SHAMBHUNATH INSTITUTE OF ENGINEERING AND TECHNOLOGY

Subject Code: RCS 085 Subject: SPEECH AND NATURAL LANGUAGE PROCESSING

B.Tech.: IV year SEMESTER: VIII

FIRST SESSIONAL EXAMINATION, EVEN SEMESTER, (2019-2020)

Branch: Computer Science & Engineering

Time –1hr 30 min Maximum Marks – 30

SOLUTION

SECTION – A

Define NLP, NLU & NLG.

Natural language processing (NLP) starts with a library, a pre-programmed set of algorithms that plug into a system using an API, or application programming interface. Basically, the library gives a computer or system a set of rules and definitions for natural language as a foundation.

Natural language understanding (NLU) is a smaller part of natural language processing. Once the language has been broken down, it's time for the program to understand, find meaning, and even perform sentiment analysis.

The task of Natural Language Generation (NLG) is to generate natural language from a machine-representation system such as a knowledge base or a logical form. To simplify this, NLG is like a translator that converts data into a "natural language representation", that a human can understand easily.

Name any two general NLP libraries?

1- NLTK: A general library for NLP written in python.
2- OpenNLP: A library written in JAVA that implement many different NLP tools.
3- Stanford CoreNLP: A library including many of the NLP tools developed at Stanford.

What is ELIZA?

ELIZA is an early natural language processing computer program created from 1964 to 1966 at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum. ELIZA was one of the first chatterbots and one of the first programs capable of attempting the Turing test.

NLTK platform is used for building programs in which language?

Python

List some Java tools for training NLP models?

1- OpenNLP

2- StanfordNLP
3- CogCompNLP

SECTION – B

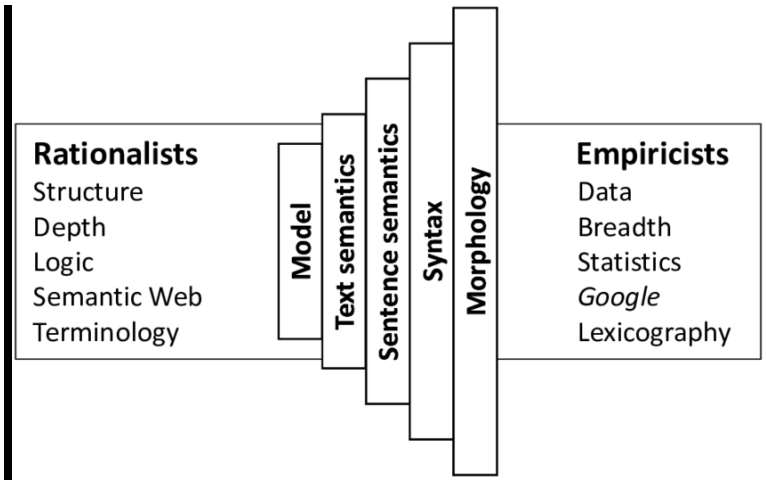
Explain the Named entity recognition (NER)?
Named entity recognition (NER) , also known as entity chunking/extraction , is a popular technique used in information extraction to identify and segment the named entities and classify or categorize them under various predefined classes. SpaCy has some excellent capabilities for named entity recognition.
NER systems have been created that use linguistic grammar-based techniques as well as statistical models such as machine learning. Hand-crafted grammar-based systems typically obtain better precision, but at the cost of lower recall and months of work by experienced computational linguists . Statistical NER systems typically require a large amount of manually annotated training data. Semi-supervised approaches have been suggested to avoid part of the annotation effort.

Why we use Regular expression?
A regular expression is a method used in programming for pattern matching. Regular expressions provide a flexible and concise means to match strings of text. For example, a regular expression could be used to search through large volumes of text and change all occurrences of "cat" to "dog".

Regular expressions are used for syntax highlighting systems, data validation and in search engines such as Google, to try to determine an algorithmic match to the query a user is asking.

Regular expressions are also known in short form as regex or regexp.

Differentiate between the rationalist and empiricist approaches to natural language processing?



Rationalists generally develop their view in two ways. First, they argue that there are cases where the content of our concepts or knowledge outstrips the information that sense experience can provide. Second, they construct accounts of how reason in some form or other provides that

additional information about the world. Empiricists present complementary lines of thought. First, they develop accounts of how experience provides the information that rationalists cite, insofar as we have it in the first place. (Empiricists will at times opt for skepticism as an alternative to rationalism: if experience cannot provide the concepts or knowledge the rationalists cite, then we don't have them.) Second, empiricists attack the rationalists' accounts of how reason is a source of concepts or knowledge.

