

---

# Machine Learning Natural Language

Srijith P. K.

Computer Science and Automation  
Indian Institute of Science

# NLP System : IBM Watson

---

Question Answering System

Quiz show Jeopardy!

- “The first person mentioned by name in 'The man in the Iron mask' is this hero of a previous book by the same author”



# Natural Language Processing

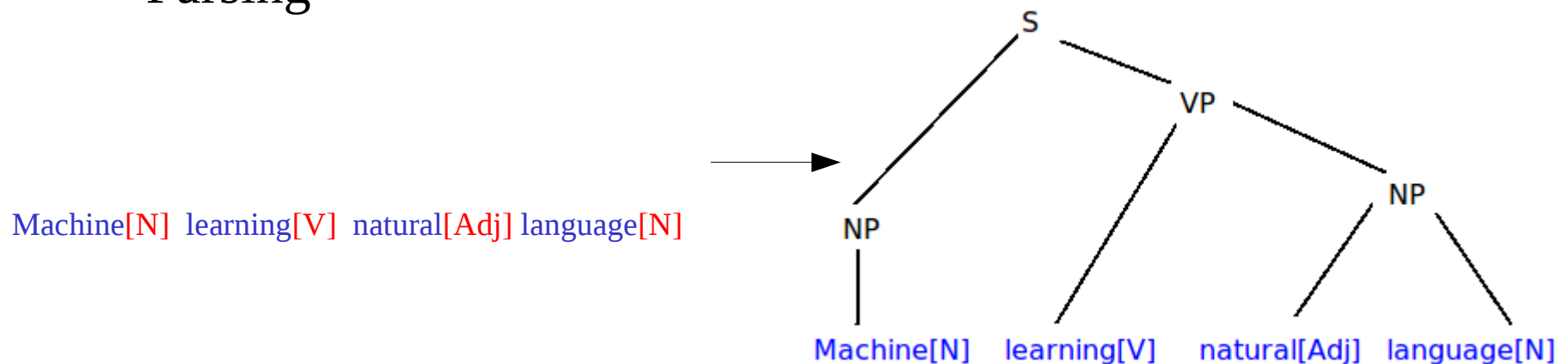
---

- NLP focuses on developing systems that allow computers to perform useful tasks involving human language
  - Also called **Computational Linguistics**
- NLP applications
  - Information Retrieval
  - Question Answering
  - Machine Translation
  - Information Extraction

# NLP : Tasks

---

- Segmentation : words, sentences
- Morphology : plural “boy” “boys” , “agree” ---> “agreement”  
Stemming "fishing", "fished", "fish", "fisher" ---> "fish"
- Syntactic Analysis : structural relationships between words
  - Part of Speech (POS) Tagging  
Machine[N] learning[V] natural[Adj] language[N]
  - Parsing



# NLP : Tasks

---

- Semantics

- Word Sense Disambiguation : “I went to *bank*”
- Semantic role labelling :  
“*Mary*[Agent] sold the *book*[goods] to *John*[Receipient]”

- Pragmatics : how language is used to accomplish goals

- I’m sorry Dave, I’m afraid I can’t do that [Polite]
- I can't do that [Rude]

- Discourse

Coreference Resolution : linking pronouns/abbreviations to entities

“I saw *Scott* yesterday. *He* was fishing by the lake.”

“*Indian Institute of Science* is a public institution located in Bangalore.

