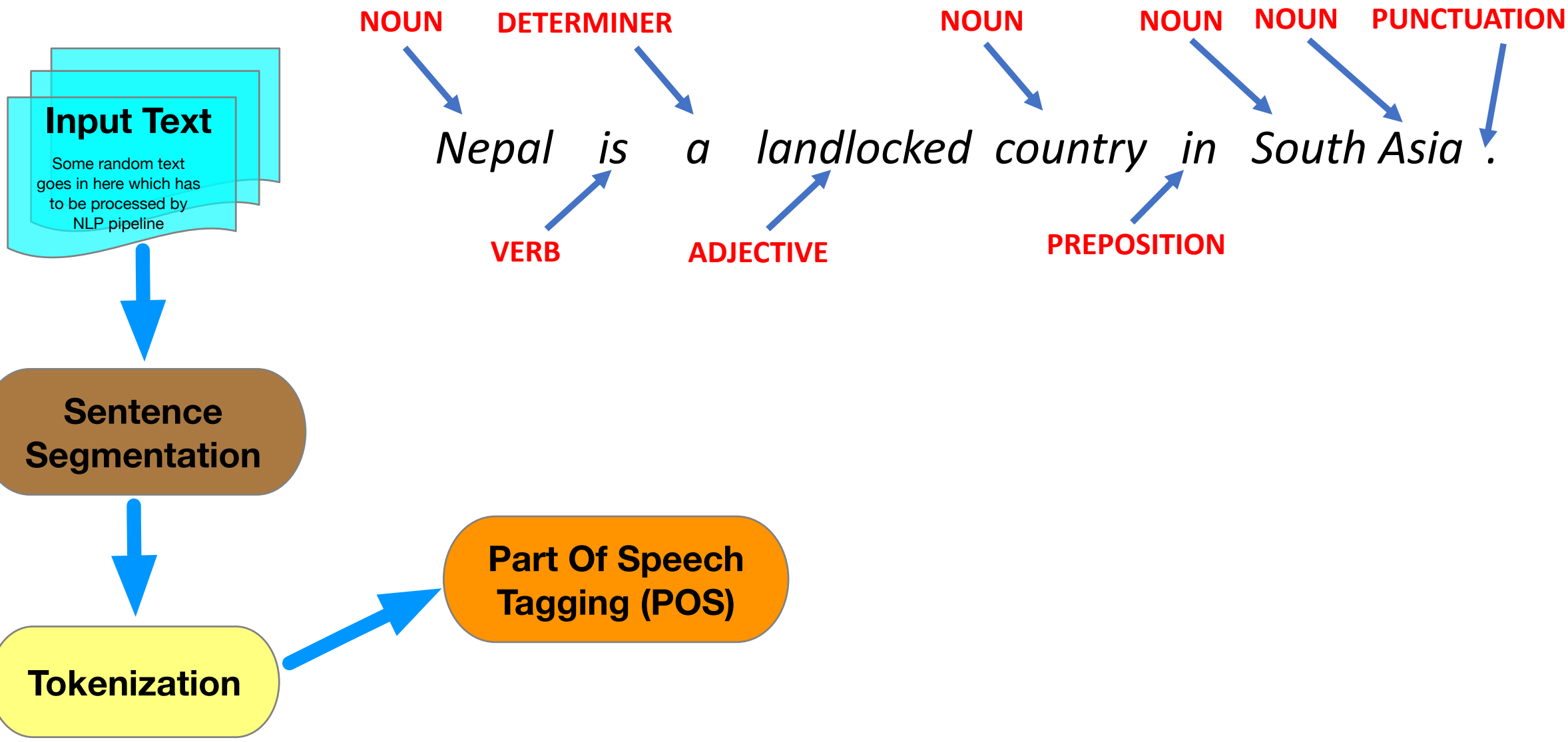
Longest word in Sanskrit!!

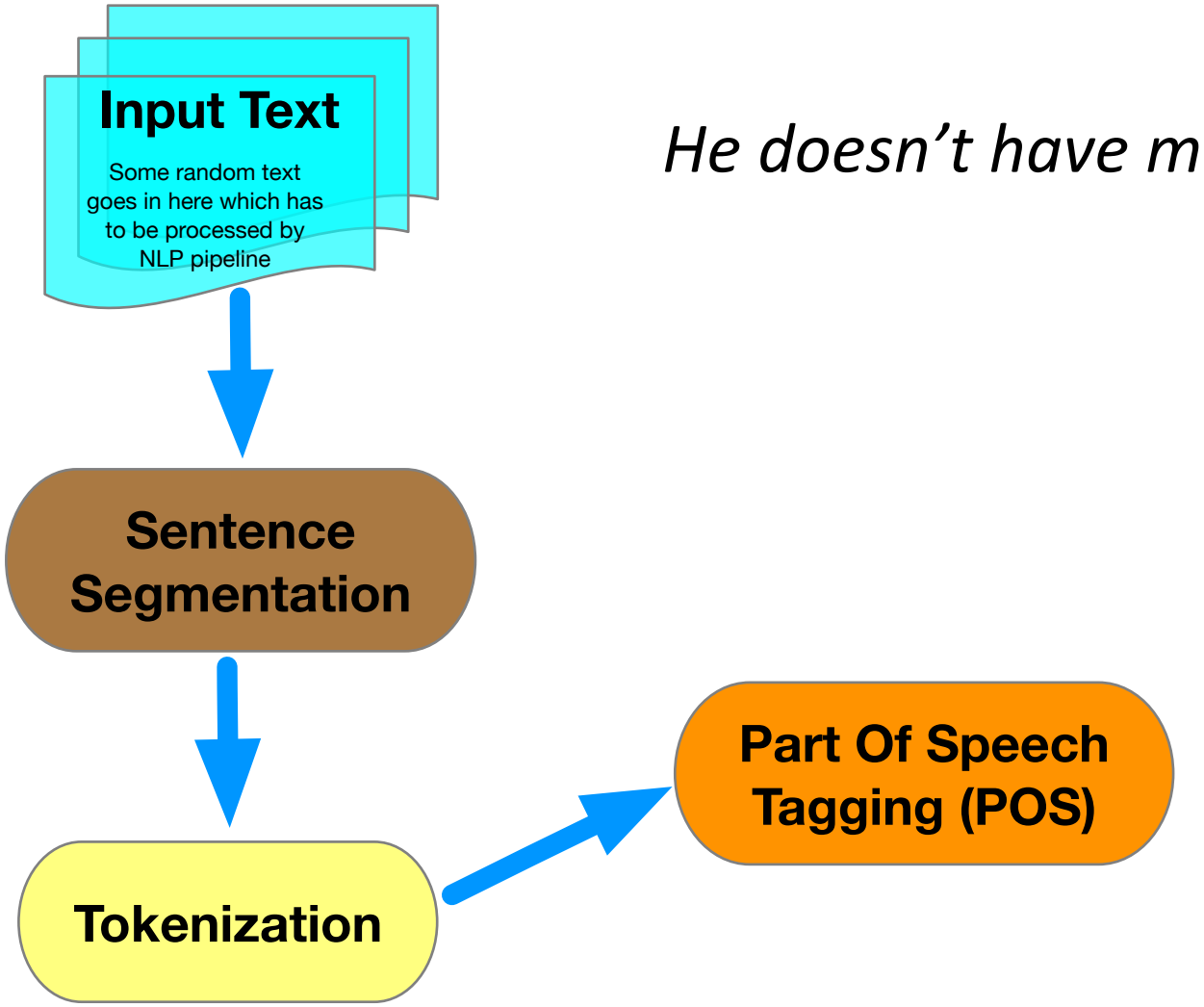
निरन्तरान्धकारित-दिगन्तर-कन्दलदमन्द-सुधारस-बिन्दु-सान्द्रतर-घनाघन-वृन्द-सन्देहकर-स्यन्दमान-मकरन्द-बिन्दु-बन्धुरतर-माकन्द-तरु-कुल-तल्प-कल्प-मृदुल-सिकता-जाल-जटिल-मूल-तल-मरुवक-मिलदलघु-लघु-लय-कलित-रमणीय-पानीय-शालिका-बालिका-करार-विन्द-गलन्तिका-गलदेला-लवङ्ग-पाटल-घनसार-कस्तूरिकातिसौरभ-मेदुर-लघुतर-मधुर-शीतलतर-सलिलधारा-निराकरिष्णु-तदीय-विमल-विलोचन-मयूख-रेखापसारित-पिपासायास-पथिक-लोकान्

Challenge: How do you tokenize long compound words?

