Assignment - I

1. Explain about words and their components.

- words is most languages are the smallest linguistic units that can form a complete utterance by themselves.
- Three important terms which are integral part of a word are :-

Phonemes:- the distinctive units of sound in spoken language

Graphemes :- the smallest unit of a written language which corresponds to a phoneme

Morphemes :- the minimal part of a word that delivers aspects of meaning to the word

Tokens

Tokens are the building blocks of natural language

Tokenization is a way of seperating a piece of text into smaller units called tokens.

Here tokens can be either words, characters, or subwords.

Lexemes

There are a lot of alternative forms that can be expressed for a given word

such set are called lexemes or lexical items.

They constitute the lexicon of a language

Lexemes are divided by their lexical categories such as verb, noun, adjective, adverb etc

The citation form of a lexeme by which it is identified is called lemma.

morphemes :-

The structural components that associate the properties of word forms are called morphs.

- The morphs that by themselves represent some aspect of the meaning of a word are called morphemes of some function.

4 Typology :-

- Morphological typology divides languages in groups. Here we outline the typology that is based on quantitative relations between words, their morphemes and their features.

Isolating or analytic languages include no or relatively few words that would comprise more than one morpheme

Synthetic languages can combine more morphemes in one word and are further divided into agglutinative and fusional language

agglutinative languages have morphemes associated with only a single function at a time.

Fusional languages are defined by their feature per morpheme ratio higher than one

In accordance with the notions about word formation processes mentioned earlier

concatenative languages linking morphs and

Irregularity

Morphological Parsing provides generalization and abstraction in the world of words.

Irregular morphology can be seen as enforcing some extended rules - the nature of which is phonological, over the underlying or prototypical regular word forms.

the irregular verbs in english tend to take different forms in the past or in the present. participle depending on the origin of the word.

Ambiguity:-

words forms that look the same but has distrinct functions or meaning are called homonyms.

Ambiguity is present in all aspects of morpholugical processing and language processing at large

Morphological parsing is not concerned with complete disambiguation of words in their context. however it can effectively restrict the set of valid interpretations of a given word form.

- the morphological phenomenon that some words or word classes show instance of systematic homonymy is called syncretism.
- neutralization is about syntactic irrelevance being reflected in morphology.
- uninflectedness is about morphology being unresponsive to a feature that is syntactically relevant

Productivity

- In one view language can be seen as simply a collection of utterances actually pronounced or written
- This data set can be the linguistic copora a finite collection of linguistic data
- the members of the corpus are the word types
- the original instances of the word form is the word token
- The negation is a productive morphological operation in some language.

Explain about the different morphological models.

morphological parsing is a process by which word forms of a language are associated with corresponding linguistic descriptions

There are many approaches to designing and implementing morphological models

A lot of domain specific programming languages have been created that can be very useful in implementing theoritical problems with minimal programming effort.

Dictionary Lookup.

A dictionary ~~lookup~~ is a data structure that directly enables obtaining precomputed word analysis.

Dictionaries can be implemented as lists, binary search trees, tries, hash tables etc.

Dictionaries enumerate the set of associations between word forms and their desciptions.

the generative power of the language is not exploited when implemented in the form of a dictionary

-the problem with dictionary based approach is how the associated annotations are constructed and how informative and accurate they are.

Finite state morphology

these are directly compiled into finite state transduuers.

The set of possible sequences accepted by the transducer defines the input language and the set of possible sequences emitted by the transducer defines the output language.

Morphological operations and processes in human languages can be expressed in finite state terms

Finite state tools can be used to a limited extent in morphological analyzers or generators.

Unification Based Morphology

Unification based approaches to morphology are inspired by two things:

The formal linguistic frameworks like head driven phase structure grammar

Languages for lexical knowledge representation like DATR

The concepts and methodologies of these formalisms are closely connected to logic programming

Morphological models of this kind are typically formulated as logic programs and unification is used to solve the system of constraints imposed by the model.

Functional morphology

It defines its models using principles of functional programming and type theory.

It treats morphological operations and processes as pure mathematical functions and organizes the lnguistic as well as abstract elements of a model into distinct types of values and typeclasses.

Functional morphology can be used for the implementation of:

Morphological parsing

Morphological generation

Lexicon browsing etc

Along with parsing described in the previous section we canalso describe Infection I, derivation D as lookup L as functions of these generic types:

I : lexeme → {Parameter} → form

D : lexeme → {parameter} → {lexeme}

L : content → {lexeme}

until now the focous is on finding the structure of words in diverse languages supposing we know what we are looking for.

we now consider the problem of discovering and inducing word structure without the human insight.

There are several challenges issues abot deducing the word structure just from the forms.

Explain about the different methods finding the structure of documents.

As we all know words form sentences.
sentences can be related to each other by explicit discourse connectives such as therfore
sentences form paragraphs
Automatic extraction of structure of documents help in:
Parsing
machine translation
semantic role labelling,
Here we discuss about two topics:-
sentence boundary detection:- The task of deciding where sentences start and end given a sequence of characters
Topic segmentation:- The task of determining when a topic starts and ends in a sequence of sentences.
Features of input are local characteristics that give evidence toward the presence or absence of a sentence or a topic boundary
such as:
punctuation mark
A pause in a speech
A new word in a document

Sentence Boundary detection

Sentence boundary detection deals with automatically segmenting a sequence of word tokens into sentence units.

In written text in english and a few languages the beginning of a sentence usually marked with an upper case letter and the end of a sentence is marked with:

a period (.), a question mark (?) and an exclamation mark (!)

capitalized letters — distinguish proper nouns

periods — used in abbrevations & numbers

other punctuation marks — used inside proper names.

Human participants when tried to re punctuate mono case texts performed at an F1 measure of about 80% which shows how difficult the task is.

code switch also affects technical texts for which the punctuation signs can be redefined

sentence segmentation can be stated as classification problem.

Topic Boundary Detection

Topic segmentation is the task of automatically dividing a stream of text or speech into topically homogeneous blocks

Topic segmentation is an important task for applications like:

Information extraction and retrieval

Text summarization

Topics are not typically flat but occur in a semantic hiearchy.

It is difficult to segment the text into a predefined number of topics

In text, topic boundaries are usually marked with distinct segmentation cues like headlines and paragraph breaks.

speech provides other cues such as pause duration and speaker change.