What makes natural language processing difficult?

In a normal conversation between humans, things are often unsaid, whether in the form of some signal, expression, or just silence. Nevertheless, we, as humans, have the capacity to understand the underlying intent of the conversation, which a computer lacks. A second difficulty is owing to ambiguity in sentences. This may be at the word level, at the sentence level, or at the meaning level.

Ambiguity at Word Level :

Consider the word won't. There is always an ambiguity associated with the word. Will the system treat the contraction as one word or two words, and in what sense (what will its meaning be?).

Ambiguity at Sentence Level :

Consider the following sentences: Most of the time travelers worry about their luggage. Without punctuation, it is hard to infer from the given sentence whether "time travelers" worry about their luggage or merely "travelers." Time flies like an arrow.

The rate at which time is spent is compared to the speed of an arrow, which is quite difficult to map, given only this sentence and without enough information concerning the general nature of the two entities mentioned.

Ambiguity at Meaning Level :

Consider the word tie. There are three ways in which you can process (interpret) this word: as an equal score between contestants, as a garment, and as a verb.

SECTION – C

3-(a) Why we model a language? Discuss main approaches used for language modeling?

Language Modeling (LM) is one of the most important parts of modern Natural Language Processing (NLP). Language model is required to represent the text to a form understandable from the machine point of view. *Language Modeling (LM)* is one of the most important parts of modern Natural Language Processing (NLP). There are many sorts of applications for Language Modeling, like: Machine Translation, Spell Correction Speech Recognition, Summarization, Question Answering, Sentiment analysis etc. Each of those tasks require use of *language model*. Language model is required to represent the text to a form understandable from the machine point of view.

1. Rule-based

Rule-based approaches are the oldest approaches to NLP. Why are they still used, you might ask? It's because they are tried and true, and have been proven to work well. Rules applied to text can offer a lot of insight: think of what you can learn about arbitrary text by finding what words are nouns, or what verbs end in -ing, or whether a pattern recognizable as Python code can be identified. Regular expressions and context free grammars are textbook examples of rule-based approaches to NLP.

Rule-based approaches:

- tend to focus on pattern-matching or parsing
- can often be thought of as "fill in the blanks" methods
- are low precision, high recall, meaning they can have high performance in specific use cases, but often suffer performance degradation when generalized

2. "Traditional" Machine Learning
"Traditional" machine learning approaches include probabilistic modeling, likelihood maximization, and linear classifiers. Notably, these are not neural network models (see those below).

Traditional machine learning approaches are characterized by:

- training data - in this case, a corpus with markup
- feature engineering - word type, surrounding words, capitalized, plural, etc.
- training a model on parameters, followed by fitting on test data (typical of machine learning systems in general)
- inference (applying model to test data) characterized by finding most probable words, next word, best category, etc.
- "semantic slot filling"

3. Neural Networks
This is similar to "traditional" machine learning, but with a few differences:

- feature engineering is generally skipped, as networks will "learn" important features (this is generally one of the claimed big benefits of using neural networks for NLP)
- instead, streams of raw parameters ("words" -- actually vector representations of words) without engineered features, are fed into neural networks
- very large training corpus

Specific neural networks of use in NLP include recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

4. The grammar-based approach traditionally implies that a human is involved in the process of stepwise system development and improvement. The biggest advantage of formal grammar is that there is always a way to check whether the system could process a query placed by a user and how it could do that. And since all the rules are written by people, any reported bug is easy to localize and fix by adjusting the rules in the related module.

Grammar rules can be developed in a very **flexible** manner, for example, through the extension of translation rules and synonyms base, and can easily be updated with new functions and data types, with no significant changes to the core system. This approach to query analysis is based on the development and extension of the existing rules, so the system **doesn't require a massive training corpus**, compared to the machine learning-based approach.

