



# Natural Language Processing


# Text is everywhere!

## Social media




# Text is everywhere!

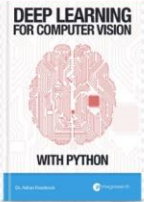
Research papers, news, etc.




efficiency of quantum vs. classical annealing in nonconvex learning problems



United States Patent 2018/004  
Class: 2018/004  
Date: 2018



DEEP LEARNING FOR COMPUTER VISION  
WITH PYTHON




Libération of cancer heat mapping technology

Scientific documents


Patent articles

Books


Review articles



The Sciences



Trading Our Future for Fossil Fuels



GitHub  
Stack Overflow  
ORCID

Magazines

Blogs

Crowd-sourced platforms

# Diversity of Languages (Worldwide)

How many Languages are spoken today?

---

<sup>1</sup>Source: <https://www.ethnologue.com/guides/how-many-languages>

# Diversity of Languages (Worldwide)

How many Languages are spoken today?

7,111<sup>1</sup>

---

<sup>1</sup>Source: <https://www.ethnologue.com/guides/how-many-languages>

# Diversity of Languages (Worldwide)

How many Languages are spoken today?

7,111<sup>1</sup>

Can we understand the majority of the world's data? What percentage?

<sup>1</sup>Source: <https://www.ethnologue.com/guides/how-many-languages>

# Diversity of Languages (Worldwide)

How many Languages are spoken today?

7,111<sup>1</sup>

Can we understand the majority of the world's data? What percentage?

Rank ↕	Language ↕	Native speakers in millions 2007 (2010) ↕	Percentage of world population (2007) ↕
1	<b>Mandarin</b> (entire branch)	935 (955)	14.1%
2	<b>Spanish</b>	390 (405)	5.85%
3	<b>English</b>	365 (360)	5.52%
4	<b>Hindi</b> <sup>[b]</sup>	295 (310)	4.46%
5	<b>Arabic</b>	280 (295)	4.23%
6	<b>Portuguese</b>	205 (215)	3.08%
7	<b>Bengali</b> (Bangla)	200 (205)	3.05%
8	<b>Russian</b>	160 (155)	2.42%
9	<b>Japanese</b>	125 (125)	1.92%
10	<b>Punjabi</b>	95 (100)	1.44%

<sup>1</sup>Source: <https://www.ethnologue.com/guides/how-many-languages>

# Diversity of Languages (India)

Can we understand the majority of the India's data?



Total Languages: > 1650 (2011 consensus)

Hindi: 57.1%

English: 10.6%

Bengali: 8.9%

Marathi: 8.2%

Telugu: 7.8 %

...

...



# The goals of NLP

**French Sentence:** Tu Bois un Coca-cola

# The goals of NLP

**French Sentence:** Tu Bois un Coca-cola

**English Translation:** You drink a Coca-cola

# The goals of NLP

**French Sentence:** Tu Bois un Coca-cola

**English Translation:** You drink a Coca-cola

We try to understand a foreign language using some known keywords

# The goals of NLP

**French Sentence:** Tu Bois un Coca-cola

**English Translation:** You drink a Coca-cola

We try to understand a foreign language using some known keywords

## Goals of NLP

- Deep understanding of broad language constructs.
- Achieve human-like comprehension of texts/languages.
- Make computer systems to understand, draw inferences from, summarize, translate and generate accurate and natural human text and language.

# Some Applications: Language Translation

The screenshot shows the Google Translate interface. At the top is the Google logo. Below it, the word 'Translate' is on the left, and 'Turn off instant translation' with a star icon is on the right. The source language is set to 'English' and the target language is 'Hindi'. The input text is 'Let's go out for a date'. The translated text is 'चलो एक तारीख के लिए बाहर जाओ'. Below the translation, there is a small Hindi text 'चाओ एक ताारेख के ले बाहर जाओ'.

Google

Translate Turn off instant translation

English Spanish Hindi Detect language

English Spanish Hindi Translate

Let's go out for a date


चलो एक तारीख के लिए बाहर जाओ

23/5000

Suggest an edit

चाओ एक ताारेख के ले बाहर जाओ

# Language Translation



Translate

Turn off instant translation

EnglishSpanishHindiDetect language

EnglishSpanishHindi

Translate

मेरे हाथ में तेरा हाथ हो सारी जन्नतें मेरे साथ हो

49/5000

All hands should be with me in my hand.

☆🔊🔊🔊

Suggest an edit

mere haath mein tera haath ho saaree jannaten mere saath ho

# Language Translation is not easy even for humans

## Pepsi Chinese blunder

“Come alive with the Pepsi Generation”, when translated into Chinese meant,  
“Pepsi brings your relatives back from the dead.”