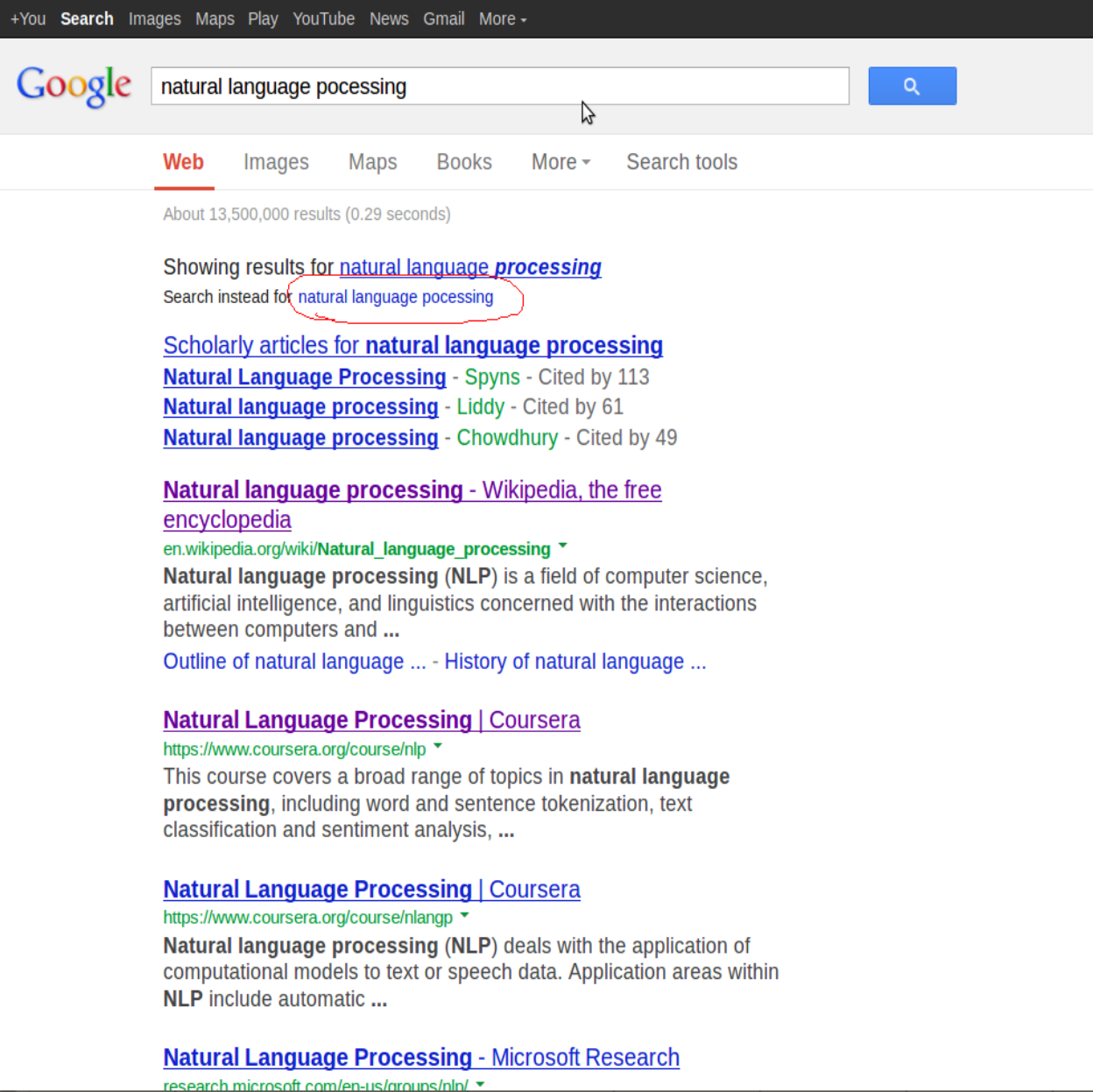
*IISc.* was established in 1909.”

Named Entity recognition (NER) : person, location, price, product

*Mohandas Karamchand Gandhi* was born in *Porbandar, Gujarath*

# NLP application : Information Retrieval

- Stemming
- Spell checking
- Query expansion
- Word sense disambiguation



The screenshot shows a Google search interface. At the top, there are navigation links: '+You Search Images Maps Play YouTube News Gmail More'. The search bar contains the text 'natural language processing' and a blue search button with a magnifying glass icon. Below the search bar, there are tabs for 'Web', 'Images', 'Maps', 'Books', 'More', and 'Search tools'. The search results show 'About 13,500,000 results (0.29 seconds)'. The first result is 'Showing results for [natural language processing](#)' with a red circle around the text. Below this, there are several search results:

- [Scholarly articles for natural language processing](#)
- [Natural Language Processing - Spyns](#) - Cited by 113
- [Natural language processing - Liddy](#) - Cited by 61
- [Natural language processing - Chowdhury](#) - Cited by 49
- [Natural language processing - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Natural\\_language\\_processing](https://en.wikipedia.org/wiki/Natural_language_processing)
- Natural language processing (NLP)** is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and ...  
[Outline of natural language ... - History of natural language ...](#)
- [Natural Language Processing | Coursera](#)  
<https://www.coursera.org/course/nlp>  
This course covers a broad range of topics in **natural language processing**, including word and sentence tokenization, text classification and sentiment analysis, ...
- [Natural Language Processing | Coursera](#)  
<https://www.coursera.org/course/nlangp>  
**Natural language processing (NLP)** deals with the application of computational models to text or speech data. Application areas within NLP include automatic ...
- [Natural Language Processing - Microsoft Research](#)  
[research.microsoft.com/en-us/groups/nlp/](https://research.microsoft.com/en-us/groups/nlp/)

