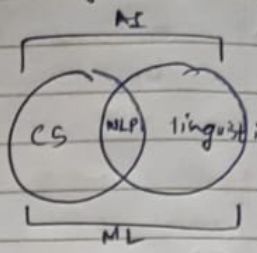
NLP

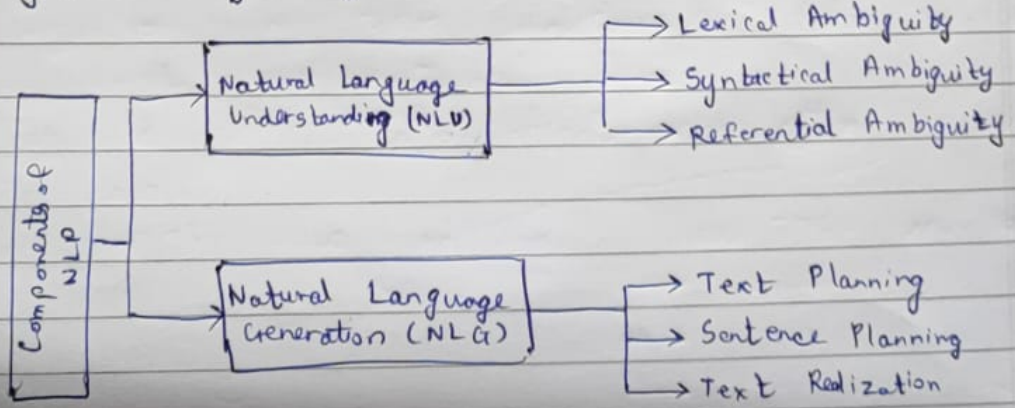
NLP is field of AI & CS where ML & linguistics both are comprised.
Used for unstructured data.



* Applications of NLP -
  - speech recognition.
  - Recommendation
  - Summarization
  - Categorization
  - Social Network Analysis.

* Stages of NLP -
  - Phonetics & (study of sound).
  - Morphology (Word level analysis, study of morphen)
  - Lexical analysis
  - Syntax analysis (Part of speech tagging)
  - Semantic analysis (Word sense disambiguity removal).
  - Pragmatic & Discourse (Pronoun resolution).

* Generic diagram of NLP -



NLG

It is process of producing meaningful phrases & sentences in form of natural lang. for some internal representation.

* Statistical approach to hand ambiguity.
  1) Probabilistic
  2) Part of speech.
     - Rule Based
     - Markov model
     - Maximum entropy
     - HMM based tagger
  3) Machine Learning Approach

Challenges of NLP-
Breaking sentence
Tagging pos & generating dependency graphs
Building appropriate vocabulary
Word sense disambiguation.
Pronoun resolution
Extracting named entities.

Type is unique token / concept and token is number of words in a sentence.
Type to token ratio indicate how often a new word is used.
High TTR - news headline low TTR - conversation.

* ~~Zipf~~ Zipf Law -

Its a statistical distribution in certain dataset such as words in linguistic corpus in which frequency of words are inversely proportional to its rank.

$\text{Freq} \propto \frac{1}{\text{rank}}$ $M = \frac{1}{\sqrt{\text{Rank}}}$

* Heaps Law -

It describes number of distinct words in a document as function of document level.

$$W1 = k N^{\beta}$$

$k = 10 - 100$ no. of tokens

$\beta = 0.4 - 0.6$.

Q. In a corpus, a word with rank 4 has freq. of 600 What will be rank with freq 300

→ Using zipf's law. $\text{Freq} \propto \frac{1}{r}$ $\therefore \frac{f_1}{f_2} = \frac{r_2}{r_1}$

$\therefore \frac{600}{300} = \frac{r_2}{4}$ $\therefore r_2 = 8$

In a sentence "the only thing we have to fear is fear itself" Token to type ratio.

Tokens = 10 Type = 9 Token to type = $\frac{10}{9} = 1.11$

Let Rank of 2 words w_1 & w_2 be 1600 & 400 resp. Let M_1 & M_2 represent ~~no~~ no. of meaning of w_1 & w_2. What is ratio of $M_1 : M_2$.

$\frac{R_1}{R_2} = \frac{1600}{400}$ $M \propto \frac{1}{\sqrt{R}}$ $\therefore \frac{M_1}{M_2} = \frac{\sqrt{R_2}}{\sqrt{R_1}}$

$\therefore \frac{M_1}{M_2} = \frac{1}{2}$

Lexical Ambiguity -
It occurs when word carries different sense i.e. having more than one meaning & sentence can be interpreted differently depending on its correct sense. It can be resolved using part of speech tagging.
eg.
Will will will will's will.

Syntactical Ambiguity -
When we see more than one meaning in a sequence of words. It is termed as gramatical ambiguity.
eg. X met Y & Z. They went to restaurant.