---

# Language Translation is not easy even for humans

## Pepsi Chinese blunder

“Come alive with the Pepsi Generation”, when translated into Chinese meant, “Pepsi brings your relatives back from the dead.”

## KFC's Chinese blunder

KFC's slogan, “Finger lickin' good”, when translated into Chinese meant “We'll eat your fingers off.”



## Some more examples...



# Query Recommendation in Search Engines



Dhoni is

dhoni is **from which state**

dhoni is **back**

dhoni is **in which team**

dhoni is

dhoni is **retiring**

dhoni is **best**

dhoni is **photo**

dhoni is **a legend**

dhoni is **injured**

dhoni is **age**

# Spelling Correction



natural langauage



All

News

Images

Videos

Maps

More

Settings

Tools

About 1,95,00,00,000 results (0.45 seconds)

Showing results for **natural *language***

Search instead for **natural language**

## Information Extraction

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

# Sentiment Analysis



Saravana Kumar



**Worst software update Needed**

4 May 2019

Style: 3GB RAM

Size: 32GB

Colour: Elegant Blue

**Verified Purchase**

Bad software

Phone hangs a lot

Many issues and bugs in phone

Only Battery and front camera is good

Worst from MI

208 people found this helpful

Helpful

▼ 1 Comment

Report abuse

# Recent Trends: Fake news detection

## Lies

IndiaToday.in

New Delhi, April 13, 2017 | UPDATED 18:31 IST

### Uttar Pradesh government ends reservation for SC, ST, OBC candidates in private medical colleges

The Uttar Pradesh government made an announcement to end caste-based reservations in private medical and dental colleges of up. An order has been passed to do away with the quota for candidates belonging to SC, ST and OBC categories.



Uttar Pradesh CM Yogi Adityanath

#### RELATED STORIES

- Union Cabinet approves IIPe Visakhapatnam's status as institute of national importance
- Delhi HC suggests outsiders should be barred from entering JNU
- 25,000 Indian students to be trained in cyber security by Cisco
- Haryana boy makes 8 to IIM-A with 89.67 percentile

## Truth

### 'Caste-based reservation never existed in private medical colleges in Uttar Pradesh'

Shalvee Sharda | TNN | Apr 13, 2017, 07:27 PM IST



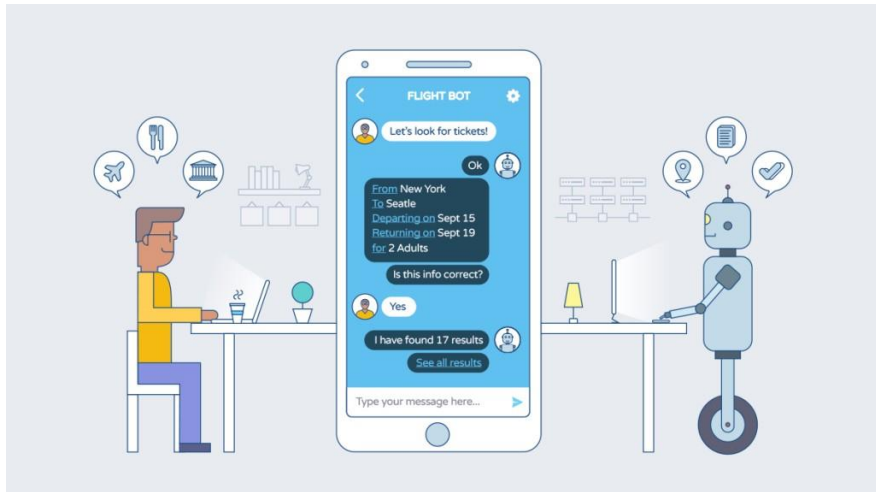
#### HIGHLIGHTS

- UP medical education department refuted reports of Yogi government abolishing reservations in private medical colleges
- It said caste-based reservation was never a part of the admission process in private colleges in UP



File photograph used for representational purpose.

# Recent Trends: Chatbots



## Other Goals

- Text Summarization
- Opinion dynamics
- Spam detection
- ...



# Other Goals

- Text Summarization
- Opinion dynamics
- Spam detection
- ...

Natural Language Technology not yet perfect

But still good enough for several useful applications

# Why is NLP hard?

## Compounding

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

195 characters (with 428 characters when transliterated into the roman writing system).

# Why is NLP hard?

## Ambiguity

### Lexical Ambiguity

The presence of two or more possible meanings within a single word.



"I saw her duck."

### Syntactic Ambiguity

The presence of two or more possible meanings within a single sentence or sequence of words.