# NLP application : Question Answering

- Determine type of question and answer
- Parse the question and identify relations  
POS tagging, Parsing, named entity recognition



**AnswerBus**

Who is the prime minitser of india ?

Type in your question! We only want to offer the best possible answers to you.

---

**Possible Answers**

**More Resources from Wiki and Web**

[Manmohan Singh \(prime minister of India\) -- Encyclopedia Britannica](#) (www.britannica.com/EBchecked/topic/936615/Manmohan-Sing...)

Indian economist and politician, who became prime minister of India in 2004. A Sikh, he was the first non-Hindu to occupy the office. Singh attended Punjab ...  
[PMs of India - Prime Minister's Office](#) (pmindia.nic.in/pmsfindia.php)

india.gov.in. Prime Ministers of India. Name, Tenure, Party.  
[Indira Gandhi \(prime minister of India\) -- Encyclopedia Britannica](#) (www.britannica.com/EBchecked/topic/225198/Indira-Gandhi)

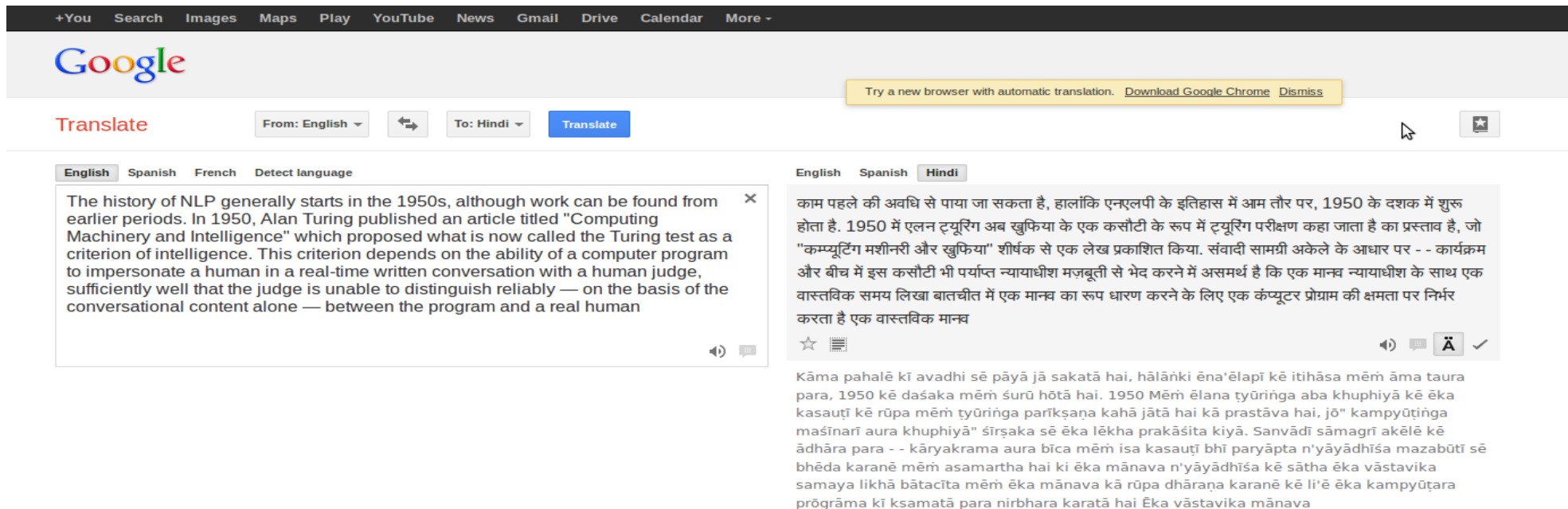
Politician who served as prime minister of India for three consecutive terms (1966 -77) and a fourth term from 1980 until she was assassinated in 1984. She was ...  
[1947-2009: List of Prime Ministers of India - Oneindia News](#) (news.oneindia.in/2009/05/22/1947-2009-list-of-prime-min...)

