pronounced or written. Hence for the linguistic corpra, parale and performance data set is pratical.

* Such corpora are a finite collection of linguistic data that are studied with empirical methods. It can be used for comparison when linguistic models are developed.

2. What is morphology and explain morphological models.

Morphology :-

- It is a study of structure of word and word formation.

Morpheme :-

- The smallest unit of meaning full word.

Type of morpheme :-

* Free morpheme
* Bound morpheme.

Freea morpheme :

- Word can stand by themself. It has own meaning. Standalone.

Example :- boy, girl, car, beauty.

Types :-
* Lexical morpheme
* Functional morpheme.

* Lexical morpheme :-
- It classify the words using parts of speech.

Example :- POS.

* Functional morpheme :-
- It built the class of the words using part of Speech. That have grammatical function.

Example :-
Preposition - of, to
conjunction - but, and.
determiners - this, that
pronoun - I, U, verb - is, can

Bound morpheme :-
- Words cannot stand alone by themselfs.
* It only occur as part of word.
* It must connected to other morpheme to create a word.
* Both derivational and influction morpheme are bound morpheme

Example :- Boys, pictures

The " s " suffix in boys & prict pictures is an example of bound morpheme.

1 **Derivational morpheme :-**

- It change the category of the word and its grammatical.

Types :-

* Class changing
* Class maintaining.

Class changing :-

- Produce a derived form of another class.

Ex:- Teacher, development, national,

- The above example (er, ment) are class changing suffixes.
- In teach → teacher

teach ↓ verb

teacher ↓ noun.

- In teacher, a verb teach has become a noun after suffixing " er ".

Class maintaining :-

- To produce a derived form of same class.
- They do not change there class of a POS

Example :- Boyhood, childhood, principalship.
The above example (hood, ship) are class maintaining derivational suffixes.

Inflection morpheme :-
It change a word interm of grammer but does not create a new word.
. Never change the grammatical category of word.

-skip- (base form) skipping
-skipped.

Types :- There are 2 types.
* regular inflection morpheme
* irregular inflection morpheme.

Morphological model.

Domain Specific Language (DSL) :-
* A domain specific language (DSL) is a specialized programming language that is used for a single purpose.
* Various domain-specific languages have been created for achieving intuitive and minimal programming effort.

* Pragmatically, a DSL may be specialized to a particular problem domain, a particular problem representation technique, a particular solution technique, or other aspects of a domain.
* Examples of such domain-specific programming languages are HTML, SQL AWK etc.,..

<u>Dictionary lookup as morphological model.</u>

* Morphological model needs a systems in which analysing a word form in reduced kept in sync with more sophisticated model of the language.
* A dictionary is understood as a data structure that directly enables obtaining some precomputed results.
* The data structure can be optimized for efficient lookup.
* Hence dictionary lookup is constructed as one of the effective morphological model.

* Finite-state morphological :-

* Finite-state morphological models are the morphological models in which the specifications written by humans programmers.
* The finite state morphological model can be used for multiple natural languages.
* The tools used are XFST and Lex tools.

3. Explain sentence boundary & topic boundary?

- In human language, words and sentences do not appear randomly but usually have a structure.
- For example, combinations of words from sentences - meaningful grammatical units. such as statements, requests and commands.
- Likewise in written text, sentences form paragraphs - self-contained units of discourse about a particular point or idea.
- Document structure help in breaking apart the input text or speech into topically coherent blocks that provides better organization and indexing it.

Sentence boundary :
- It is detection is the problem in natural language processing of deciding where sentences begin and end.
Sentence detection is an important task, which should be performed at the beginning of a text processing pipeline.
- Sentence boundary detection deals with automatically segmenting a sequence of word tokens into sentence units.
- Natural language processing tools often languages, the beginning of a sentence; however, sentence boundary identification can be challenging due to the potential ambiguity of punctuation marks.
- In written text in english and some other languages, the beginning of a sentence is usually marked with an uppercase letter, and the end of a sentence is explicitly marked with a period (.), a question mark (?), an exclamation marks or another type of punctuation.
- However, in addition to their role as sentence boundary, capitalized inital letters.

