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

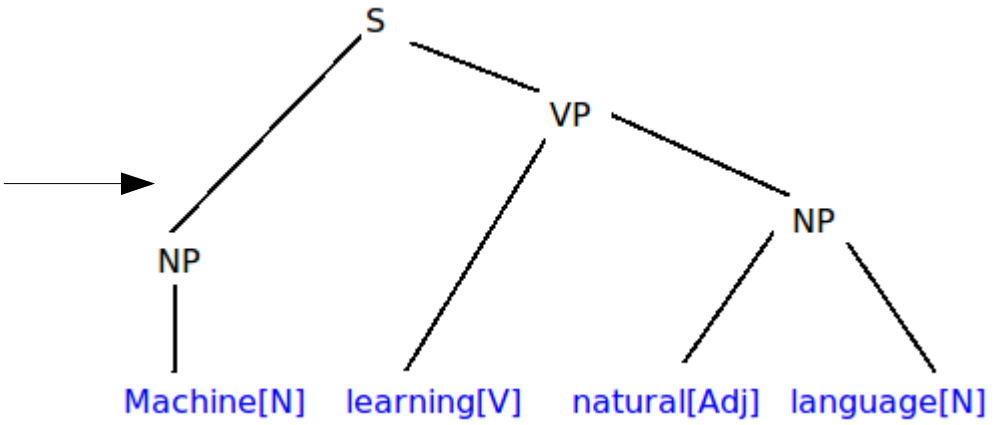
NLP : Tasks

- 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

Machine[N] learning[V] natural[Adj] language[N]



NLP : Tasks

- Semantics
  - Word Sense Disambiguation : "I went to *bank*"
  - Semantic role labelling :
    "Mary[Agent] sold the book[goods] to John[Recepient]"

- 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."

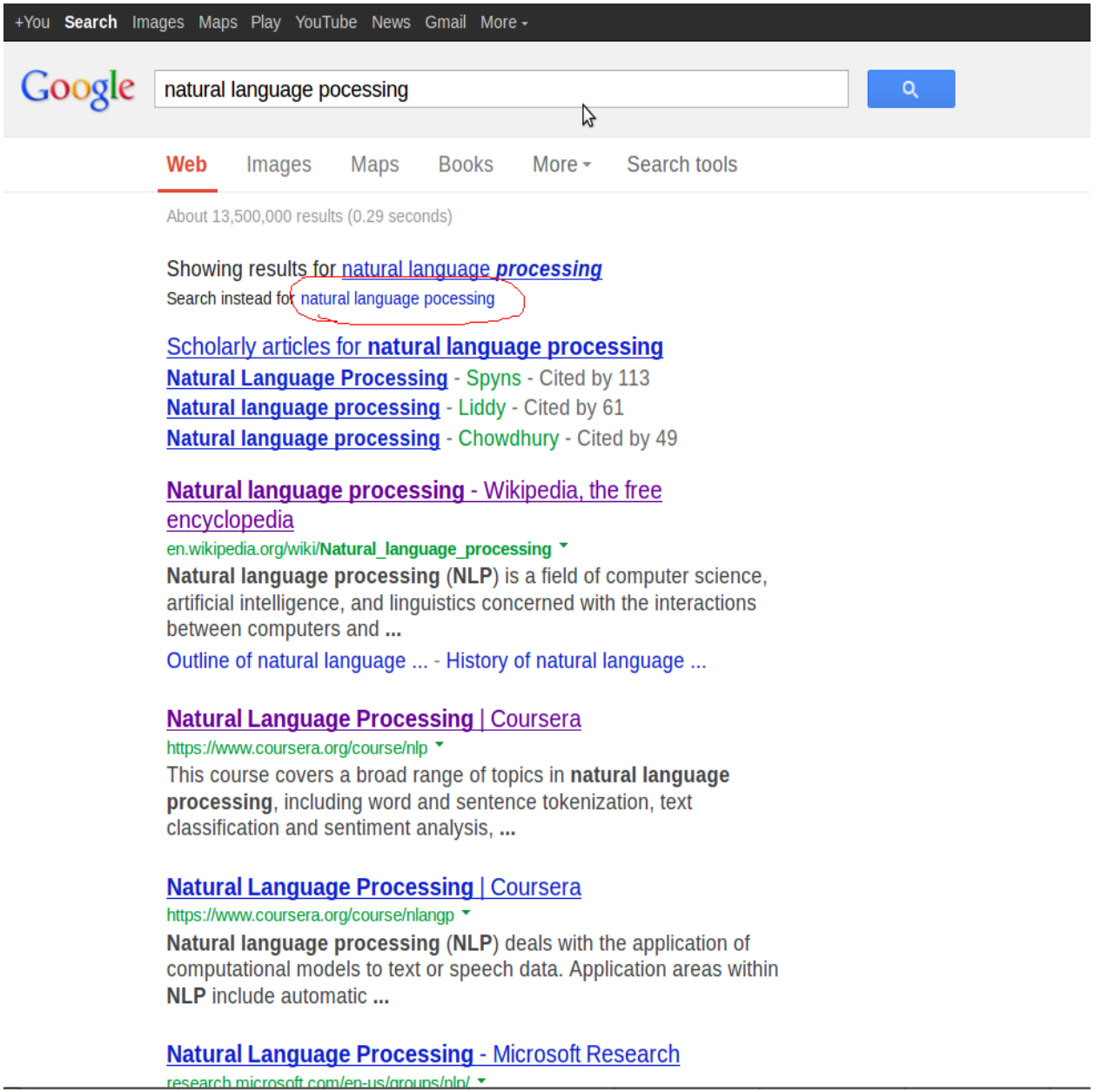
  "Indian Institute of Science is a public institution located in Bangalore.
  IISc. was established in 1909."

