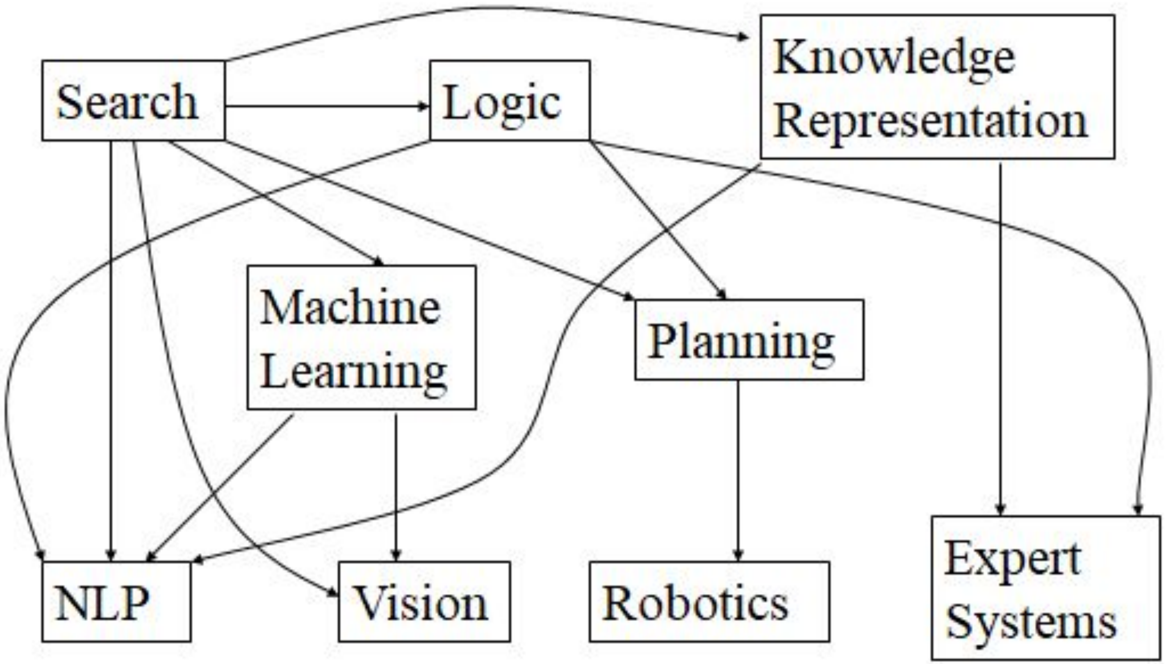
Natural Language Processing

Perpectivising NLP: Areas of AI and their inter-dependencies



AI is the forcing function for Computer Science

Stages of language processing

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

Two Views of NLP

1. Classical View: Layered Processing; Various Ambiguities (already discussed)
2. Statistical/Machine Learning View

Uncertainty in classification: **Ambiguity**

- *Visiting aunts can be a nuisance*
  - Visiting:
    - *adjective or gerund* (POS tag ambiguity)
  - Role of *aunt:*
    - *agent of visit* (aunts are visitors)
    - *object of visit* (aunts are being visited)
- Minimize uncertainty of classification with **cues** from the sentence

What *cues?*

- Position with respect to the verb:
  - *France* to the left of *beat* and *Brazil* to the right: agent-object role marking (English)
- Case marking:
  - *France ne (Hindi); ne (Marathi): agent role*
  - *Brazil ko (Hindi); laa (Marathi): object role*
- Morphology: *haraayaa (hindi); haravlaa (Marathi):*
  - *verb POS tag as indicated by the distinctive suffixes*

Cues are like *attribute-value pairs* prompting machine learning from NL data

- Constituent ML tasks
  - Goal: classification or clustering
  - Features/attributes (word position, morphology, word label *etc.*)
  - Values of features
  - Training data (corpus: annotated or un-annotated)
  - Test data (test corpus)
  - Accuracy of decision (precision, recall, F-value, MAP *etc.*)
  - Test of significance (sample space to generality)