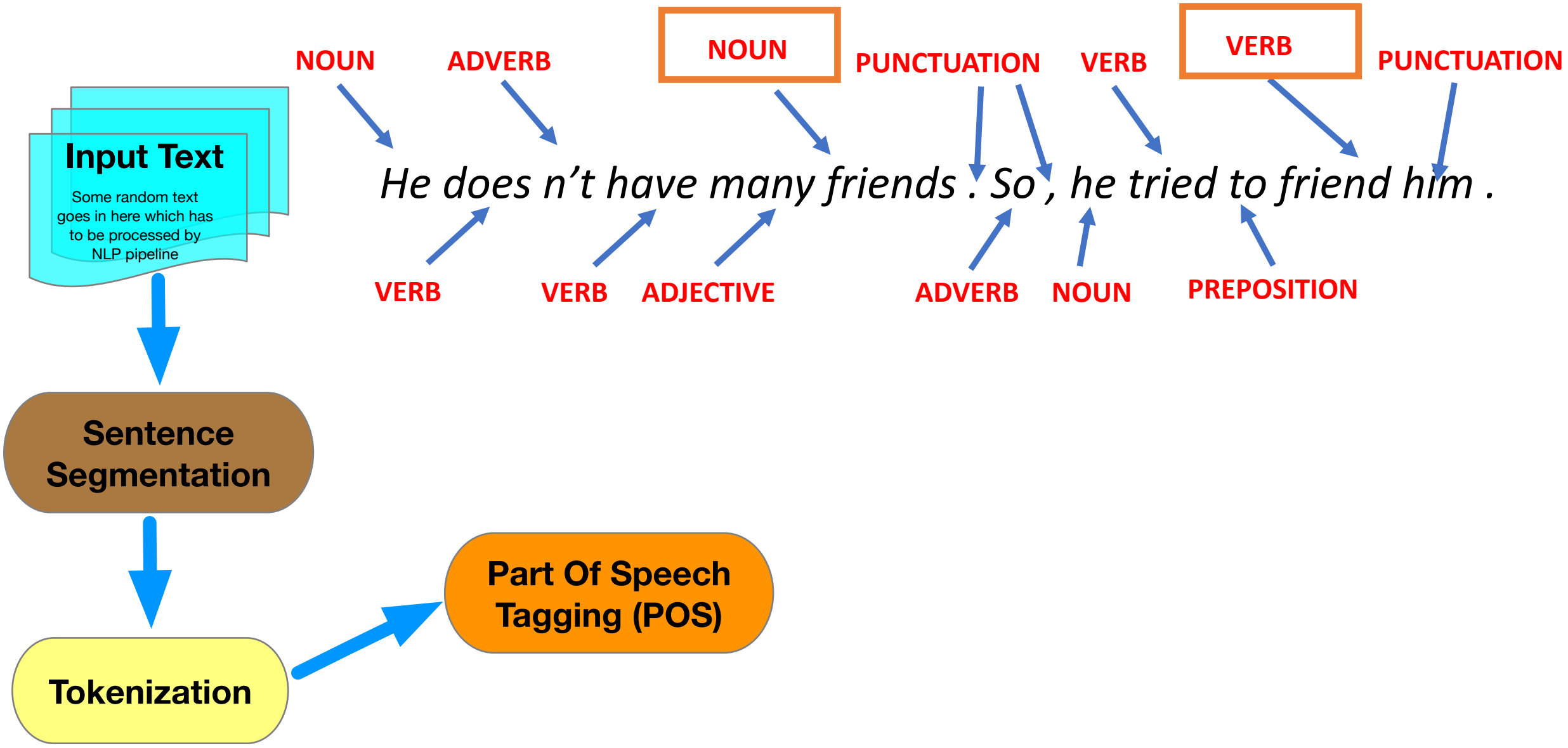


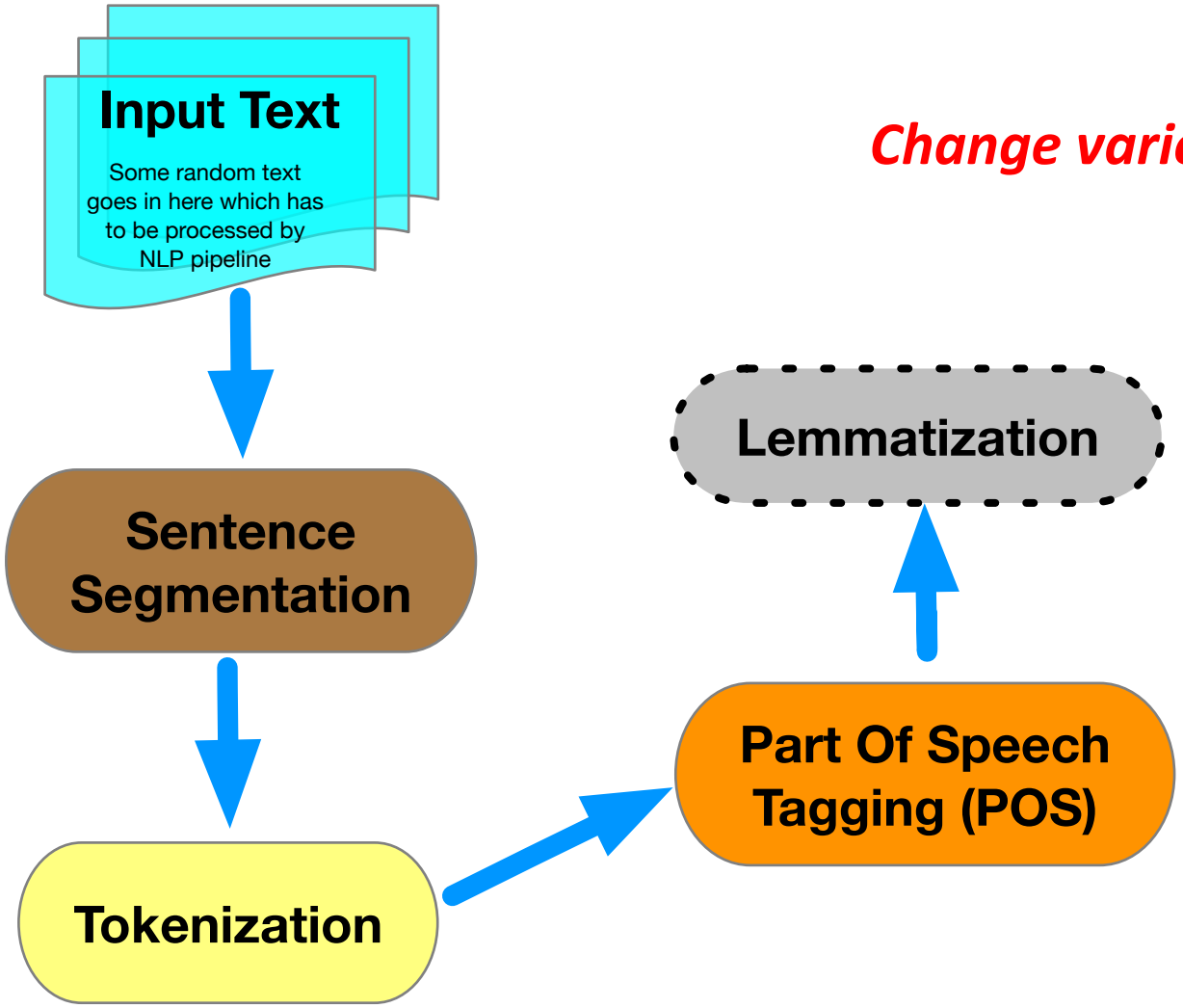
Assign grammatical categories (Part of Speech) to each token.
e.g. Verb, Noun, Adjective, Adverb, etc.


NOUN
DETERMINER
NOUN
NOUN
NOUN
PUNCTUATION
Nepal is a landlocked country in South Asia .
VERB
ADJECTIVE
PREPOSITION
Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)



He doesn't have many friends . So, he tried to friend him.


NOUN
ADVERB
NOUN
PUNCTUATION
VERB
VERB
PUNCTUATION
He does n't have many friends . So , he tried to friend him .
VERB
VERB
ADJECTIVE
ADVERB
NOUN
PREPOSITION
Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)