Referential Ambiguity -
Very often text mentions as an entity, and then refers to it again, possibly in diff. sentence, using another word. Pronouns can cause ambiguity.
eg. Boy told his father about theft. He was very upset.

What is size of unique words in a doc where total words = 1200 $k = 3.71$ & $\beta = 0.69$

$|V| = k N^{\beta}$ i.e. $= 3.71 \times (1200)^{0.69}$

$\therefore |V|$ i.e. $= 494.32 \approx 494$

Module 2

Word Level Analysis

* Morphology -

Morphology is study to understand how word is formed.
Morphen is smallest unit of word also called as stem word.

Morphens are of 2 type -

1) Free Morphen -

Free morphens are independent word having its own meaning. Eg. yellow, and, or.

2) Bound Morphen -

The morphens which do not have own meaning.
Eg. suffix, prefix.
They are of 2 types

a) Inflectional -

It is a morphological process in which if a word gets combined with free morphen it will not change part of speech
Eg. a prefix.

Derivational -

In this if a word is combined with free morphen it will change part of speech.
Eg. suffix.

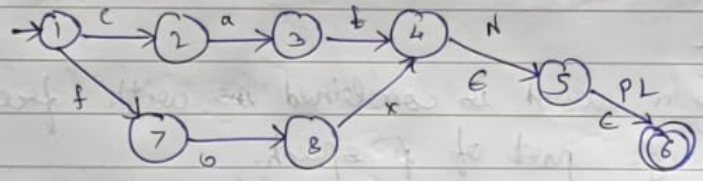
FST

input	Morphological output
cats	cat + N + PL
cat	cat + N + SG
geese	goose + N + PL
caught	catch + V + past
foxes	fox + N + PL

FST is used to build morphological analyser.
It translates strings from one language to another.
It has two levels -

1) Lexical
2) surface.

C	A	T	N	PL
C	A	T	S	



Ngram model -

Ngram model is a language model to compute probability of sequence of words in a sentence it uses n-1 words of prior context.

Q.1. Design trigram model to predict probability of full sentence.

<s> Michael & Jack played at the playground </s>

from given corpus -

<s> The school was open </s>

<s> Michael & Jack went to the school. </s>

<s> The playground at the school was huge </s>

<s> Bob & Jack played at the playground </s>

<s> Bob, Michael & Jack were friends </s>

$P(\text{michael} \mid \langle s \rangle) = \frac{1}{5}$ $\left\{ \frac{\text{Michael after } \langle s \rangle}{\text{Total } \langle s \rangle} \right\}$

$P(\& \mid \langle s \rangle \text{ Michael}) = \frac{1}{1}$ $P(\text{Jack} \mid \text{Michael \&}) = \frac{2}{2} = 1$

played | & Jack at

$P(\text{~~& Jack played~~}) = \frac{1}{3}$ $P(\text{~~Jack play~~} \mid \text{Jack played}) = \frac{1}{1}$

$P(\text{the} \mid \text{played at}) = \frac{1}{1}$ $P(\text{playground} \mid \text{at the}) = \frac{1}{2}$

$P(\langle /s \rangle \mid \text{the playground}) = \frac{1}{2}$

$P(S) = \frac{1}{5} \times 1 \times \frac{1}{3} \times \frac{1}{3} \times 1 \times 1 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{60}$

Perplexity -
Best language model is the one that predicts unseen text. It can be evaluated based on perplexity.
Lower the perplexity better the model.
It is inverse probability of text data, normalized by number of words.

$$\text{Perplexity} = P(S)^{-1/n} = \left(\frac{1}{60}\right)^{-1/9}$$
$$= (60)^{1/9}$$
$$= 1.576$$

Bigram.

$P(\text{Michael} \mid \langle s \rangle) = \frac{1}{5}$ $P(\& \mid \text{Michael}) = \frac{2}{2}$

$P(\text{Jack} \mid \&) = \frac{3}{3}$ $P(\text{played} \mid \text{Jack}) = \frac{1}{3}$

$P(\text{at} \mid \text{played}) = \frac{1}{1}$ $P(\text{the} \mid \text{at}) = \frac{2}{2}$

$P(\text{playground} \mid \text{the}) = \frac{2}{5}$ $P(\langle /s \rangle \mid \text{playground}) = \frac{1}{2}$

$$P(S) = \frac{1}{5} \times 1 \times 1 \times \frac{1}{3} \times \frac{1}{2} \times \frac{2}{5} \times 1 \times \frac{1}{2}$$
$$= \frac{1}{75}$$

$$\text{Perplexity} = \left(\frac{1}{\frac{1}{75}}\right)^{-1/9} = (75)^{1/9} = 1.6156$$

Find the total count of unique bigram for which likehood would be estimated.