3-(b)The applications utilizing NLP includes Speech Recognition and Speech Synthesis: explain their features with example.
we have looked at some essential linguistic concepts, we can return to NLP. Computerized processing of speech comprises

• speech synthesis
• speech recognition.
One particular form of each involves written text at one end of the process and speech at the other, i.e.
• text-to-speech or TTS
• speech-to-text or STT.
There are also applications where text is not directly involved, such as the reproduction or synthesis of fixed text, as in the 'speaking clock', or recognition which involves selecting from audio menus via single word responses.
TTS requires the conversion of text into semi-continuous sounds. The most obvious approach seems to be to store digitized sound sequences representing complete words or phrases, and then simply select and output these sounds. For those who travel by train, a good example of this approach is the automated announcement system used at many UK railways stations, including New Street Station.1 This approach has the advantage of simplicity and can also produces very natural-sounding speech; however it also has several serious drawbacks.
• In a limited domain (e.g. speaking weather forecasts or train announcements), it is feasible to construct a complete dictionary giving the digitized pronunciations of all relevant words or phrases. However, this is impossible for general text, since natural languages contain a very large number of words, many of which occur only rarely, so that a huge dictionary would be required. Furthermore, natural languages are constantly changing, with new words being introduced (especially in the names of commercial products). Such words could not be handled by a system relying on pre-stored sounds.
• Natural speech does not consist of separate words with pauses between them. 'Blending' words together is essential for natural sounding speech. In most languages, the correct choice of allophone at a word boundary depends on the neighbouring word. For example, in SEE, /r/ is not pronounced unless immediately before a vowel. Thus in the sentence tear the paper in half, the /r/ in tear is not pronounced. However in the phrase tear it in half, it is. This phenomenon means that combining pre-recorded words or phrases into new combinations can seriously lower the quality of the output.
Speech-to-text (STT) is considerably more difficult than the reverse. The simplest approach is to require the user to speak in discrete words and then attempt to recognise them 'whole' (perhaps by digitizing and matching with a dictionary of stored sounds). With sufficiently slow speech, variation in pronunciation caused by phonological rules operating across word boundaries should be minimal. A major problem of course is that speaking in this way is highly unnatural. Whole word recognition can also be applied to continuous speech. It requires considerably more processing, since the position at which a match should be attempted is unknown, so that multiple matches will be needed. Furthermore to take into account variations in pronunciation caused by neighbouring words, some 'looseness' will be required in the matching process.
Speech waveform (electrical signal) Extracted features (usually at fixed time intervals) 'Unit' likelihoods (probability of input being each of the possible 'units') Words Signal processing, analogue and/or digital Recognition of 'units', e.g. by NNs Statistical models (N-grams, Hidden Markov Models)

4-(a) In Information retrieval which NLP techniques are most commonly used?

Text Embeddings are real valued vector representations of strings. We build a dense vector for each word, chosen so that it's similar to vectors of words that appear in similar contexts. Word embeddings are considered a great starting point for most deep NLP tasks.

Machine Translation is the classic test of language understanding. It consists of both language analysis and language generation. Big machine translation systems have huge commercial use, as global language is a $40 Billion-per-year industry. To give you some notable examples:

- Google Translate goes through 100 billion words per day.
- Facebook uses machine translation to translate text in posts and comments automatically, in order to break language barriers and allow people around the world to communicate with each other.
- eBay uses Machine Translation tech to enable cross-border trade and connect buyers and sellers around the world.
- Microsoft brings AI-powered translation to end users and developers on Android, iOS, and Amazon Fire, whether or not they have access to the Internet.
- Systran became the 1st software provider to launch a Neural Machine Translation engine in more than 30 languages back in 2016.

Dialogue and Conversations

A lot has been written about conversational AI, and a majority of it focuses on vertical chatbots, messenger platforms, business trends, and startup opportunities (think Amazon Alexa, Apple Siri, Facebook M, Google Assistant, Microsoft Cortana). AI's capability of understanding natural language is still limited. As a result, creating fully-automated, open-domain conversational assistants has remained an open challenge.

4-(b) Differentiate between Natural language and formal language with suitable example.

Formal languages are very different from natural languages: Natural languages exist in the real world, in flesh-and-blood communities of language users. The grammar of a natural language like English is incredibly complex. We discover the grammar of natural language through empirical investigation.PL is a formal language, an artificial language**.** The grammar of an artificial language like PL is incredibly simple. We don't discover the grammar of an artificial language, we stipulate it — we define it however we want. The properties of formal languages are mathematically stipulated. Formal languages exist as mathematical abstractions. Metaphysically, they have more in common with computer code than real-world spoken language.