May 22, 2009 ... indian prime ministers, manmohan singh, indian prime minister, ab vajpayee, indira gandhi, lok sabha, jawaharlal nehru, gulzari lal nanda, ...  
[Prime Minister of India - NNDB.com](#) (www.nndb.com/gov/751/000047610/)

Prime Minister of India. GOVERNMENT OFFICE. Prime Minister ...

# NLP application : Machine Translation

- Sentence alignment
- POS tagging
- Parsing
- Sentence generation grammars
- Named Entity Recognition (“New Delhi”)



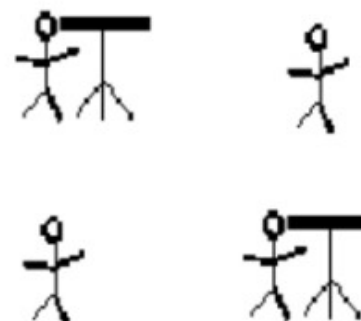
The screenshot shows the Google Translate interface. The source text in English is: "The history of NLP generally starts in the 1950s, although work can be found from earlier periods. In 1950, Alan Turing published an article titled "Computing Machinery and Intelligence" which proposed what is now called the Turing test as a criterion of intelligence. This criterion depends on the ability of a computer program to impersonate a human in a real-time written conversation with a human judge, sufficiently well that the judge is unable to distinguish reliably — on the basis of the conversational content alone — between the program and a real human". The translated text in Hindi is: "काम पहले की अवधि से पाया जा सकता है, हालांकि एनएलपी के इतिहास में आम तौर पर, 1950 के दशक में शुरू होता है। 1950 में एलन ट्यूरिंग अब खुफिया के एक कसौटी के रूप में ट्यूरिंग परीक्षण कहा जाता है का प्रस्ताव है, जो "कम्प्यूटिंग मशीनरी और खुफिया" शीर्षक से एक लेख प्रकाशित किया। संवादी सामग्री अकेले के आधार पर - - कार्यक्रम और बीच में इस कसौटी भी पर्याप्त न्यायाधीश मजबूती से भेद करने में असमर्थ है कि एक मानव न्यायाधीश के साथ एक वास्तविक समय लिखा बातचीत में एक मानव का रूप धारण करने के लिए एक कंप्यूटर प्रोग्राम की क्षमता पर निर्भर करता है एक वास्तविक मानव".



# NLP is hard

- Natural language is **ambiguous**
- Sentence Segmentation : “I went out with Mr. Smith.”
- Syntactic

“Flies[Noun/Verb] like flower[Noun/Verb]”



“I saw the man with the telescope” vs

“I saw the man with the telescope”

- Semantic

“I put the **plant** in the window” vs “Ford put the **plant** in Mexico”

- Ambiguity is Explosive

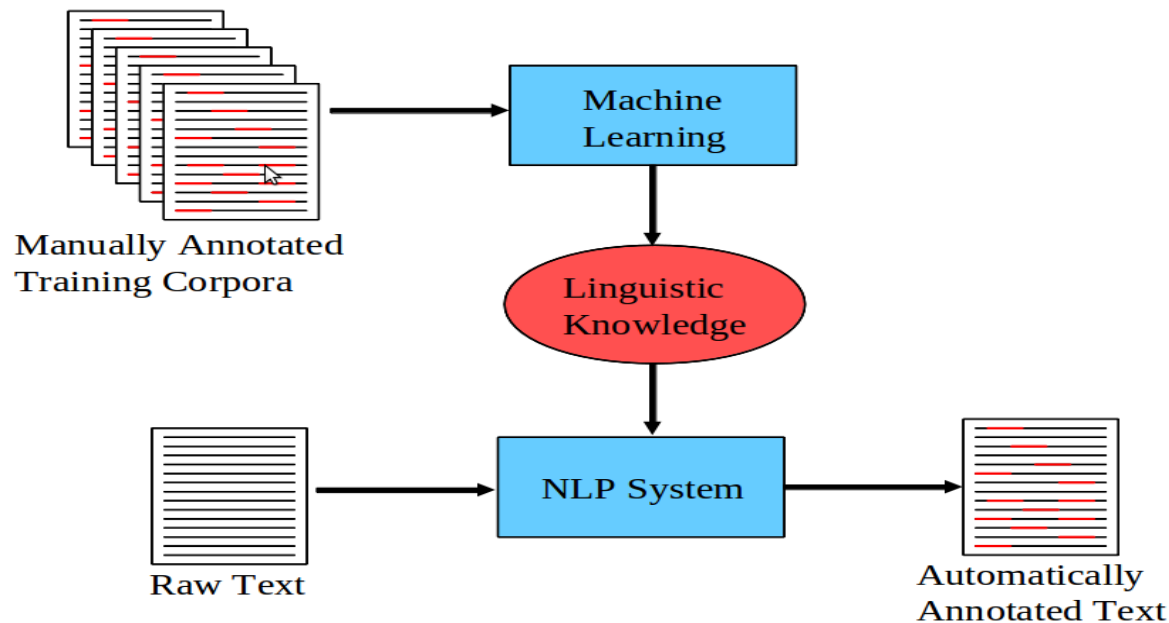
“I saw the man on the hill with the telescope.”: 4 parses