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



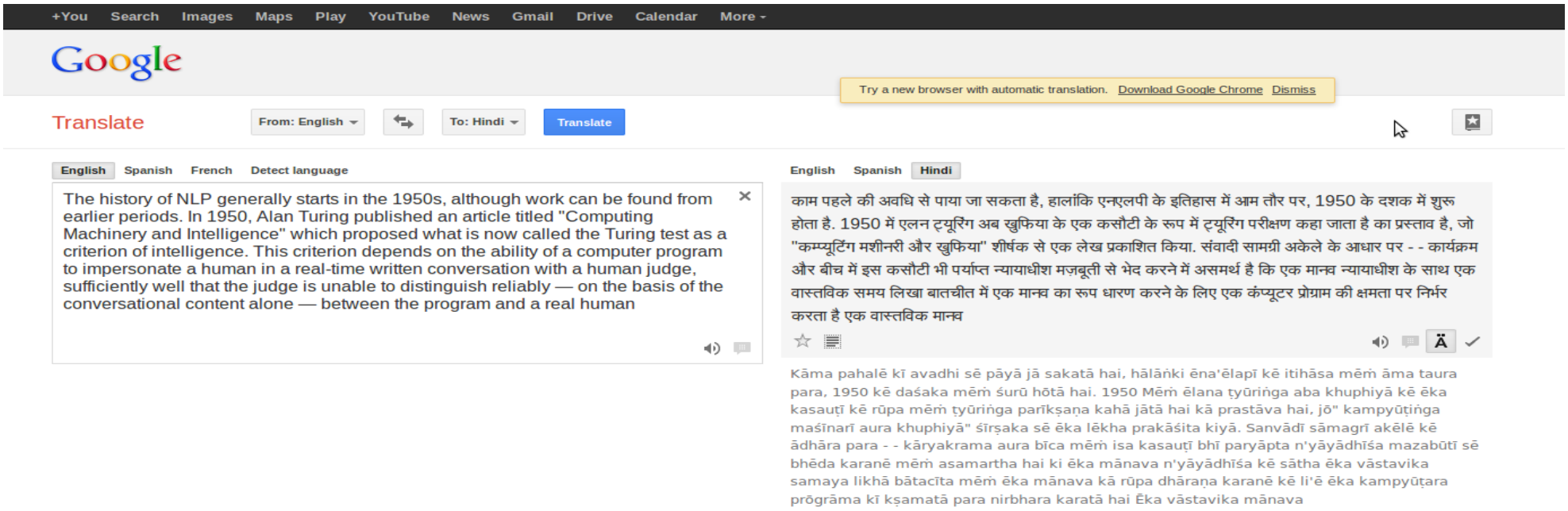
NLP application : Question Answering

- Determine type of question and answer
- Parse the question and identify relations