"The chicken is ready to eat."

# Why is NLP hard?

## Ambiguity



# Why else is NLP hard?

## Shorthand text



# Why else is NLP hard?

Non-standard English

# Why else is NLP hard?

## Segmentation Issues

the New York New Haven Railroad

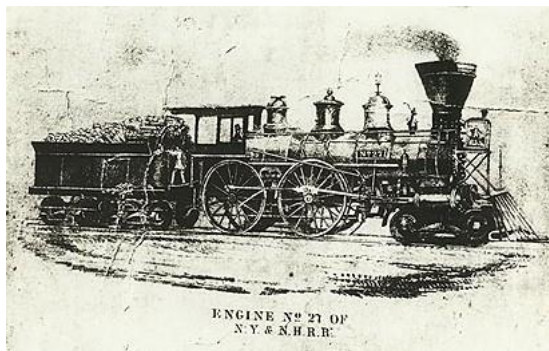
# Why else is NLP hard?

## Segmentation Issues

the New York New Haven Railroad

the [New] [York New] [Haven] [Railroad]

the [New York] [New Haven] [Railroad]





# Why else is NLP hard?

## Idioms

- Dark horse
- Ball in your court
- Burn the midnight oil

# Why else is NLP hard?

**Idioms** : An expression whose meaning is different from the meanings of the individual words in it.

## Idioms Example

- Dark horse
- Ball in your court   Burn the midnight oil
- 

**Neologisms**: A new word, phrase or expression, or a new meaning of a familiar word

- Unfriend
- Retweet
- Google/Skype/photoshop

# Why is NLP hard?

## New Senses of a word

- That's *sick* dude!
- Giants

# Why is NLP hard?

## New Senses of a word

- That's *sick* dude!
- Giants ... *multinationals, conglomerates, manufacturers*

# Why is NLP hard?

## New Senses of a word

- That's *sick* dude!
- Giants ... *multinationals, conglomerates, manufacturers*

## Tricky Entity Names

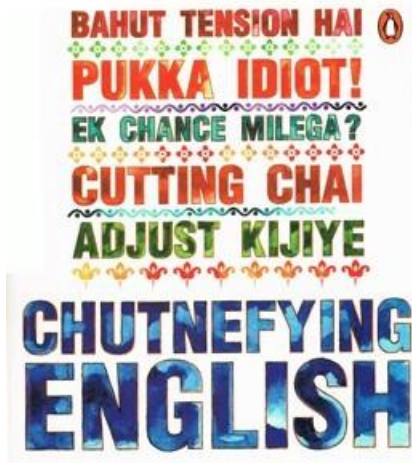
- Where is *A Bug's Life* playing ...
- *Let It Be* was recorded ...

# Why is NLP hard?

Code Mixing/switching



Romanization



# What we do in NLP?

## Create annotated corpora

Brown Corpus, Google Books Ngram Corpus, Reuters Newswire Topic Classification, IMDB Movie Review Sentiment Classification, Project Gutenberg, etc.

## Create Models/Algorithms

LDA, BERT, CKY, Edit Distance, CRF++, etc.

## Create Tools

CoreNLP, NLTK, Gensim, SpaCy, etc.

# Stages in NLP (*traditional view*)

- *Phonetics and phonology*
- *Morphology*
- *Lexical Analysis*
- *Syntactic Analysis*
- *Semantic Analysis*
- *Pragmatics*
- *Discourse*

*Source: IITB NLP Course by Pushpak Bhattacharyya*



# Phonetics

- Processing of speech
- Challenges
  - Homophones: *bank (finance)* vs. *bank (river bank)*
  - Near Homophones: *maatras* vs. *maatra (hin)*
  - Word Boundary
    - *aajaayenge (aa jaayenge (will come) or aaj aayenge (will come today)*
    - *I got [ua]plate*
  - Phrase boundary
    - *mtech1 students are especially exhorted to attend as such seminars are integral to one's post-graduate education*
  - Disfluency: *ah, um, ahem etc.*

# Morphology

- Word formation rules from *root* words
- Nouns: Plural (*boy-boys*); Gender marking (czar-czarina)
- Verbs: Tense (*stretch-stretched*); Aspect (e.g. *perfective sit-had sat*); Modality (e.g. *request khaanaa → khaaie*)
- First crucial first step in NLP
- Languages rich in morphology: e.g., Dravidian, Hungarian, Turkish
- Languages poor in morphology: Chinese, English
- Languages with rich morphology have the advantage of easier processing at higher stages of processing
- A task of interest to computer science: *Finite State Machines for Word Morphology*