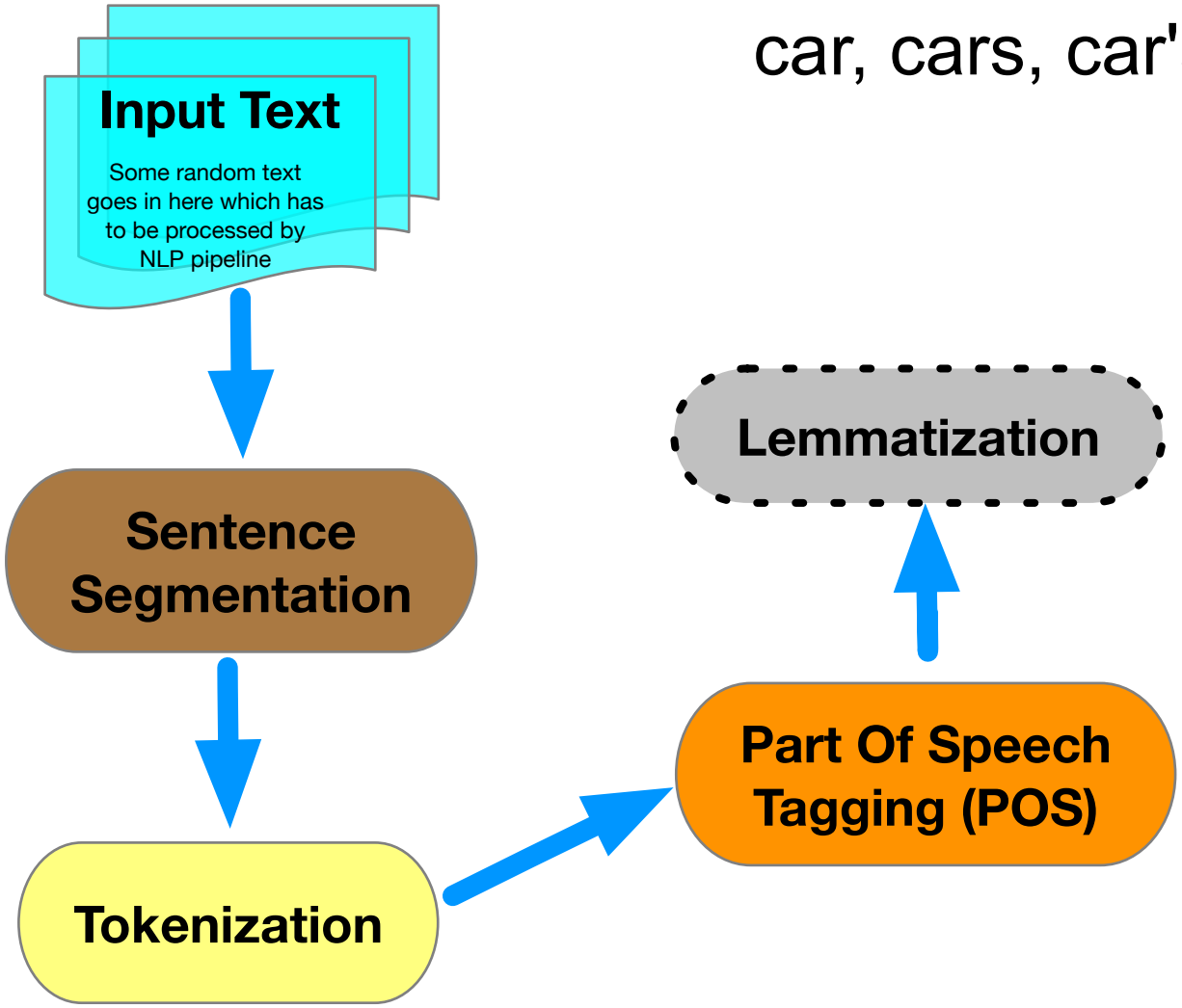


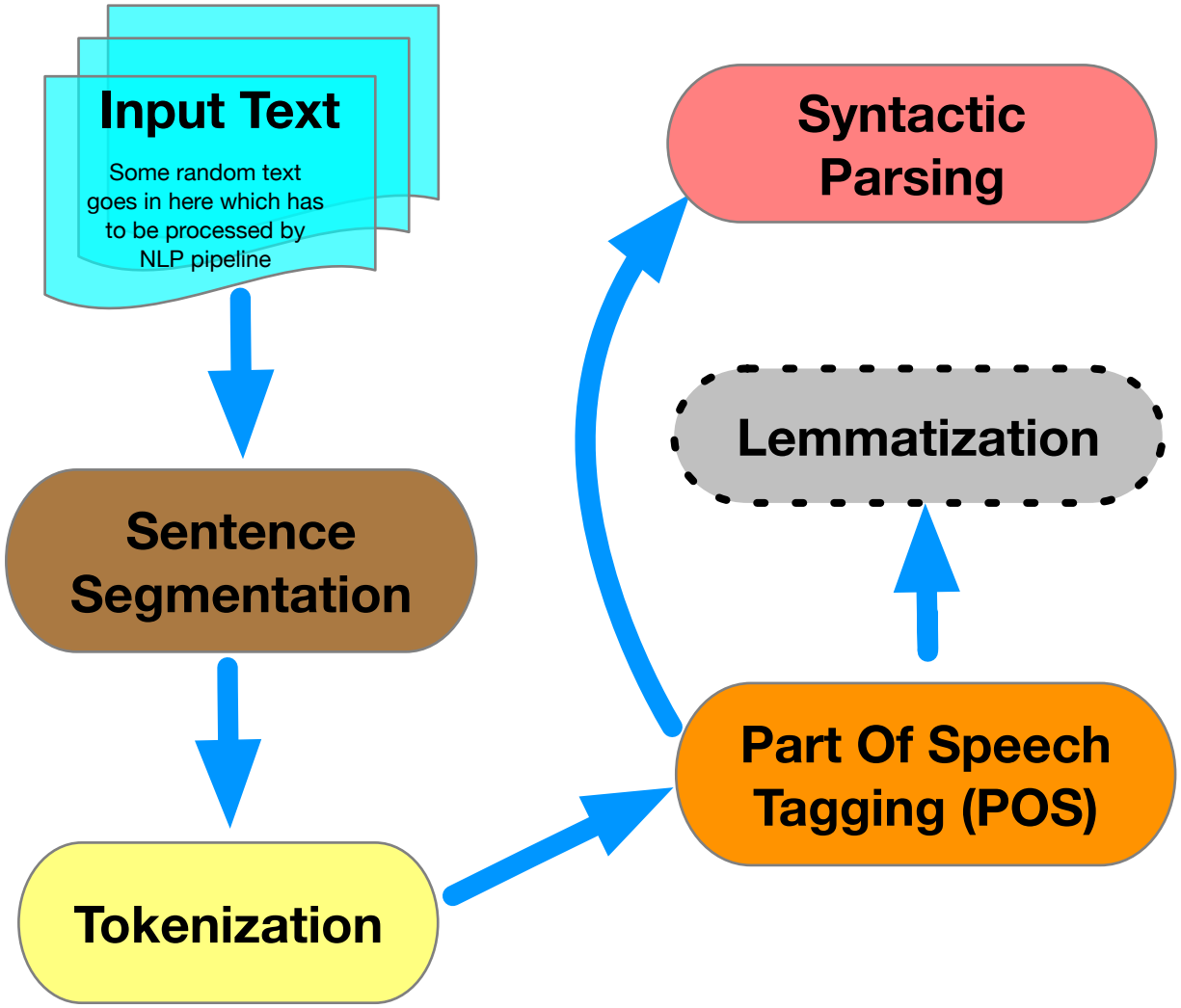
Change various forms of a word to its base form

