01-11-23
Wednesday

UNIT-I

Words and their components

1. Tokens
2. Lexemes
3. Morphomes
4. Typology

① **Tokenization**: It is a way of separating a piece of text into smaller units called Tokens

* Token may be either words, characters or subwords

Ex: NLP is sub branch of AI

Tokens = 6.

② **Lexemes**:

The set of alternate forms that can express a word are called lexical items

Ex: The lexem "play" can have many forms such as playing, play, plays, etc

③ **Morphomes**:

* It is smallest unit of a word provides a specific meaning to the string of letters
* There are two types of morphomes

(i) Free morphome:-

un+happy = unhappy

play+ing = playing

(ii) Bound morphomes:

Dis+place+ment = Displacement

Un+employ+ment = unemployment

im+prison+ment = imprisonment

④ Typology:-

* Typology is a study, the ways in which the language of the world vary in their patterns
* It is concern with discovery what grammatical patterns are common to many languages. they are specified according to three criteria,

(i) Geneological families
(ii) Areal families
(iii) Geographical families

(i) Geneological families:

This classification indicates the historical connections between the languages and it uses the historical and linguistic criteria as a bases

Ex: In Europe the Basque language is called language isolates, as it cannot relate to any other language

(ii) Areal families:

These kinds of languages are creating language unions such as Balkan language union, encompassing Macedonian, Bulgarian, Serbian, Albanian languages

(iii) Geographical families:

These covers the languages that are close by and have developed similar characteristics In terms of structure

03/11/23 Friday

Issues and challenges

① Irregularity
② Ambiguity
③ Productivity

① Irregularity :-

Irregularity means existence of words and structures that are not described appropriately by prototypical linguistic model.

Ex:- eat ate eats eating eaten

↓ ↓ ↓ ↓ ↓

~~V1(past)~~ past(V_2) eat+s eat+ing eat+en

① watch watches watching watched

↓ ↓ ↓

V_1 watch+ing watch+ed

② write wrote writing written

↓ ↓ ↓

V_2 V_1+ing V_3

② Ambiguity:

Types

① Lexical (semantic) [meaning of the word]

② syntantical (structural)

③ Anaporic

④ Pragmatic

① Lexical Ambiguity:

The ambiguity of a single word is called Lexical Ambiguity (or) ~~when~~ when the meaning of the word ~~can be~~ themselves misinterpreted

(or)

semantic ambiguity happens when the sentence contains an ambiguous words or phrases

Ex: watch (having 2 meanings).

① wrist watch
② watching movie

Ex: 2
Live - (Live match)
Live - (existing)

Ex: 3 plant - (planting trees)
plant - (solar plant, steel plant, water plant)
bat - (cricket bat)
bat - (mammal)

The car hit the pole while it was moving

* The sentence is having two meanings,
① the car, while moving hit the pole
② the car hit the pole, while pole was moving.

② Syntactical Ambiguity:

This type of ambiguity occurs when sentence is parsed in different ways

Ex: The man saw the girl with the telescope.

* It is ambiguous whether the man saw the girl carrying telescope or he saw her through his telescope

Ex: This book costs a large amount

Ex: The chicken is ready to eat

Ex:- Ram bought green shirts & shoes

③ Anaporic Ambiguity:

This kind of ambiguity arises due to the use of Anapora entities

Ex: The horse ran up the hill. It was steep. It got tired

Here the anaporic reference "it is ambiguous"

Ex: Jyoshna invited sudeshna for a visit, but she told her, she had to go to work

Ex: I am allergic to tomatoes. Also fish

④ Pragmatic Ambiguity:

This type of ambiguity occurs when phrases gives multiple interpretation.

Ex:- I like you too.

Multiple interpretations are I like you just you like me

(or) I like you someone else do that.

~~④ Pragmatic Ambiguity~~:

Ex:- Mother : what you want to eat

son : If icecream is good this time of year

Ex: I tried to reach her and on the phone, but she didnot answer.