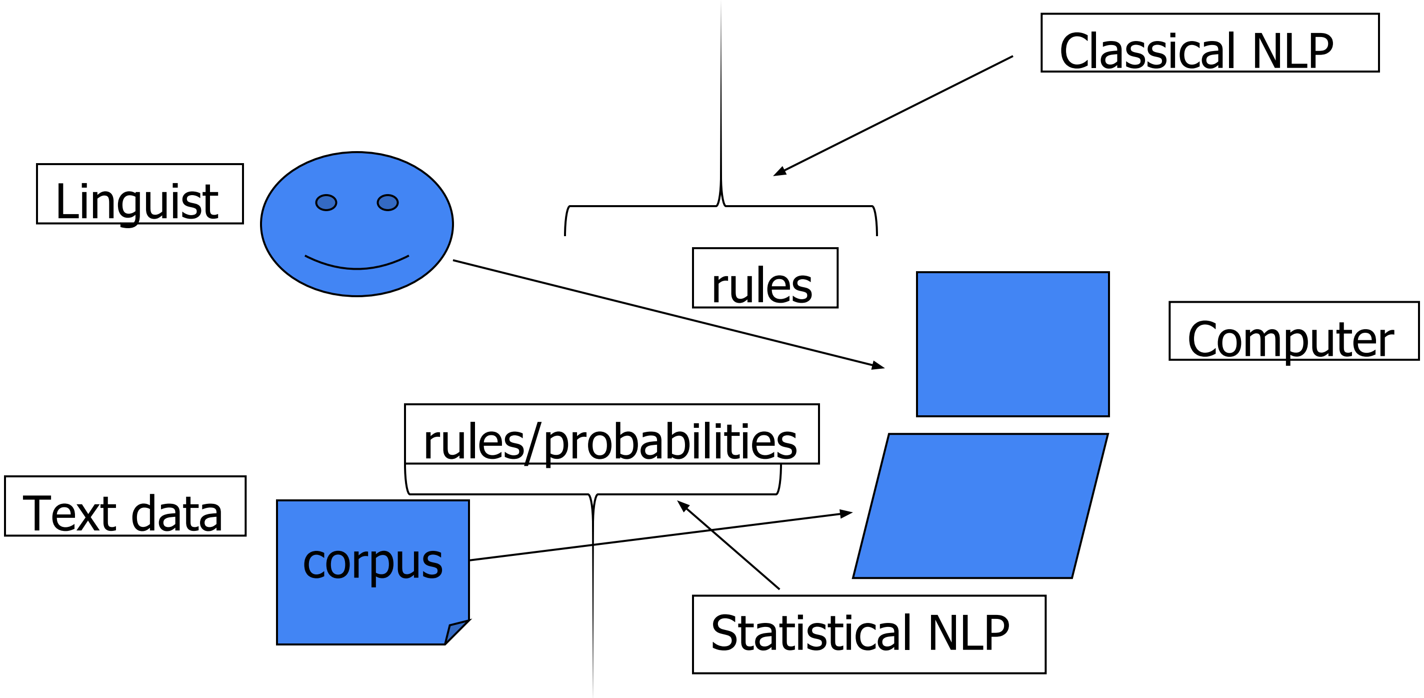
What is the output of an ML-NLP System *(1/2)*

- Option 1: A set of rules, *e.g.*,
  - *If the word to the left of the verb is a noun and has animacy feature, then it is the likely* ***agent*** *of the action denoted by the verb.*
    - *The child broke the toy* (*child* is the agent)
    - *The window broke* (*window* is not the agent; inanimate)

What is the output of an ML-NLP System *(2/2)*

- Option 2: a set of probability values
  - *P(agent|word is to the left of verb and has animacy) > P(object|word is to the left of verb and has animacy)> P(instrument|word is to the left of verb and has animacy) etc.*

How is this different from classical NLP



Classification appears as sequence labeling

A set of Sequence Labeling Tasks: *smaller to larger units*

- *Words*:
  - Part of Speech tagging
  - Named Entity tagging
  - Sense marking
- *Phrases*: Chunking
- *Sentences*: Parsing
- *Paragraphs*: Co-reference annotating

Example of word labeling: POS Tagging

<s>

Come September, and the UJF campus is abuzz with new and returning students.

</s>

<s>

Come_VB September_NNP ,_, and_CC the_DT UJF_NNP campus_NN is_VBZ abuzz_JJ with_IN new_JJ and_CC returning_VBG students_NNS ._.

</s>

Example of word labeling: Named Entity Tagging

<month_name>
September
</month_name>

<org_name>
UJF
</org_name>

Example of word labeling: Sense Marking

Word	Synset	WN-synset-no
come	*{arrive, get, come}*	*01947900*
	.	
	.	
	.	
abuzz	*{abuzz, buzzing, droning}*	*01859419*

Example of phrase labeling: Chunking

Come July, and $\boxed{\text{the UJF campus}}$ is

abuzz with new and returning students .