LEMMA

am, are, is ⇒ be

car, cars, car's, cars' ⇒ car

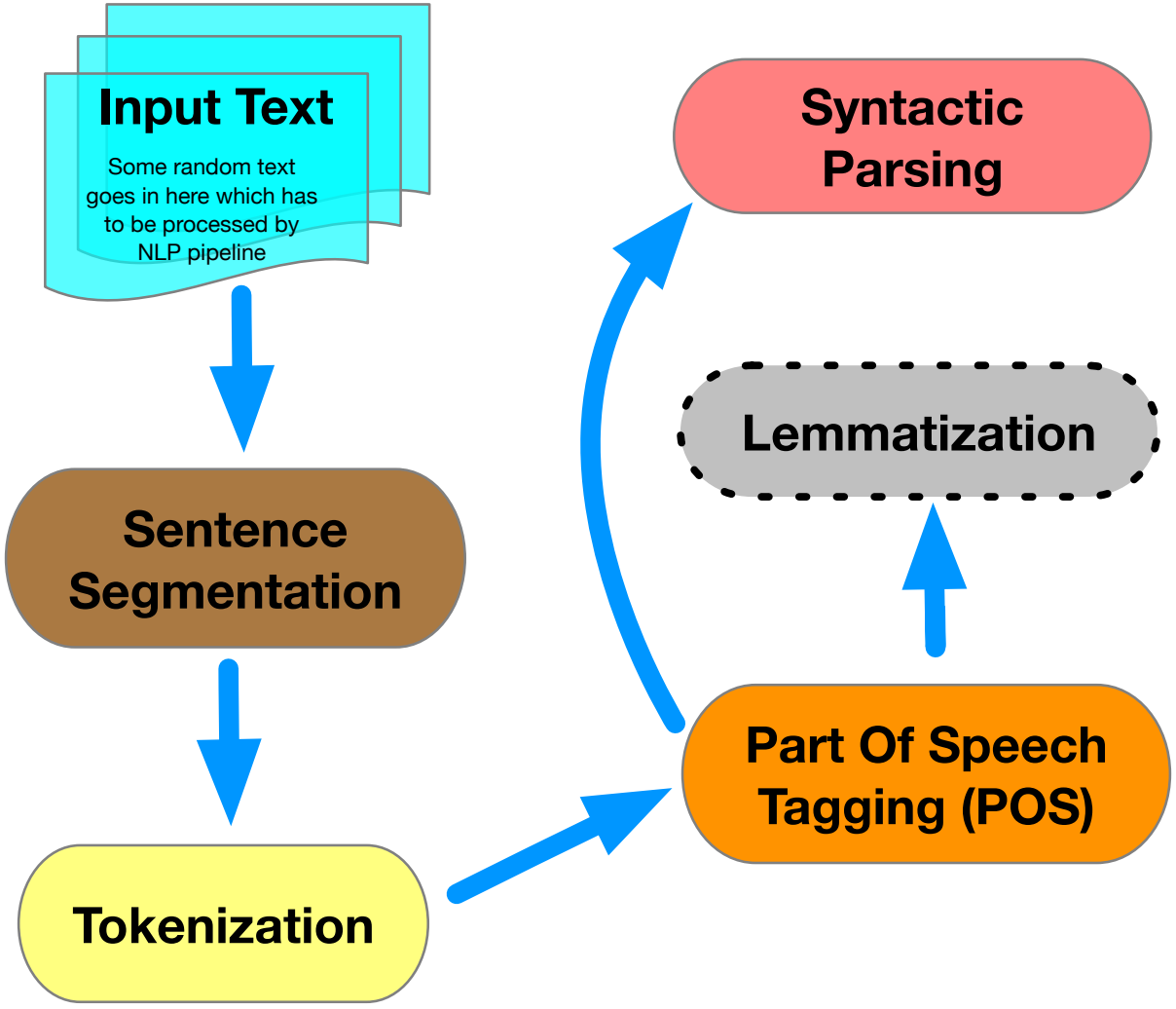


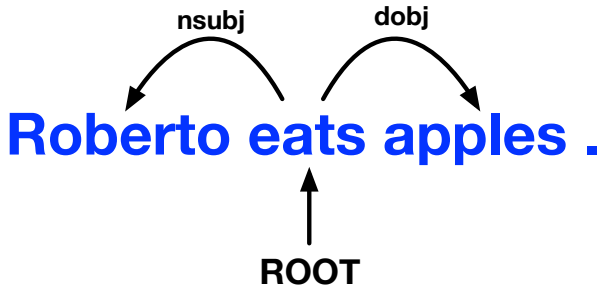


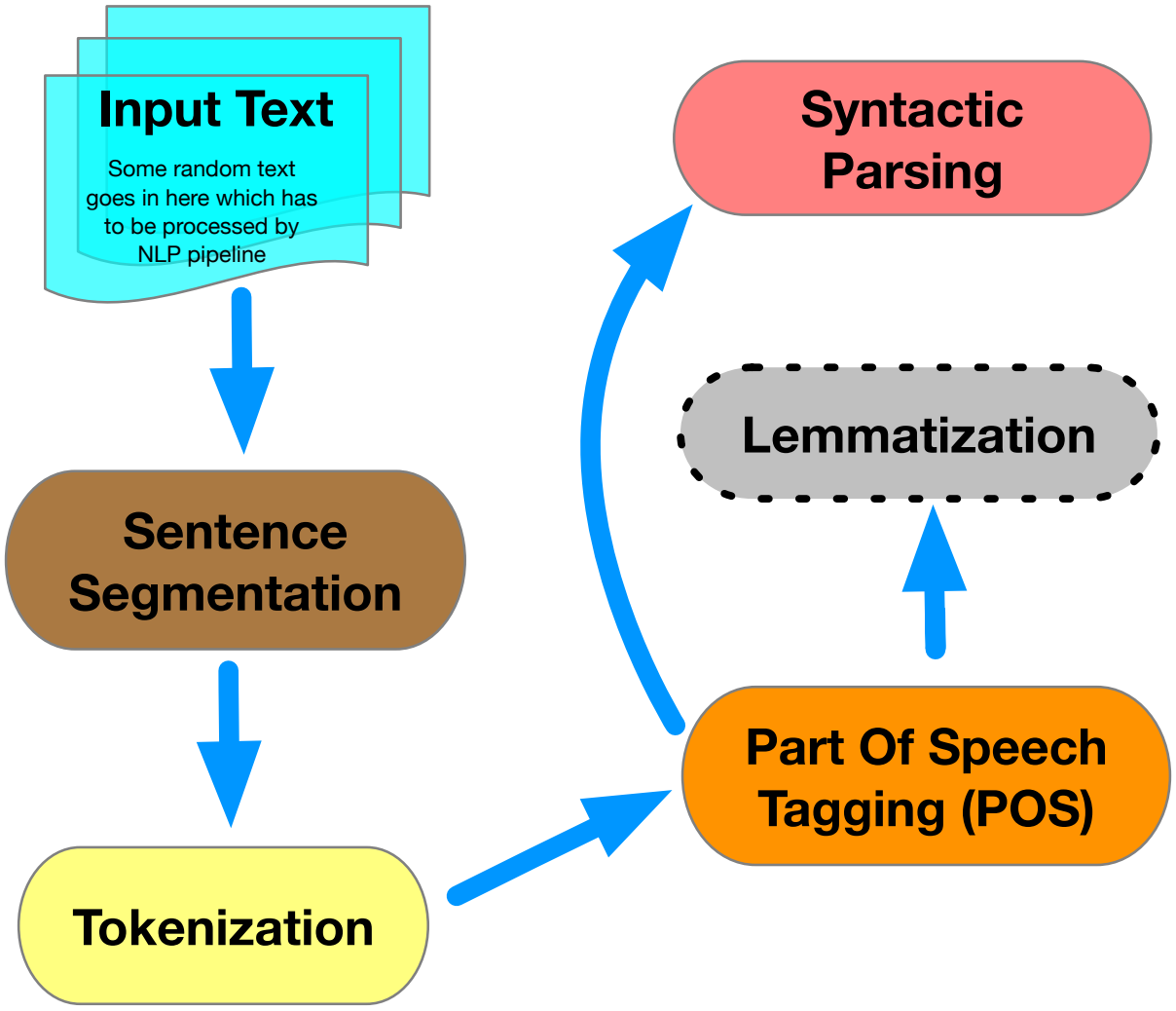
Sentences in a language follow the grammar which is formally specified via the Syntax (rules of the language)

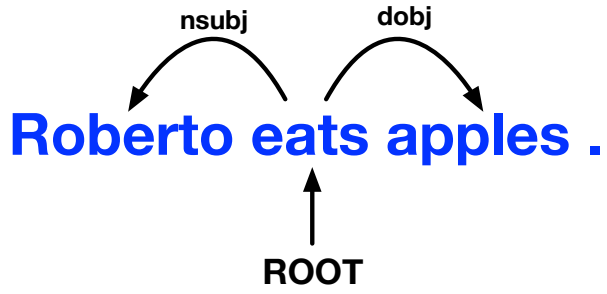
Parsing is the process for syntactically analyzing the sentences

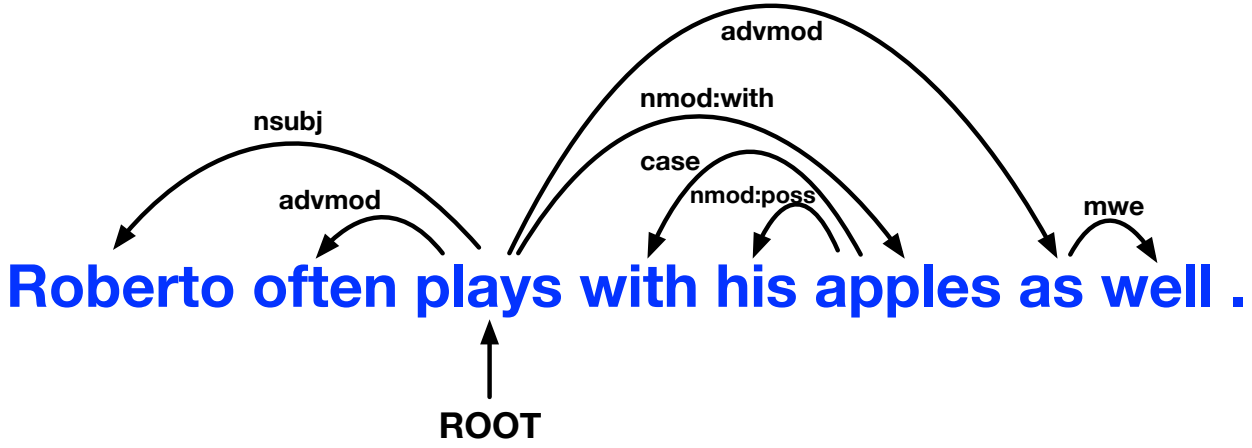
One of the popular Parsing method is Dependency Parsing