# Machine Learning Natural Language

---

- “Rules” in language have numerous exceptions and irregularities
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- Use **machine learning** methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.



# Machine Learning POS Tagging

---

- Lowest level of syntactic analysis
- Useful for Parsing and word sense disambiguation
- Ambiguity in POS tagging

Flies[Noun] like[Verb] flower[Noun]

Time flies[Verb] like[Prep] an arrow.

**Learning** : Train models on human annotated corpora like the Penn Treebank.

1 Pierre/NNP Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/, will/MD join/VB the/DT board/NN as/IN a/DT nonexecutive/JJ director/NN Nov./NNP 29/CD ./.

2 Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN ./.

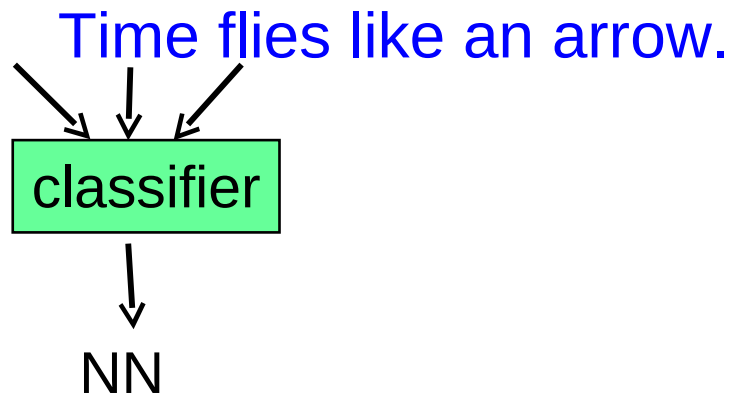
3 Rudolph/NNP Agnew/NNP ,/, 55/CD years/NNS old/JJ and/CC chairman/NN of/IN Consolidated/NNP Gold/NNP Fields/NNP PLC/NNP ,/, was/VBD named/VBN a/DT nonexecutive/JJ director/NN of/IN this/DT British/JJ industrial/JJ conglomerate/NN ./.

# POS Tagging

---

## Classification

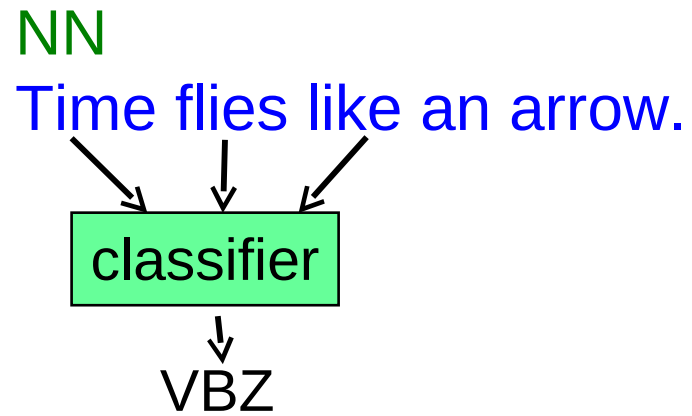
Classify each word independently but use as input features, information about the surrounding words.