<s> Alice went went to the cafe </s>
<s> Bob was waiting for Alice. </s>
<s> Alice & Bob went to the museum </s>

<s> Alice	<s> Bob	<s> Alice
Alice went	Bob was	Alice &
went to	was waiting	& Bob
to the	waiting for	Bob went
the cafe	for Alice	went to
cafe </s>	Alice </s>	to the
		the museum
		museum </s>

Unique bigrams = 17

Consider same corpus as Q.1 Design bigram & check which sentence has highest prob.

<s> Michael played at the playground </s>
<s> Bob went to the school. </s>
<s> The school was huge. </s>
<s> Jack went to the playground. </s>

<s> Michael played at the playground </s>

Bigram

$P(\text{Michael} \mid \langle s \rangle) = \frac{1}{5}$ $\quad P(\text{played} \mid \text{Michael}) = 0$

~~P = 0~~ $P(S) = 0$

<s> Bob went to the school </s>

$P(\text{Bob} \mid \langle s \rangle) = \frac{2}{5}$ $\quad P(\text{went} \mid \text{Bob}) = 0$

$P(S) = 0$

<s> The school was huge </s>

$P(\text{The} \mid \langle s \rangle) = \frac{2}{5}$ $\quad P(\text{school} \mid \text{The}) = \frac{3}{5}$

$P(\text{was} \mid \text{school}) = \frac{2}{3}$ $\quad P(\text{huge} \mid \text{was}) = \frac{1}{2}$

$P(\langle /s \rangle \mid \text{huge}) = \frac{1}{1}$

$P(S) = \frac{2}{5} \times \frac{2}{5} \times \frac{2}{3} \times \frac{1}{2} \times 1 = \frac{2}{25}$

<s> Jack went to the playground </s>

$P(\text{Jack} \mid \langle s \rangle) = 0$

$P(S) = 0$

* Laplace smoothening -

When a particular bigram or trigram does not occur in corpus data, the probability of that word will be 0. When prob of any word will be 0, the overall contribution of other words will also be 0. To overcom this we use laplace smoothening, also called as add one.

~~Step 1~~ Given same lang. model & corpus consider laplace smoothening & find prob of <s>Michael played at the playground </s>

Unique bigrams -

<s> the	<s> Michael	the playground
the school	Michael &	playground at
school was	& Zack	at the
was open	Zack went	was huge
open </s>	went to	huge </s>
school </s>	to the	
<s> Bob	Bob Michael	
Bob &	Zack were	
Zack played	were friends	
played at	friends </s>	
playground </s>		

Unique = 25

$P(\text{Michael} \mid \langle s \rangle) = \frac{1+1}{5+26} = \frac{2}{31}$ $P(\text{played} \mid \text{Michael}) = \frac{1}{28}$

$P(\text{at} \mid \text{played}) = \frac{2}{27}$ $P(\text{the} \mid \text{at}) = \frac{3}{28}$

$P(\text{playground} \mid \text{the}) = \frac{3}{31}$ $P(\langle /s \rangle \mid \text{playground}) = \frac{2}{28}$

$$P(S) = \frac{2}{31} \times \frac{1}{28} \times \frac{2}{27} \times \frac{3}{28} \times \frac{3}{31} \times \frac{2}{28}$$

$$P(S) = 1.264 \times 10^{-7}$$

Q.4 For a corpus maximum likelihood estimate (MLE) for bigram "battery life" is 0.27 & freq. of word battery is 800. After applying laplace smoothening, the MLE for "battery life" is 0.025. What is vocabulary size of corpus.

MLE is a method to estimate parameter of an assumed prob. distribution, given some observed data. Value that makes the observed data is called most probable estimated data.

$$P(w_i \mid w_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}$$

without laplace

$$P(\text{life} \mid \text{battery}) = \frac{\text{Count}(\text{battery, life})}{\text{Count}(\text{battery})}$$

with laplace.

$$P(\text{life, battery}) = \frac{\text{Count}(\text{battery,}}{\text{Count}(\text{battery}}$$

$$0.025 = \frac{0.27 \times 800 + 1}{800 + \text{uniq}} \quad \therefore \quad \text{uniq} = \frac{0.27 \times 800 + 1}{0.025} - 800$$

$$= 8680 - 800$$

$$\therefore \text{uniq.} = 7880$$

* Spelling correction -

Minimum added distance between 2 strings using insertion, deletion, ~~substitute~~ & substitution with cost 1, 1, 2.
It can be done using DP.

$$D(i,j) = \min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1) + \begin{cases} 2, & s_i \neq s_j \\ 0, & s_i = s_j \end{cases} \end{cases}$$

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	11
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	9	10	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

↑ j

→ i

$$D(1,1) = \min \begin{cases} D(0,1)+1 = 2 \\ D(1,0)+1 = 2 \\ D(0,0)+2 = 2 \end{cases} = 2$$