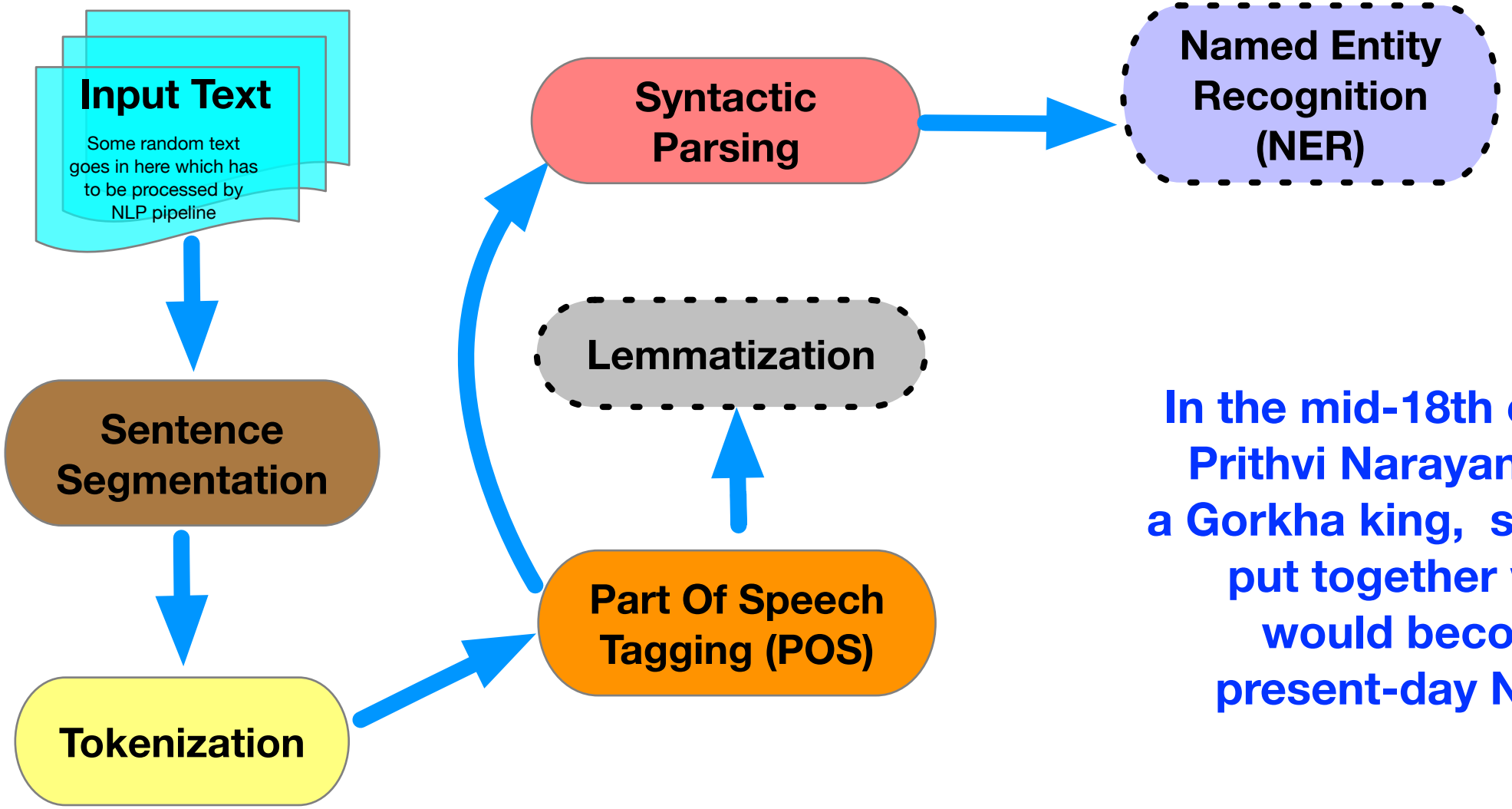

Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)
Lemmatization
Syntactic Parsing


nsubj
dobj
Roberto eats apples .
ROOT

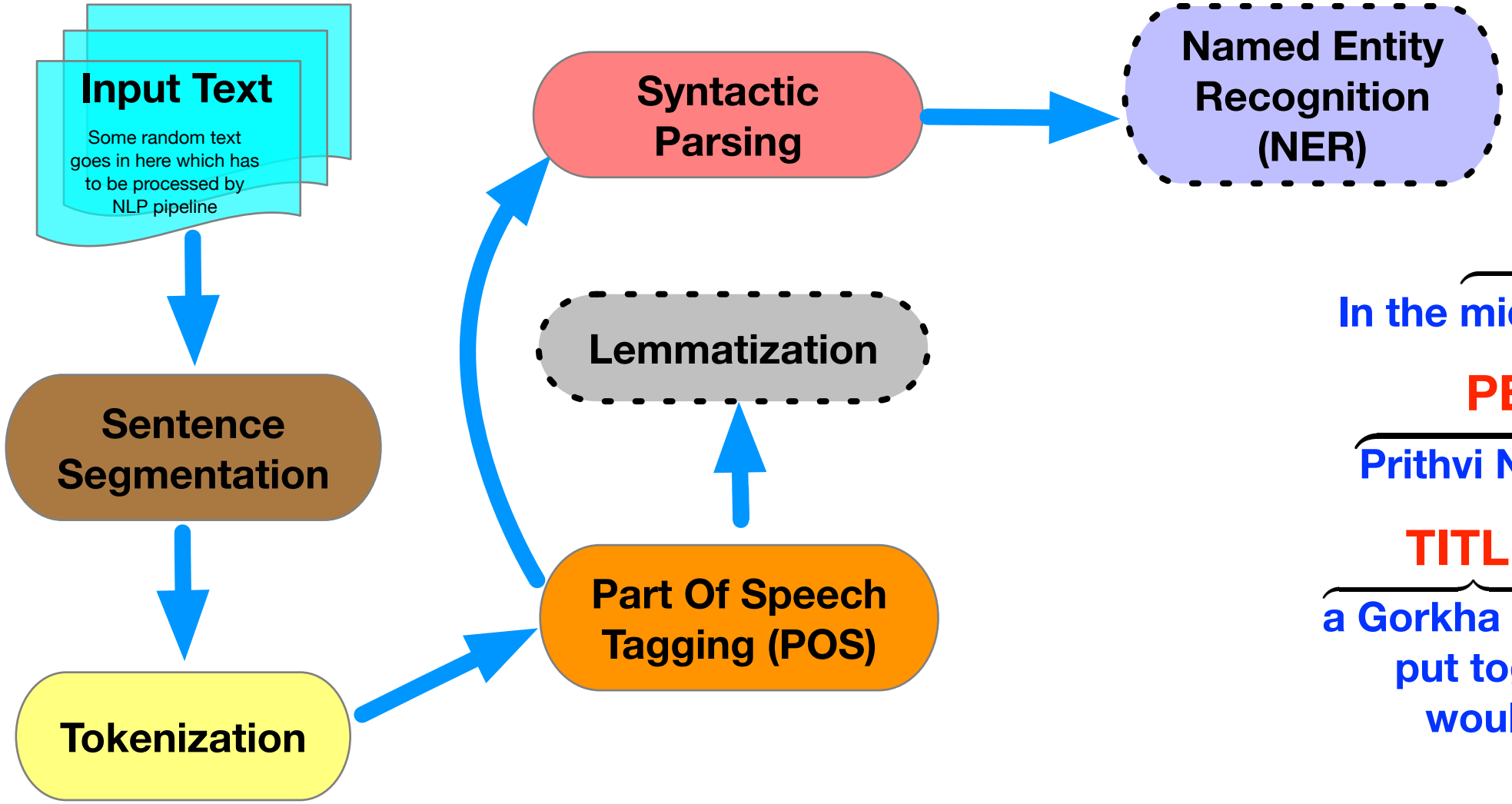

Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)
Lemmatization
Syntactic Parsing

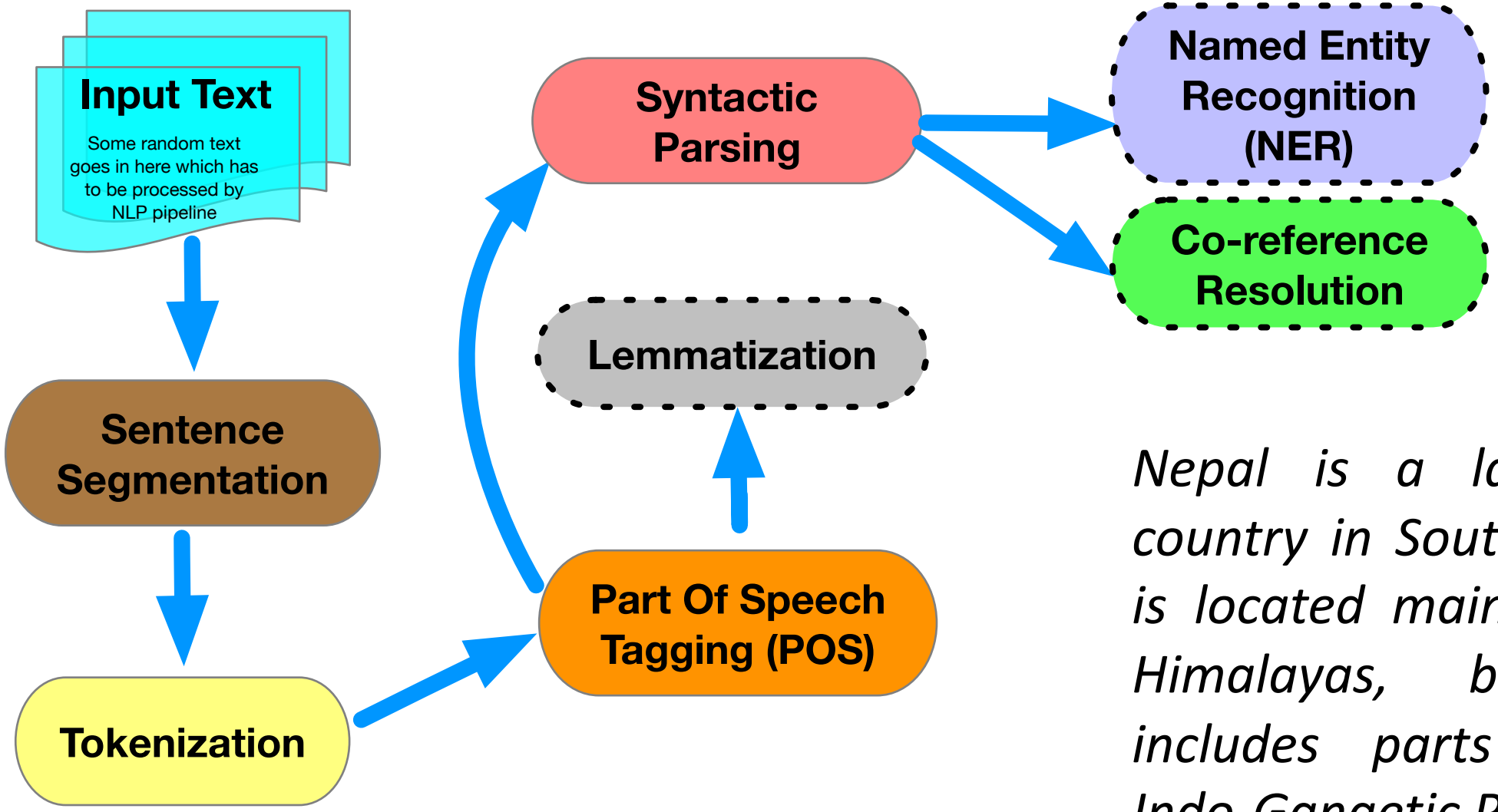

nsubj
dobj
Roberto eats apples .
ROOT


advmod
nmod:with
nsubj
case
advmod
nmod:poss
mwe
Roberto often plays with his apples as well .
ROOT