Example of Sentence labeling: Parsing

$[_{S1}[_{S}[_{S}[_{VP}[_{VB}$Come$][_{NP}[_{NNP}$July$]]]]$

$[_{,}$,$]$

$[_{CC}$ and$]$

$[_{S}$ $[_{NP}$ $[_{DT}$ the$]$ $[_{JJ}$ UJF$]$ $[_{NN}$ campus$]]$

$[_{VP}$ $[_{AUX}$ is$]$

$[_{ADJP}$ $[_{JJ}$ abuzz$]$

$[_{PP}[_{IN}$ with$]$

$[_{NP}[_{ADJP}$ $[_{JJ}$ new$]$ $[_{CC}$ and$]$ $[_{VBG}$ returning$]]$

$[_{NNS}$ students$]]]]]]$

$[_{.}$.$]]]$

Handling labeling through the Noisy Channel Model

w

$(w_n, w_{n-1}, \dots, w_1)$ $(t_m, t_{m-1}, \dots, t_1)$

Sequence *w* is transformed into sequence *t*.

Bayesian Decision Theory and Noisy Channel Model are close to each other

- Bayes Theorem : Given the random variables A and B,

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

$P(A \mid B)$ Posterior probability

$P(A)$ Prior probability

$P(B \mid A)$ Likelihood

Corpus

- A collection of text called *corpus*, is used for collecting various language data
- With annotation: more information, but manual labor intensive
- Practice: *label automatically; correct manually*
- The famous *Brown Corpus* contains 1 million tagged words.
- **Switchboard:** very famous corpora 2400 conversations, 543 speakers, many US dialects, annotated with orthography and phonetics

Example-1 of Application of Noisy Channel Model: Probabilistic Speech Recognition (Isolated Word)[8]

- **Problem Definition : Given a sequence of speech signals, identify the words.**
- **2 steps :**
  - **Segmentation (Word Boundary Detection)**
  - **Identify the word**
- **Isolated Word Recognition :**
  - **Identify W given SS (speech signal)**

$$\hat{W} = \arg\max_{W} P(W \mid SS)$$

Identifying the word

$$\hat{W} = \arg\max_{W} P(W \mid SS)$$
$$= \arg\max_{W} P(W)P(SS \mid W)$$

- *P(SS/W)* = likelihood called "phonological model " ☐ intuitively more tractable!
- *P(W)* = prior probability called "language model"

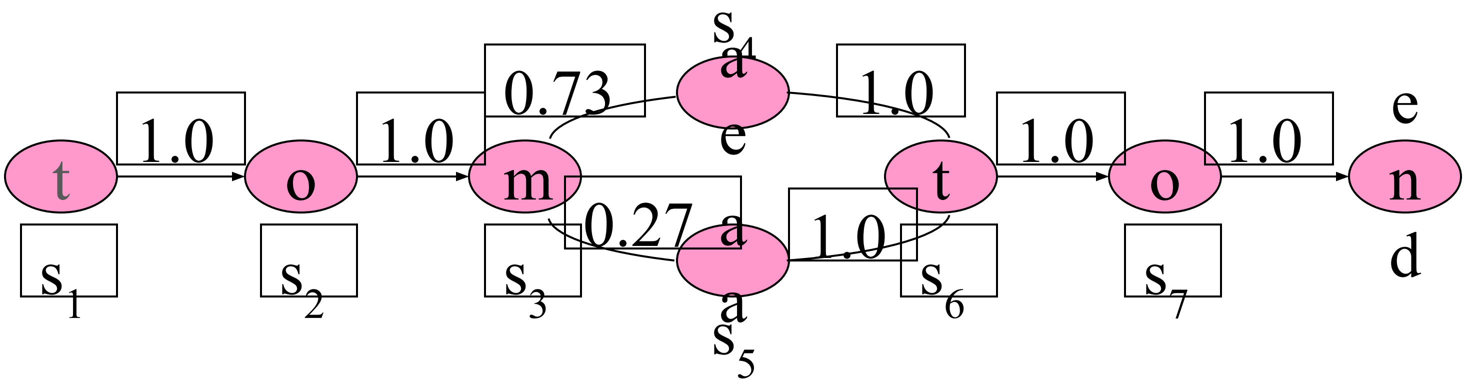
$$P(W) = \frac{\text{\# W appears in the corpus}}{\text{\# words in the corpus}}$$

Pronunciation Dictionary

Pronunciation Automaton

Word

Tomato



- *P(SS|W)* is maintained in this way.
- *P(t o m ae t o |Word is "tomato")* = Product of arc probabilities

Discriminative vs. Generative Model

$$W^* = \underset{W}{\operatorname{argmax}}\ (P(W|SS))$$

Discriminative Model

Compute directly from $P(W|SS)$

Generative Model

Compute from $P(W).P(SS|W)$