Productivity:

1. Set objectives
2. Boost staff morale
3. Better communication
4. Learning and development
5. changing behavior

06/11/23
Monday

NLP Tools and techniques

1. Monkey Learn (diagnosis, call centers)
2. IBM watson
3. Google cloud NLP API
4. NLTK (many methods & this are implemented)
5. spacy
6. pytorch

NLP techniques

1. Lemmatization
2. Lexicon
3. NER
4. N-gram

1. Lemmatization:

Lemmatization is a process of normalize a word and remove its inflection form to arrive at the base word

Ex: baseword for gone or went is go
for better & best → good

write = wrote · written

base word

(2) Lexicon:

Lexicon is the vocabulary of a language or subject along with its usage

Ex:- medical terminology, computer[default, web, shutdown, application]

(3) NER: [~~Name~~ Named entity recognition]

NER is a popular technique used in information extraction to identify and segment the named entities and classify or categorize them under various pre-defined clauses

Ex: Real world objects (people, places, organization name)

Delhi (N) is ~~in~~ the capital (N) of India (N)

sundar pichhai is the CEO of Google

Name of the person → organization name

sudeshna belongs to kadapa

name place

(4) N-gram

N-gram is a contiguous sequence of N-tokens from a given text

* N-gram based models credit the most probable words that might follow the entered text sequence

Ex: Applications · including machine translation, auto-complete

Corpus:

* Corpus in NLP means collection of text or other digital data ~~text~~ across languages sets

Ex: Wikipedia

10/11/23
Friday

Morphological models (study of internal structure of the word)

① Dictionary Lookup (collection of meanings, words, adverbs, adj, prons)
② Finite state morphology (having nodes + edges / states / arcs)
③ Unification based morphology
④ Functional morphology
⑤ Morphology Induction

Functional
Number marker – Three cats
Posessive marker – John's book
Inflection – she walks

* Morphological model is the study of internal structure of word.

① Dictionary Lookup:

* Dictionary Lookup is the fundamental concept in NLP
* That involves searching for a word or phrase in a dictionary or lexicon to retrieve its definition, translation, parts of the speech or ~~reliable~~ relevant information
* It is often used to extract meaning of the texts

(a) Lexicon / dictionary:

* In NLP, a dictionary or lexicon is a collection of words or phrases along with associated information such as meaning, part of the speech, pronunciation, translation and more.

* Dictionaries can be simple lookup tables or complex databases

* Purpose of dictionary lookup:

* Dictionary lookup is used to provide context and understanding to words, all phrases in the text

* It can be used for various tasks like word sense disambiguition:

(1) Identifying a correct meaning of the word based on the context

(2) Language translation:- converting words or phrases from one language to another

(3) Parts of speech:- Assigning the appropriate part of the speech to word like noun, verb, etc.

Information retrieval:

* Retrieving additional information such as synonyms, antonyms or definitions.

Named Entity Recognition: Identifying named entities like names of people, places or organizations

Ex: Let consider a word bank and perform dictionary lookup

word: Bank

Part of speech: Noun

def: Financial Institute

Synonyms: FI, credit union

Antonyms: borrowers, debtor

word: bank

pos: Noun

def: side of river, lake

Synonyms: shore, edge

Antonyms: interior

word: Bank

pos: verb

def: to deposit money in FI

Synonyms: deposit, invest

Antonyms: withdraw, disburse

11/11/23

Saturday

② Finite state morphology:

Finite state morphology is a computational linguistic approaches used to model and analyze the morphology of a languages in NLP.

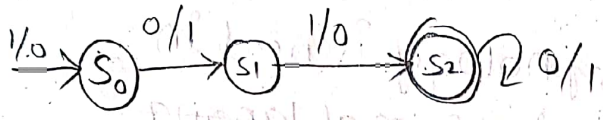
* It focus on the use of finite state machines [FSM] and finite state transducer [FST] to represent and manipulate the structure and variation of words in a language

FSM:

FSM's are computational models that consists of states, transitions and input symbols.