# POS Tagging

---

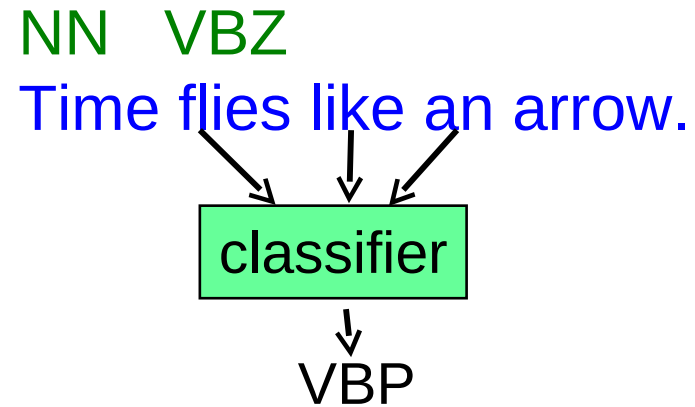
## Classification



# POS Tagging

---

- Classification



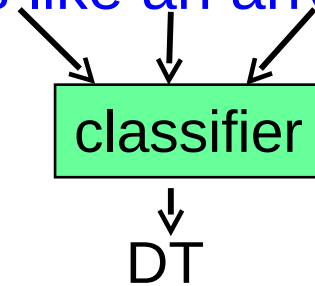
# POS Tagging

---

- Classification

NN VBZ VBP

Time flies like an arrow.



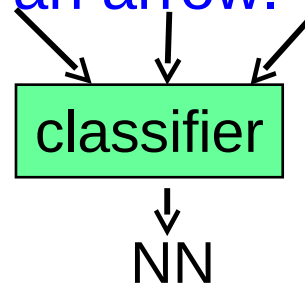
# POS Tagging

---

## Classification

NN VBZ VBP DT

Time flies like an arrow.





# POS Tagging

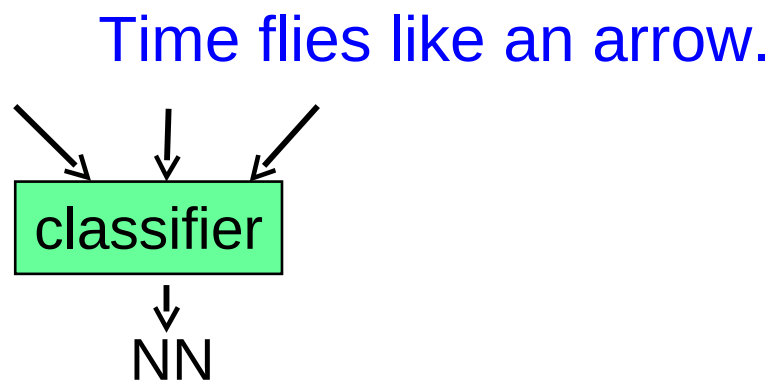
---

## Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

## Sequence Labeling

Tags of words are dependent on the tags of other words in the sentence, particularly their neighbors