Topic boundary segmentation :-

* Topic segmentation (sometimes called discourse or text segementation) is the task of automatically dividing a stream of text or speech into topically homogeneous blocks.
* That is, given a sequence of words, the aim of topic segmentation is to find the boundaries where topics.
* Topic segmentation is an imporant task for various language-understanding applications such as information extraction and retrieval and text summarization.
* In information retrieval, if long documents can be segmented into shorter, topically coherent segments, then only the segment that is about the user's query could be retrievd.
* Topic segementation is a nontrivial problem without a very high human agreement because of many natural language-related issues and hence requires a good definition.

Code switching : that is, the use of words, phrases or sentences from multiple language by multilingual speakers - is another problem that can affect the characteristics of sentences. For example, when switching is a different language, the writer can either keep the punctuation rules from the first language or resort in the code of several language.

* Code switching also affects technical texts form which the meaning of puncutations sign can be redefined is in Uniform Resource Location (URLs)

Hence code switching is considered as a problem in sentence boundary detection.

Q) Explain boundary classification problem?

* Sentence segmentation and topic segmentation have mainly been considered as a boundary classification.
* For a given boundary candidate the goal is to predict whether or not the candidate is an actual boundary.
* Let $x \in X$ be the vector of features associated with a candidate.
* And $y \in Y$ be the label predicated for that candidate.
* The label y can be b(yes) for boundary and $\bar{b}$(no) for non boundary.
* Alternatively to the binary classification problem it is possible to model boundary types finer grained categories.
* Gillick suggested that sentence segmentation in text be framed as a three-class problem sentence boundary with an abbreviation b^a and $b^{\bar{a}}$.

In this we have 2 methods :-
(i) Generative model
(ii) Discriminative model.

for Topic segementation typically instead of two n states are used,
When i...

Generative sequence model :-

* It estimate the joint distribution of the observation, $P(X, Y)$ and the labels.

* It requires specific assumptions and have good generalization properties.

Discriminative sequence model :-

* It focus on features that characterize the differences btw the labeling of the examples.

* Such methods can be used for sentence and topic in both written and spoken language.

The probability is written using the Bayes rule:

Baye's theorem

* In probability theory & statistics, bayes theorm describes the probability of an event

$$\boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

Naive Bayes classifer

$$\boxed{P(y|x_1, \ldots x_j) = \frac{P(x_1, \ldots x_j|y)\, P(y)}{P(x_1, \ldots x_j)}}$$

Where,

(i) $x_1 --- x_j$ are j features that are idependent of each other.

(ii) $p(y | x_1, --- x_j)$: Posterior probability.

(iii) $p(x_1 --- x_j | y)$: Likelihood of features x_1 to x_j given.

(iv) $p(y)$: prior probabity

(v) $p(x_1 --- x_j)$: marginal probability.

Example :-

* Let's understand how does the algorithm works to the following steps.

Step 1 :

* We start by importing database and necessary depends.

* The dataset of weather includes the features (outlook, temp, humidity, windy).

* Target variable "play".

* Now we need to predict whether the players will play (or) not.

* Based on given weather conditions.

Outlook	Temp	Humidity	Windy	Play
Rainy	hot	high	f	no
Rainy	hot	high	t	no
overcast	hot	high	f	yes
Sunny	mild	high	f	yes
Sunny	cool	normal	f	yes
Sunny	cool	normal	t	no
overcast	cool	normal	t	yes
Rainy	mild	high	f	no
Rainy	cool	normal	f	yes
Sunny	mild	normal	f	yes
Rainy	mild	normal	t	yes
overcast	mild	high	t	yes
overcast	hot	normal	f	yes
sunny	mild	high	t	no

step 2 :-

*Calculate prior probability of classes

p(y)

yes = 9 No = 5 total = 14

p(yes) = 9/14 = 0.642

p(no) = 5/14 = 0.35

step 3 :-

*Calculate the likelihood table for all features.