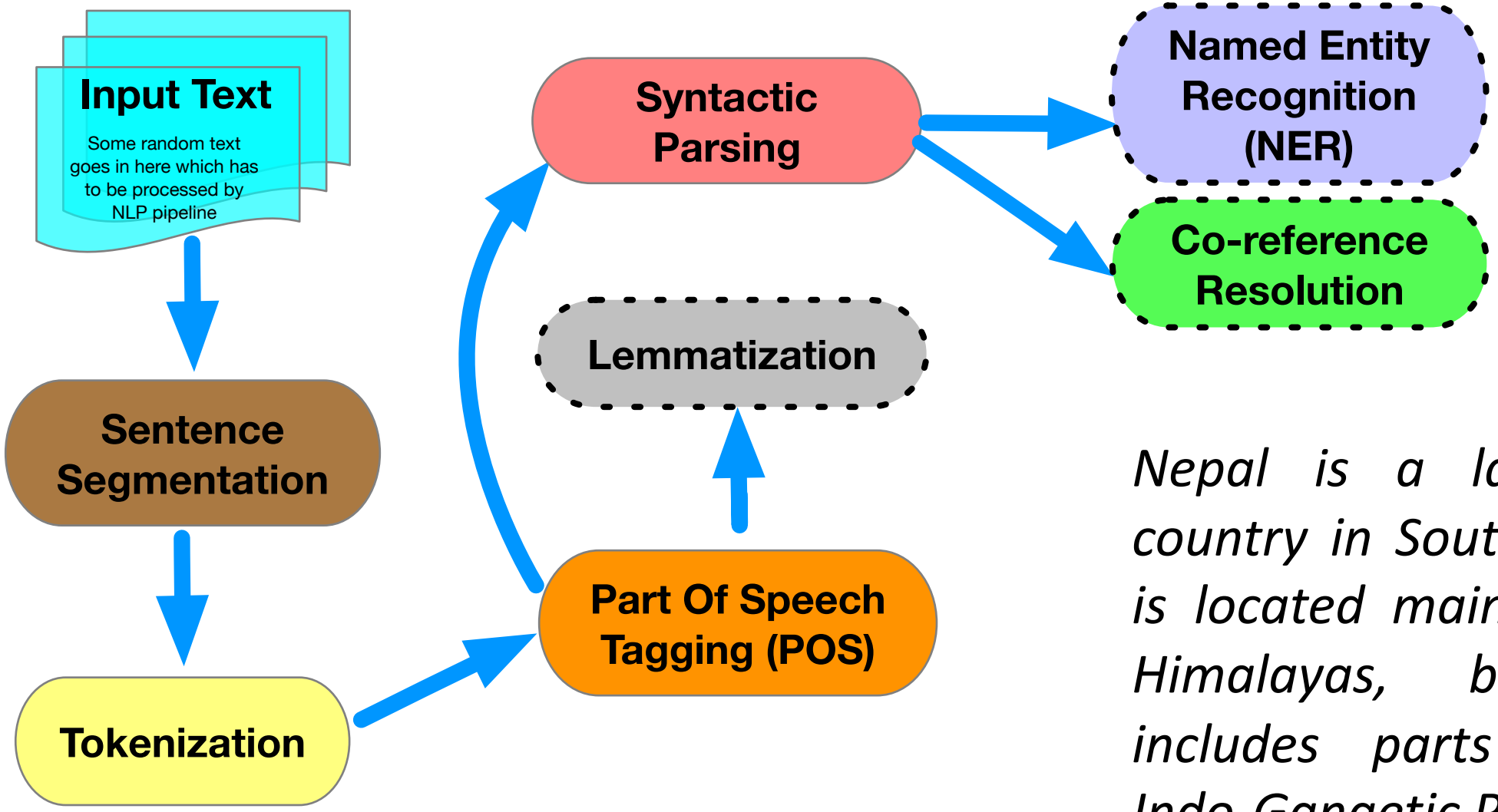


In the mid-18th century, Prithvi Narayan Shah, a Gorkha king, set out to put together what would become present-day Nepal.


Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)
Lemmatization
Syntactic Parsing
Named Entity Recognition (NER)
DATE
In the mid-18th century,
PERSON
Prithvi Narayan Shah,
TITLE
a Gorkha king, set out to put together what would become
LOCATION
present-day Nepal.

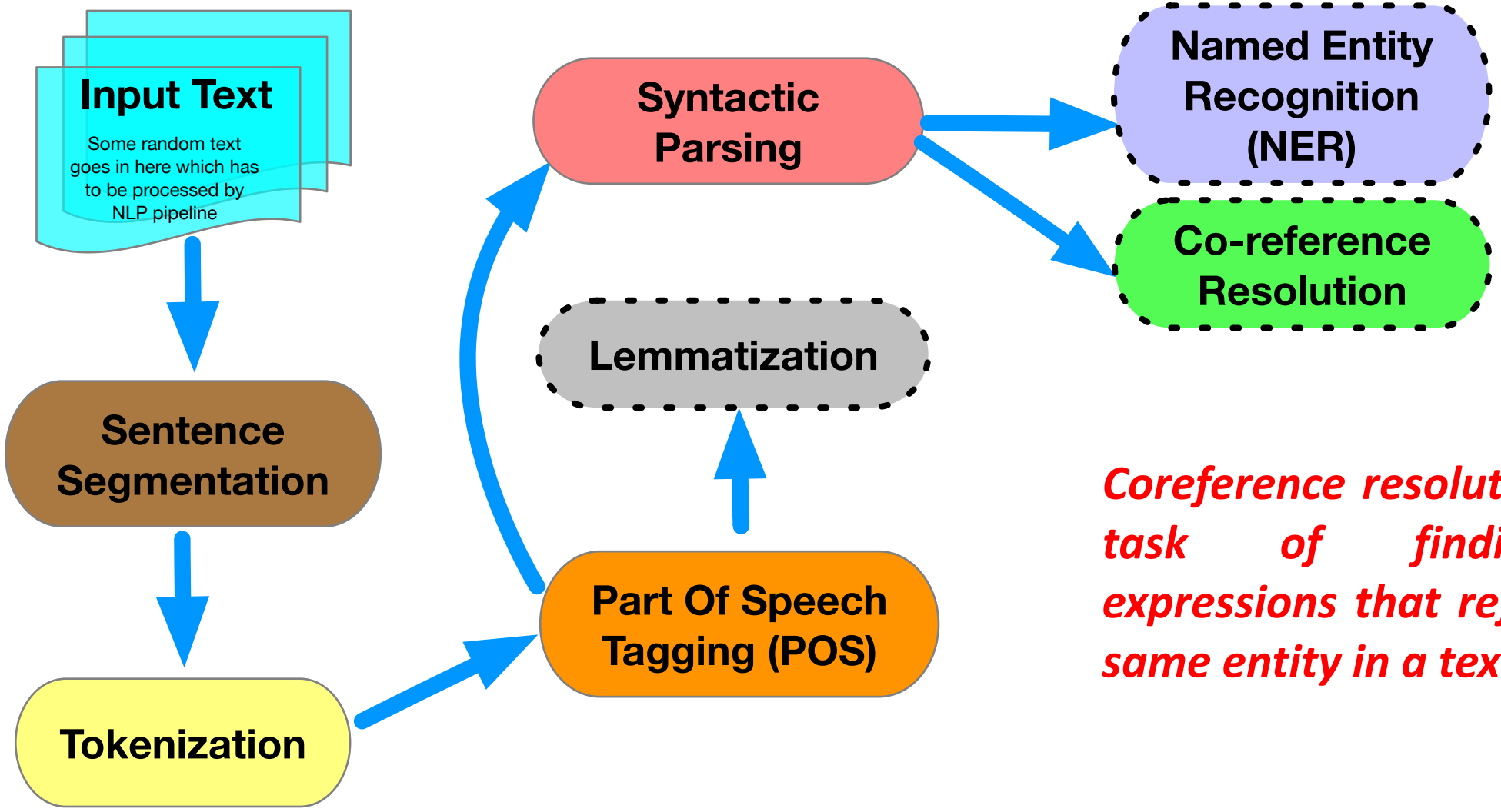


Nepal is a landlocked country in South Asia. It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain.