  POS tagging, Parsing, named entity recognition

NLP application : Machine Translation

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



NLP is hard

- Natural language is ambigious
- 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

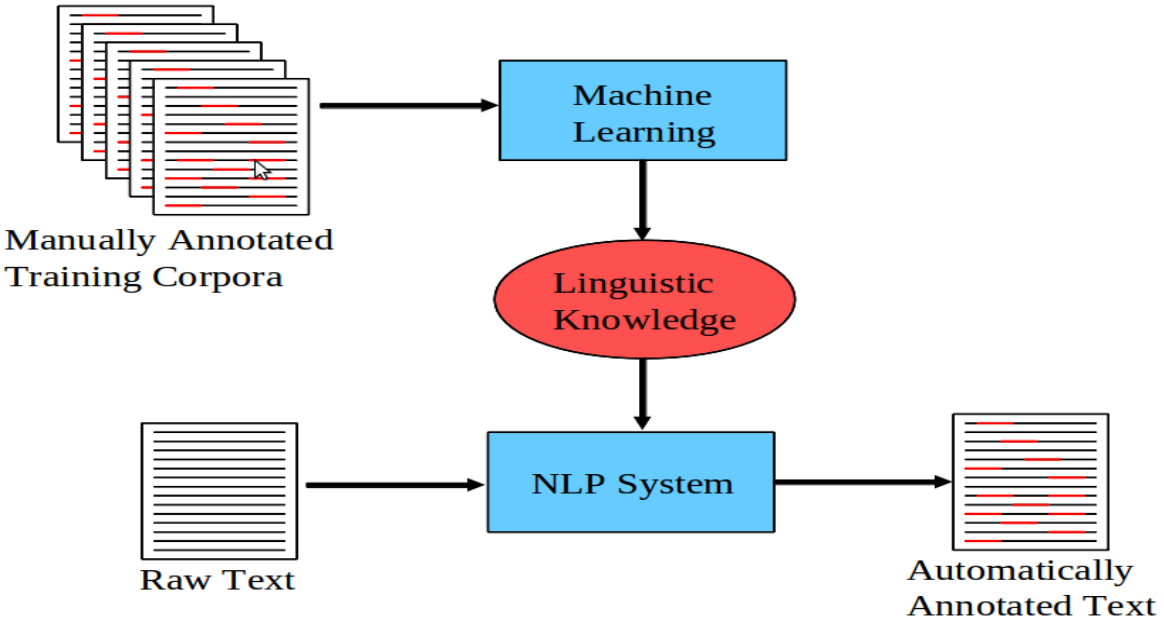
  "I put the plant in the window" vs "Ford put the plant in Mexico"

- 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.

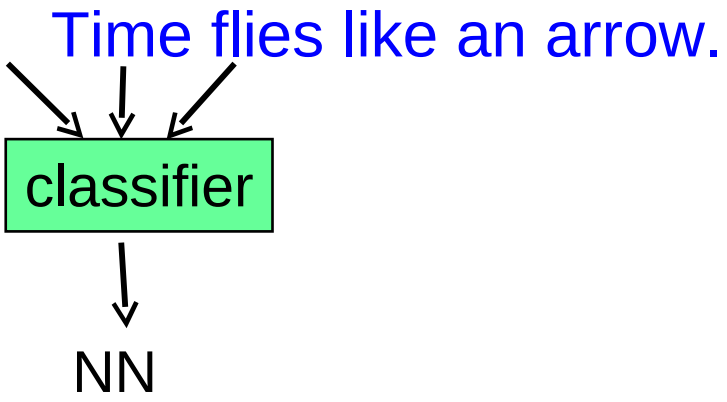
Learning : Train models on human annotated corpora like the Penn Treebank.