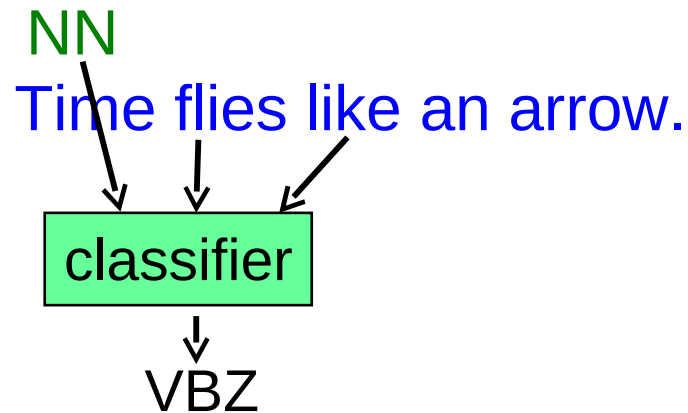
# POS Tagging

---

## Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

## Sequence Labeling



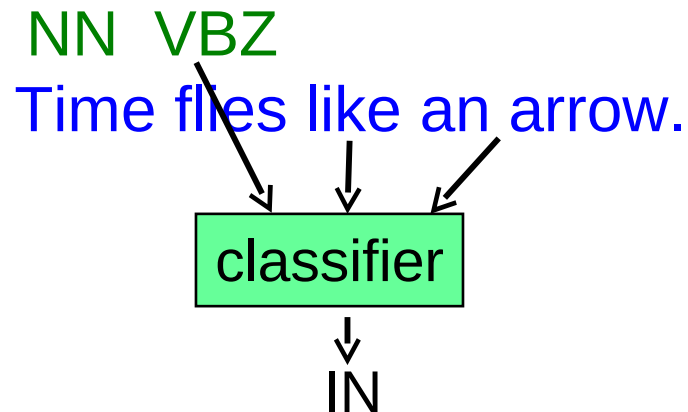
# POS Tagging

---

## Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

## Sequence Labeling



# POS Tagging

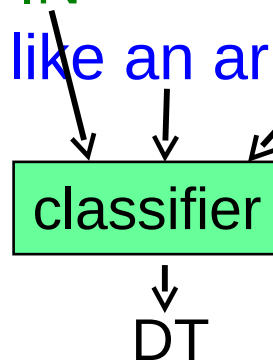
---

## Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

## Sequence Labeling

NN VBZ IN  
Time flies like an arrow.



# POS Tagging

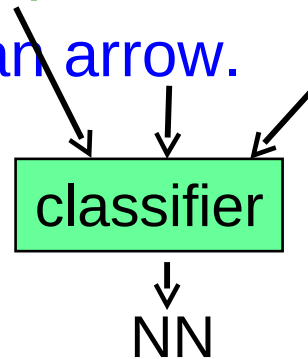
---

## Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

## Sequence Labeling

NN VBZ IN DT  
Time flies like an arrow.



# Sequence Labeling

---

Classification

NN VBZ VBP DT NN  
Time flies like an arrow.

Sequence Labeling

NN VBZ IN DT NN  
Time flies like an arrow.

POS Tagging is best modeled as a **sequence learning problem** than as a classification problem

- Information Extraction, Named Entity recognition

**Statistical models:** Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)

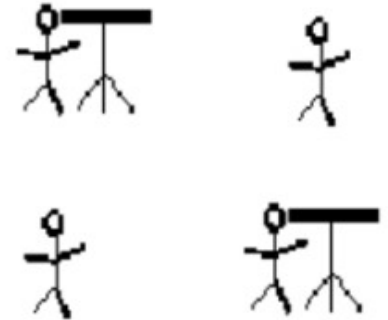
# Parsing

- Ambiguity

“I saw the man with the telescope” vs

“I saw the man with the telescope”

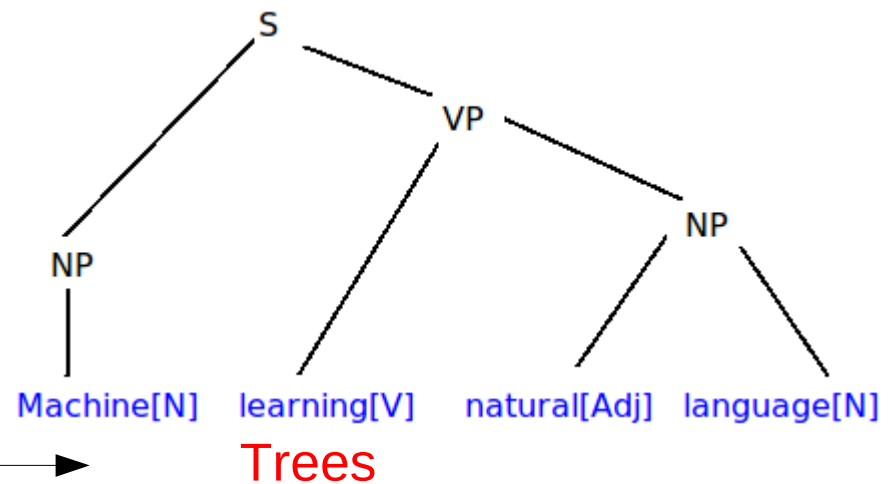
Probabilistic Context Free Grammars (PCFG)



- Structured Prediction

Machine learning natural language

Strings



**Statistical models:** Conditional Random Field, Structured perceptrons, Structured support vector machines

# Machine learning for NLP

---

- Transfer Learning, domain adaptation
  - Adapting a model learned on a resource rich language to resource scarce language
- Deep learning
  - Unsupervised learning of useful features
- Conferences : Association of Computational Linguistics(ACL), Computational Linguistics (COLING), Empirical Methods in NLP (EMNLP)
- Software tools
  - Stanford CoreNLP, openNLP, NLTK, Lingpipe



# References

---

Daniel Jurafsky and James H. Martin (2008). Speech and Language Processing

Christopher D. Manning and Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing.