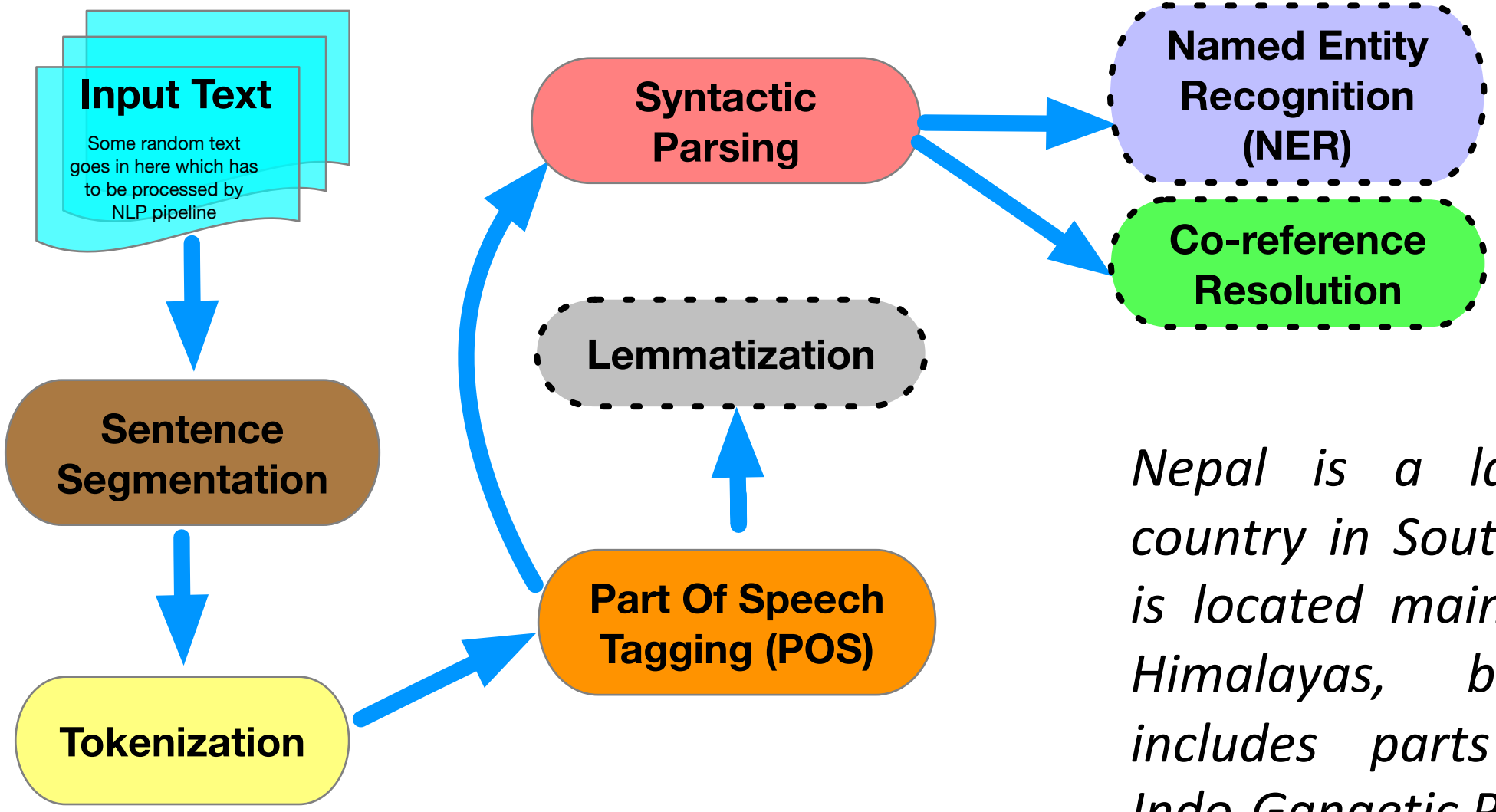
WHAT IS "It" REFERRING TO?

Nepal is a landlocked country in South Asia. It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain.



Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

Definition: https://nlp.stanford.edu/projects/coref.shtml



Nepal is a landlocked country in South Asia. It is located mainly in the Himalayas, but also includes parts of the Indo-Gangetic Plain. **Nepal**

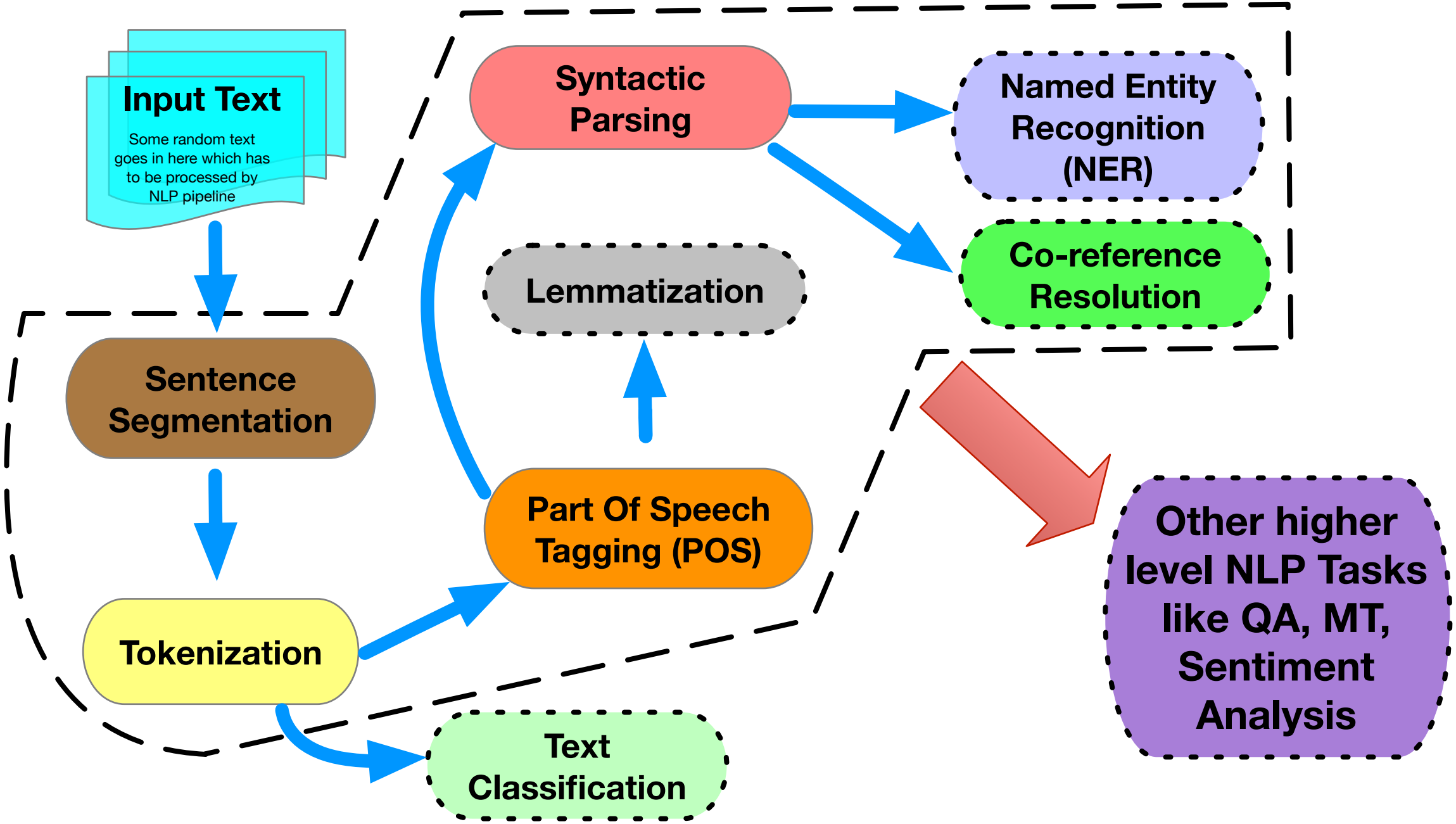
*Soleimani is the most powerful chief in Iran. Trump said that **he** is directly and indirectly responsible for the deaths of millions of people.*

Image: https://www.thedailybeast.com/

Who is "he" REFERRING TO?