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

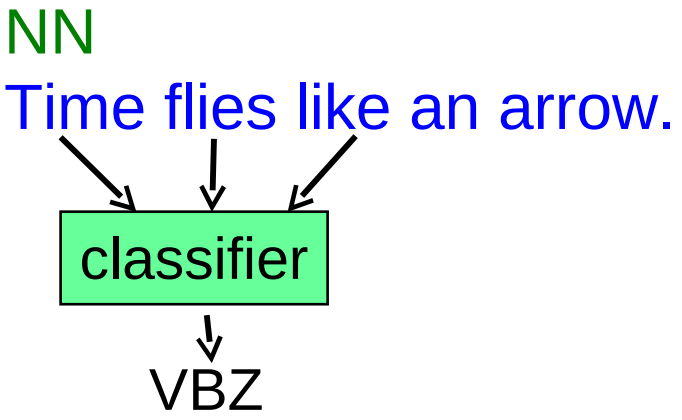
Classification

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



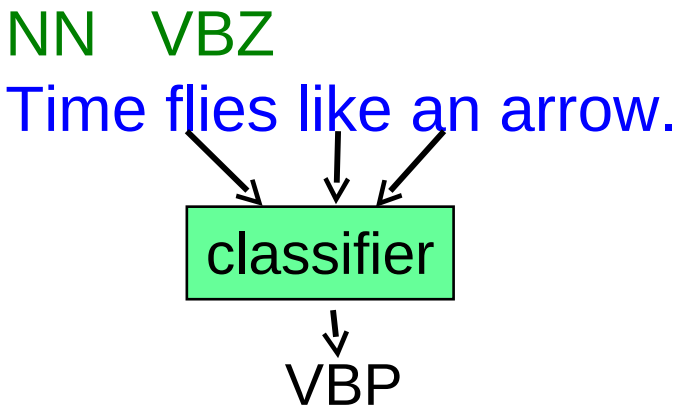
POS Tagging

Classification



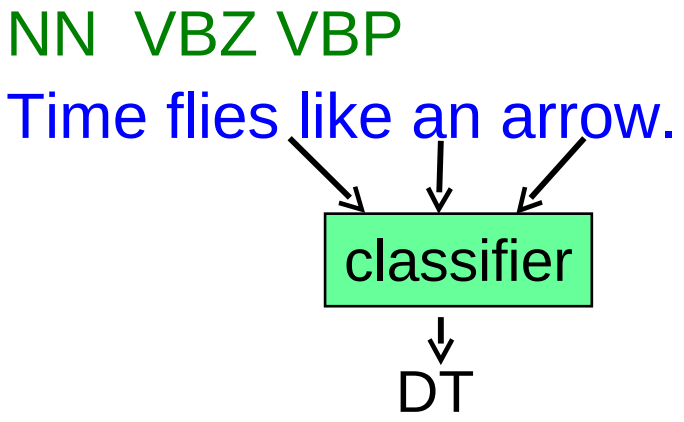
POS Tagging

- Classification



POS Tagging

- Classification

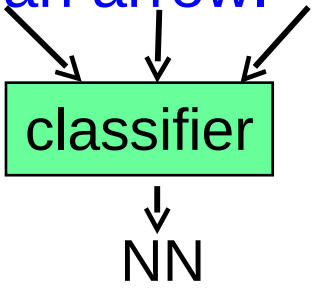


POS Tagging

Classification

NN VBZ VBP DT

Time flies like an arrow.



POS Tagging

Classification

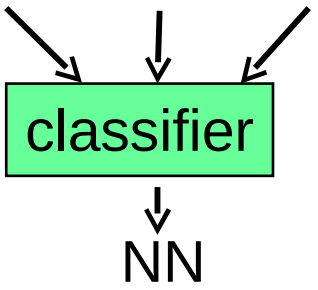
NN VBZ VBP DT NN

Time flies like an arrow.

Sequence Labeling

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

Time flies like an arrow.



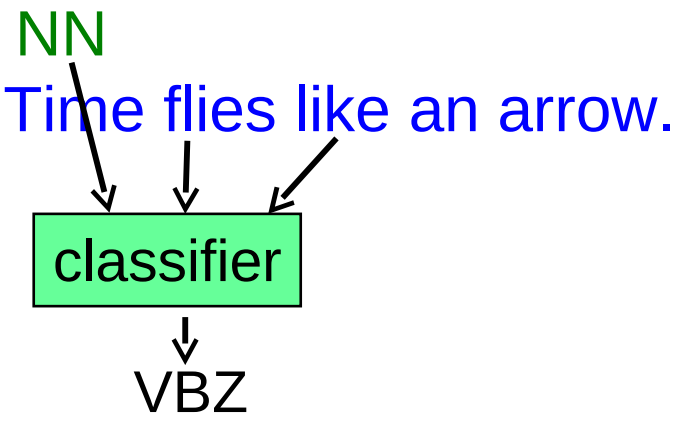
POS Tagging

Classification

NN VBZ VBP DT NN

Time flies like an arrow.