Machine Learning Methods in Natural Language Processing

[http://www.cs.columbia.edu/~mcollins/papers/tutorial\\_colt.pdf](http://www.cs.columbia.edu/~mcollins/papers/tutorial_colt.pdf)

Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann and Yasemin Altun (2005), Large Margin Methods for Structured and Interdependent Output Variables

Deep learning for NLP,

<http://www.socher.org/index.php/DeepLearningTutorial/DeepLearningTutorial>

---

Thank you

# NLP application : Information Extraction

- Identifying/Extracting specific kinds of information
- Named entities (NEs): person, location, price, product
  - Mohandas Karamchand Gandhi was born in Porbandar, Gujarath
- Coreference resolution: linking pronouns/abbreviations to entities
  - “Indian Institute of Science” <> “IISc.”
- Relations: <DOB>, <spouse>, <attribute>

## Manmohan Singh



en.wikipedia.org

Manmohan Singh is the 13th and current Prime Minister of India. He is the only Prime Minister since Jawaharlal Nehru to return to power after completing a full five-year term. A Sikh, he is the first non-Hindu to occupy the office. [Wikipedia](#)

**Born:** September 26, 1932 (age 79), Gah

**Spouse:** Gursharan Kaur (m. 1958)

**Education:** St John's College, Cambridge, Nuffield College, Oxford, More

**Children:** Amrit Singh, Upinder Singh, Daman Singh

# NLP application : Categorization

- Topical : politics, sports, business
  - Sentiment: positive, negative, neutral
- ## POS tagging to obtain adjectives

News India edition Modern

Top Stories

News near you

India

World

**Business**

Mahindra Satyam

Air India

Ashok Leyland

Indira Gandhi International Airport

Reliance Industries

HCL Enterprise

Emergency landing Mozambique

Pranab Mukherjee India


Technology

Entertainment

Sports

---


**Business**

 **Nifty holds on to 5100 as Spain surges 5%**  
NDTV - 10 minutes ago


A strong opening on European bourses, particularly Spain, helped Indian stocks maintain early gains Monday. The BSE Sensex traded nearly 150 points higher, while the Nifty index managed to stay above the 5100 mark post noon.

[Sensex touches month high; Nifty holds 5100](#) Economic Times  
[Markets hold onto gains in late noon deals](#) Business Standard

[See all 36 sources »](#)

 **5 things the numbers tell you**  
NDTV - 5 minutes ago

Car sales in India rose an annual 2.8 per cent in May, according to data from the Society of Indian Automobile Manufacturers, a seventh consecutive monthly increase but far below industry expectations, with demand hit by a hike in excise duty on the ...

 **Govt empowers Air India to sack 300 pilots on strike**  
Hindustan Times - 9 minutes ago

With Air India considering further crackdown on the striking pilots, the government today said it



## Canon PowerShot SX40 HS 12.1 MP Digital Camera

\$330 online

★★★★★ 524 reviews [+1](#) +19 Recommend this on Google

#1 in Digital Cameras

September 2011 - Canon - Point & Shoot - 12.1 megapixel - Electronic Viewfinder - Compact - CMOS - Pop-up Flash - ISO 3200

[« Back to overview](#)

## Reviews

Summary - Based on 524 reviews



What people are saying

|              |      |  |
|--------------|------|--|
| pictures     | ★★★★ | "Picture clarity is great."                    |
| features     | ★★★★ | "Product has a lot of features for the price." |
| zoom/lens    | ★★★★ | "Good if not ready for DSLR."                  |
| design       | ★★★★ | "Another point is the overall camera speed."   |
| video        | ★★★★ | "Great video quality."                         |
| screen       | ★★★★ | "Flip out screen is handy."                    |
| battery life | ★★★★ | "The battery life is pretty good."             |

## Canon PowerShot SX40 HS Review

★★★★★ By ConsumerSearch - Oct 31, 2011 - Editorial review - [ConsumerSearch](#)

Pros: Versatile lens range, excellent CMOS image sensor, shoots full 1080p HD video, optical image stabilization, lots of manual control, improved shooting speed

Cons: No RAW mode, relatively small (2.7-inch) LCD screen