*Soleimani is the most powerful chief in Iran. Trump said that **he** is directly and indirectly responsible for the deaths of millions of people.*

Image: https://www.thedailybeast.com/


Input Text
Some random text goes in here which has to be processed by NLP pipeline
Sentence Segmentation
Tokenization
Part Of Speech Tagging (POS)
Lemmatization
Syntactic Parsing
Named Entity Recognition (NER)
Co-reference Resolution
Text Classification
Other higher level NLP Tasks like QA, MT, Sentiment Analysis

Semantic Role Labeling

WHO did **WHAT** to **WHOM**

Semantic Role Labeling

WHO did **WHAT** to **WHOM**

John cooked pasta

Semantic Role Labeling

WHO did **WHAT** to **WHOM**

by what **MEANS** at what **LOCATION** and **WHEN**

John cooked pasta

Semantic Role Labeling

WHO did **WHAT** to **WHOM**

by what **MEANS** at what **LOCATION** and **WHEN**

John cooked pasta

on a stove in the kitchen at night

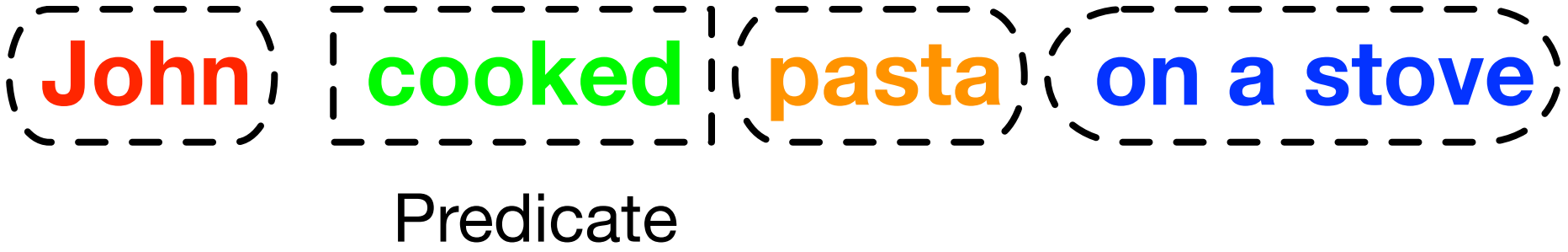
Semantic Role Labeling

- Shallow Semantic Parsing

John **cooked** **pasta** **on a stove**

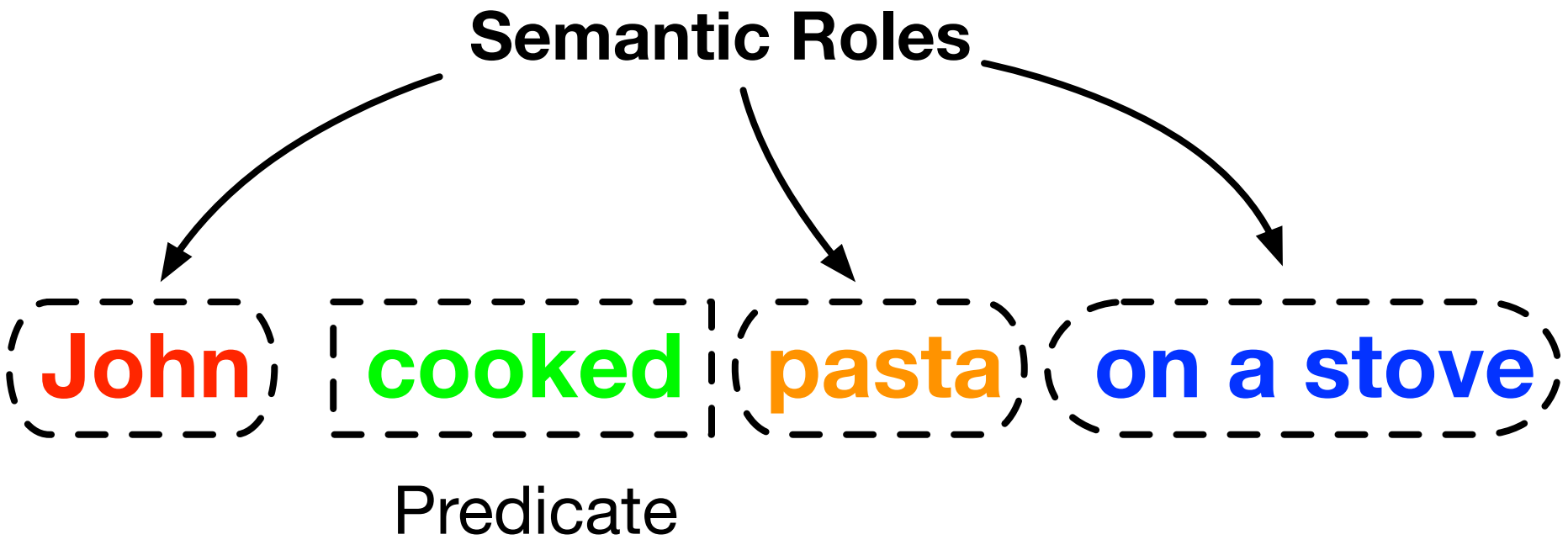
Semantic Role Labeling

- Shallow Semantic Parsing



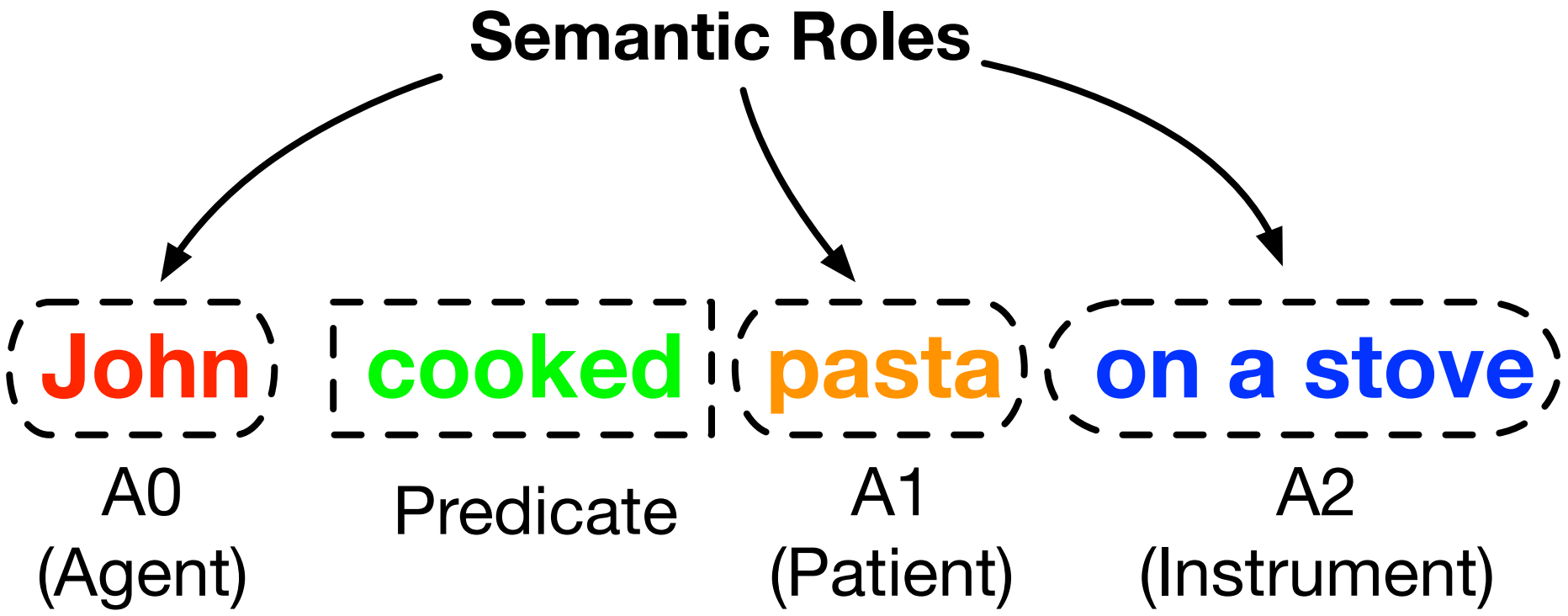
Semantic Role Labeling

- Shallow Semantic Parsing



Semantic Role Labeling

- Shallow Semantic Parsing



Semantic Frame

- Description of a type of event, relation, or entity and the participants in it.

Semantic Frame

- Description of a type of event, relation, or entity and the participants in it.
- FrameNet: https://framenet.icsi.berkeley.edu/

FrameNet is based on a theory of meaning called **Frame Semantics**, deriving from the work of Charles J. Fillmore and colleagues (Fillmore 1976, 1977, 1982, 1985, Fillmore and Baker 2001, 2010). The basic idea is straightforward: that the meanings of most words can best be understood on the basis of a **semantic frame**, a description of a type of event, relation, or entity and the participants in it. For example, the concept of cooking typically involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating_instrument). In the FrameNet project, this is represented as a **frame** called Apply_heat, and the Cook, Food, Heating_instrument and Container are called **frame elements (FEs)** . Words that **evoke** this frame, such as *fry, bake, boil,* and *broil,* are called **lexical units (LUs)** of the Apply_heat frame. Other frames are more complex, such as Revenge, which involves more FEs (Offender, Injury, Injured_Party, Avenger, and Punishment) and others are simpler, such as Placing, with only an Agent (or Cause), a thing that is placed (called a Theme) and the location in which it is placed (Goal). The job of FrameNet is to define the frames and to annotate sentences to show how the FEs fit syntactically around the word that evokes the frame, as in the following examples of Apply_heat and Revenge:

- ... [Cook the boys] ... GRILL [Food their catches] [Heating_instrument on an open fire].
- [Avenger I] 'll GET EVEN [Offender with you] [Injury for this]!

Check this: https://framenet.icsi.berkeley.edu/fndrupal/WhatIsFrameNet

Deep Learning Based Models

- Represent meaning of a character, word, sentence, paragraph, document as a vector of real values.
- The vectors are learned from (co-occurrence) statistics of the corpus.
- Have shown to be very effective in recent times.
- E.g. BERT, GPT, etc.

Summary

- It is good to know some linguistic fundamentals to develop computational models
- Words are divided into open classes and closed classes
- Based on open classes we have different phrase types.
- Verbs are very important open class category.
- Use the NLP tools for pre-processing the data.

- Next class onwards Statistical and Probabilistic Methods
  - Language Models