In the context of morphology state represents linguistic forms or morphemes

* Transitions represents transformations between these forms and input symbols represents the letters or sounds of language



FST:

* FST's are the finite state machines that can map input symbols to output symbols.
* They are commonly used in finite state morphology to represent relationship between the surface form and root form.

Ex: playing = play + ing
root surface

Purpose of Finite state Morphology:

* It is used for various linguistic tasks including stemming, lemmatization, inflection and understanding the internal structure of words
* It is particularly valuable for languages with complex inflectional processes, where words can change to base form according to their grammatical features

Ex:- ~~Noun~~ Non-pluralization

(i) singular nouns ends in -s, -o, -x and to pluralize them we add -es

Ex: bus, box, mango

(ii) singular noun ending in a consonant followed by y changes to ies Ex: baby (consonant) Monkey (vowel) ✗
babies ✓ Monkeys ✗

(3) Unification based morphology

* It employs unification as a key concept which is a process of merging or matching linguistic features and constraints to derive word forms
* Unification based morphology is used in computational linguistic, formal ~~langua~~ grammar

Unification linguistics:

* Unification is the process of merging linguistic features and constraints from different sources to create an unified representation of a linguistic object such as word.
* In this approach, features and constraints are represented using feature structures.
* which can be taught of attribute ~~phase~~ pairs

Features and constraints:

* Features and constraints describes various aspect of a words such as its lexical form, grammatical category and so on.
* Constraints specify the rules and restrictions on how features can be combined.

Purpose of unification based morphology

* The goal of unification based morphology is to model the structure and formation of words.

Ex: Features and constraints for noun pluralization

1) Features:

Lexical Form (LF): The base form of the noun is cat

Number (Num): The grammatical number of a noun [singular (or) plural]

Constraints:

Constraint 1: If the number is singular and Lexical form ends with -s, -o, -x

Constraint-2: If the number is singular, lexical form ends with y

Rule: adjective + er = comparative adjective

Ex: fast + er = faster

(4) Functional morphology:

* Morphomes are analyzed based on their roles within sentences such as how they contribute to meaning, syntax and structure
* This approach consider the functional semantic, pragmatic aspects of morphomes

Ex: (i) Number marker - ("Three cats")
the plural s indicates three numbers of cats.

(ii) Possessive marker -
In the phrase - ("John's book")
s serves as the book belongs to John.

(iii) Verbal Inflection -
In the sentence - ("she walks")
s indicates the third person singular present tense

Morphology Induction:

* Aims to automatically identify and extract morphological patterns or units from a corpus as text, without prior knowledge of the language morphology. This process involves discovering how words are formed, segmented, and inflected in a language

① Unsupervised learning: Morphology induction typically uses unsupervise machine learning approaches, such as statistical models or algorithm, to analyze a large corpus of text and identify patterns within it

② Identifying morphomes:

The goal is to identify morphomes, which are the smallest meaningful units in a language, morphomes can include prefixes, suffixes, roots and other components that contribute to word formation and meaning.

③ clustering and segmentation

It involves clustering similar words forms together and segmenting them into their constituent morphomes based on statistical or distributional patterns. This process often involves grouping words that share similar prefixes, suffixes or root.

Ex:

Unhappiness, rebuild and undo

In morphological induction process, the algorithm identify patterns such as,

un - is a prefix indicates negation

-ness - is a suffix indicates quality or state

re - is a prefix that indicates doing something again

build and do are root words

the algorithm might group these words

1. clusters 1 : unhappiness, unhappy, unhappily

Prefix : un -

suffix : -ness, -ly

cluster 2: rebuild, rebuilding, rebuilt

pre :- re -

Root : build

clusters : undo, undid, undoing

pre :- un

Root :- do

cluster 4 : happy, happier, happiest

pre : happy

Root :- -ier, -iest