Likelihood table

(i) outlook

play	overcast	rainy	sunny
yes	4/9	2/9	3/9
No	0/5	3/5	2/5
=	4/14	5/14	5/14

(ii) Temperature

play	cool	mild	hot
yes	3/9	4/9	2/9
no	1/5	2/5	2/5
	4/14	6/9	4/14

(3) Humidity

play	high	normal
yes	3/9	6/9
No	4/5	1/5
	7/14	7/14

(4) windy

play	F	T
yes	6/9	3/9
no.	2/5	3/5
	8/14	6/14

* whether the players will play or not when the wether condition are
outlook = rainy temp = mild humidity = normal
windy = true.

Calculation of postior probability.

i) for yes.

$$P(Y=yes/x) = P(yes / Rainy, mild, normal, true)$$

$$= \frac{P(Rain, mild, normal, true / yes) * P(yes)}{P(Rainy, mild, normal, true)}$$

$$= \frac{(2/9)(4/9)(6/9)(3/9) * (9/14)}{(5/14)(6/14)(7/14)(6/14)}$$

$= 0.43$

ii) for no

$$p(y=no/x) = P(no / Rainy, mild, normal, true)$$

$$= \frac{P(rain, mild, normal, true / no) * P(no)}{P(rain, mild, normal, true)}$$

$$= \frac{(3/5) * (2/5) * (1/5) * (3/5) * (5/14)}{(5/14) * (6/14) * (7/14) * (6/14)}$$

$= 0.31$

The probability for yes is more than no

∴ They can play on that day.

play = yes.

5) Discuss about hybrid approaches for word classification, complexity of approaches, performance of approaches & ?

* Nonsequential discriminative classification algorithms typically ignore the context, which is critical for the segmentation task.
* While we may add context as a feature or simply use CRFs, which inherently consider context, these approaches are suboptimal when dealing with real-valued features such as pause duration or pitch range.
* An alternative is to use a hybrid classification approach as suggested by Shribreg et al.
* The main idea is to use the posterior probability $P_c(y_i|x_i)$, for each boundary candidate, obtained from the other classifers such as boosting or CRF. by simply converting them to state observation likelihoods by dividing to their ~~pho~~ priors following the well-known Bayes rule as follows.

$$\arg_{y_i} = \frac{P_c(y_i \mid x_i)}{P(y_i)} = \arg\max_{y_i} P(x_i \mid y_i)$$

Complexity of approaches.

* Sentence topic segementation approaches can be rated in terms of complexity of their training and prediction algorithm and in terms of performance.

* Discriminative approach

* In terms of complexity, training of this approach is more complex than training of generative once because they require multiple passes.

* Generative models

* This models such as HELMs can handle multiple orders of magnitude lager longer training sets and benefit for instance from decades of news wire transcripts.

* Discriminative classifers

* They allow for a wider variety of features and perform better on smaller training sets.

Sequence approaches.

* Compared to local approaches, sequence approaches bring the additional complexity of decoding: finding the best sequence of decisions requires evaluating all possible sequences.

* Fortunately, conditional independence assumptions allow the use of dynamic programming to trade time for memory and decode ~~t~~ in polynomenal time.

Performance of the approaches.

a) Sentence segmentation in speech

- for sentence segmentation in speech, performance is usually evaluated using,
  * The ~~errot~~ error rate.
  * F1 - measure.

b) Sentence segmentation in text.

- for sentence segmentation in text, researchers have reported error rate results on a subset of the wall street journal corpus of about 27,000 sentences.
- For instance, Mikheer reports that his

rule based system performs at an error rate of 1.41%.

- Without requiring handcrafted rules or on abbreviations list, Gillick's SVM-based system obtains even fewer errors at 0.25%

c) Sentence segmentation in speech.

- For sentence segmentation in speech, Doss et al. report on the mandarin TDT4 multilingual Broadcast news speech corpus, on F1-measures using the same set of features is as of

* 69.1% for a maxent classifier
* 72.6% with adaboost
* 72.7% with SVMs.

- A combination of the three classifiers using logistic regression is also proposed.