Sequence Labeling



POS Tagging

Classification

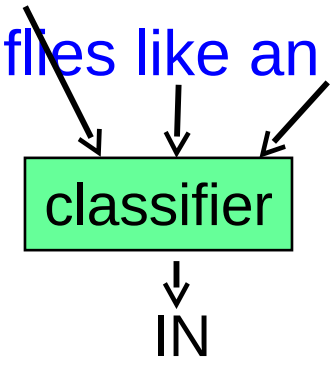
NN VBZ VBP DT NN

Time flies like an arrow.

Sequence Labeling

NN VBZ

Time flies like an arrow.



POS Tagging

Classification

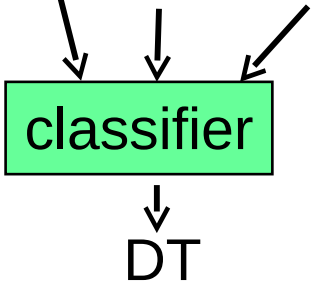
NN VBZ VBP DT NN

Time flies like an arrow.

Sequence Labeling

NN VBZ IN

Time flies like an arrow.



POS Tagging

Classification

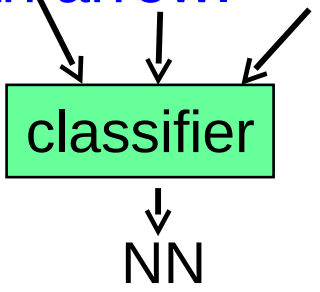
NN VBZ VBP DT NN

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

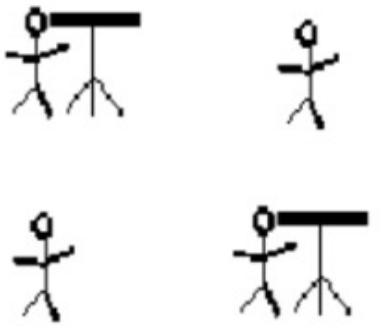
Statistical models: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)

Parsing

- Ambiguity

"I saw the man with the telescope" vs

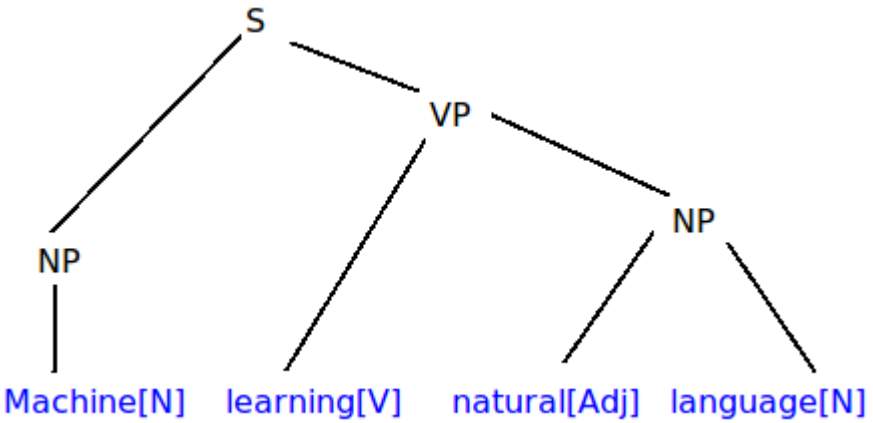
"I saw the man with the telescope"



Probabilistic Context Free Grammars (PCFG)

- Structured Prediction

Machine learning natural language →



Strings → Trees

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

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>

NLP application : Categorization

- Topical : politics, sports, business
- Sentiment: positive, negative, neutral

  POS tagging to obtain adjectives