## *Lexical Analysis*

- Essentially refers to dictionary access and obtaining the properties of the word

*e.g. dog*

*noun (lexical property)*

*take-'s'-in-plural (morph property)*

*animate (semantic property)*

*4-legged (-do-)*

*carnivore (-do)*

- Challenge: *Lexical or word sense disambiguation*

## *Lexical Disambiguation*

First step: *part of Speech Disambiguation*

*Dog as a noun (animal)*

*Dog as a verb (to pursue)*

Sense Disambiguation

*Dog (as animal)*

*Dog (as a very detestable person)*

Needs word relationships in a context

*The chair emphasized the need for adult education*

Very common in day to day communications

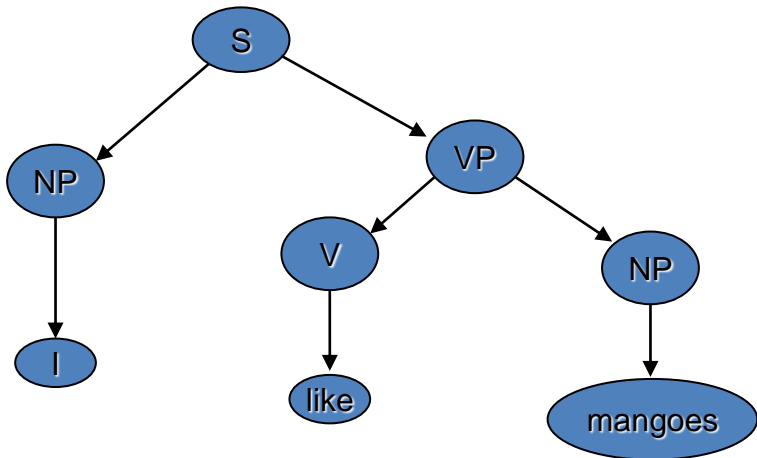
Satellite Channel Ad: *Watch what you want, when you want*

(two senses of watch)

e.g., Ground breaking ceremony/research

## *Syntax Processing Stage*

### Structure Detection



## *Parsing Strategy*

Driven by grammar

S-> NP VP

NP-> N | PRON

VP-> V NP | V PP

N-> Mangoes

PRON-> I

V-> like

## Challenges in Syntactic Processing: Structural Ambiguity

- *Scope*

1. *The old men and women were taken to safe locations (old men and women) vs. ((old men) and women)*
2. *No smoking areas will allow Hookas inside*

- *Preposition Phrase Attachment*

- *I saw the boy with a telescope (who has the telescope?)*
- *I saw the mountain with a telescope (world knowledge: mountain cannot be an instrument of seeing)*
- *I saw the boy with the pony-tail (world knowledge: pony-tail cannot be an instrument of seeing)*
- *Very ubiquitous: newspaper headline “20 years later, BMC pays father 20 lakhs for causing son’s death”*

## *Structural Ambiguity...*

### *Overheard*

*I did not know my PDA had a phone for 3 months*

### *An actual sentence in the newspaper*

*The camera man shot the man with the gun when  
he was near Tendulkar*



## *Semantic Analysis*

- *Representation in terms of*  
Predicate calculus/Semantic  
Nets/Frames/Conceptual Dependencies and  
Scripts
- *John gave a book to Mary*  
Give action: Agent: John, Object: Book,  
Recipient: Mary
- *Challenge: ambiguity in semantic role labeling*  
*(Eng) Visiting aunts can be a nuisance*  
*(Hin) aapko mujhe mithaai khilaanii padegii (ambiguous*  
*in Marathi and Bengali too; not in Dravidian languages)*

# Pragmatics

- *Very hard problem*
- *Model user intention*
  - *Tourist (in a hurry, checking out of the hotel, motioning to the service boy): Boy, go upstairs and see if my sandals are under the divan. Do not be late. I just have 15 minutes to catch the train.*
  - *Boy (running upstairs and coming back panting): yes sir, they are there.*
- *World knowledge*
  - *WHY INDIA NEEDS A SECOND OCTOBER (Tol, 2/10/07)*

# Discourse

- Processing of *sequence* of sentences
  - *Mother to John:*
  - *John go to school. It is open today. Should you bunk?*  
*Father will be very angry.*
- Ambiguity of *open*
- *bunk* what?
- *Why will the father be angry?*
  - *Complex chain of reasoning and application of world knowledge*
  - Ambiguity of *father*  
*father as parent*  
or  
*father as headmaster*

# Reference Books

- Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall. 2009.
- Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press. 1999.
- Steven Bird, Ewan Klein and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media. 2009.
- Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press. 2016.