We use formal languages as simplified models of natural language English has conjunctions, disjunctions and conditionals.

"John and Mary went to the store" expresses an English language conjunction.

"Mary will walk the dog if John agrees to maker dinner" expresses an English language conditional. But the semantics of natural language is often so complex and varied that it's not always clear how the words "and" or "if-then" are actually being used, or what semantic and syntactic rules they're actually following. Is the "and" functioning as a logical conjunction, or some other kind of conjunction? What do we even mean by a "logical conjunction"?

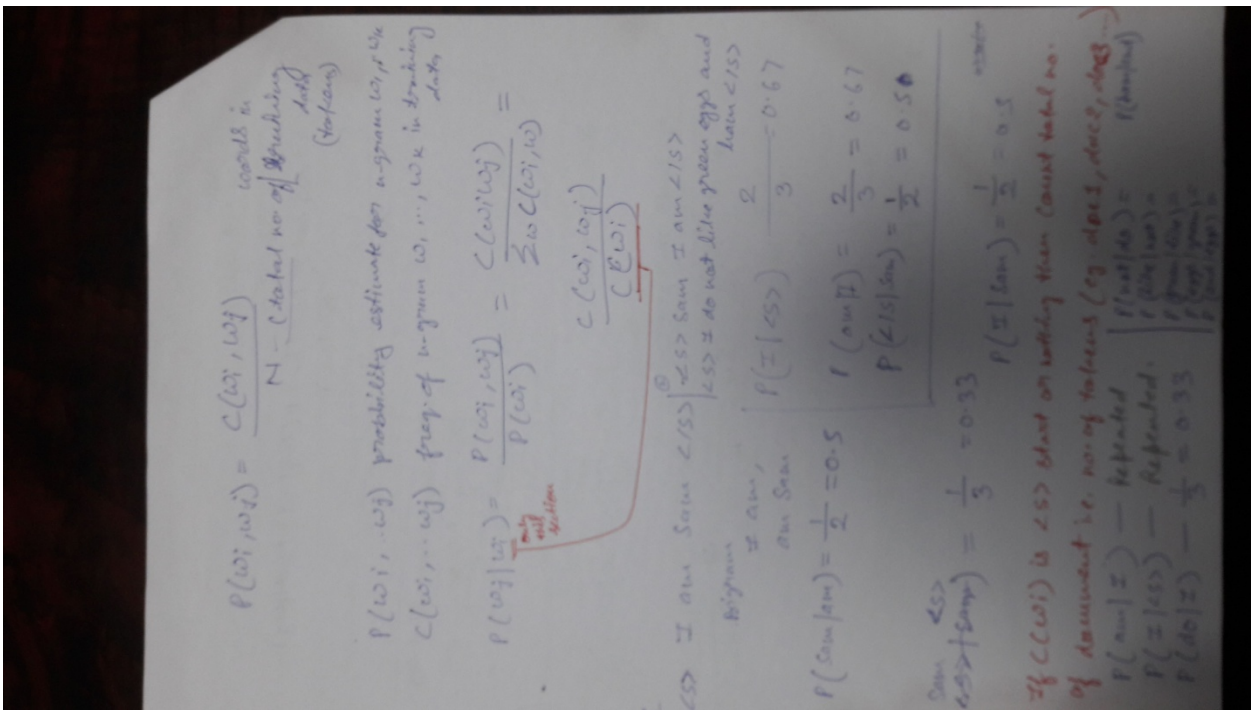
This is why we invent formal languages like PL — to help us answer these kinds of questions. Any given formal language is designed to represent or model the logical behavior of a select few natural language words. When we use it we abstract away from all other features of natural language sentences. PL, for example, was designed to model the logical behavior of conjunctions ("and"), disjunctions ("or"), conditionals ("if-then") and negations ("not-"). That's it. The whole apparatus of PL exists in order to model the logical properties of a small handful of natural language words.
Symbols help us to distinguish the model from what is being modeled.
The expressive capacity of natural language far outstrips what any formal language can represent

4- (a)Calculate MLE (maximum likelihood estimation) using bigram probabilistic model for a given toy corpus :

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>



5-(b) What are the differences between Indian languages and English?

Indian Languages (IL) refers to those varieties of English that developed on the Indian subcontinent. IL is currently the co-official language of India with Hindi, and it is the primary medium of education, law, media, and business throughout India. IL is also used for social interactions and in pan-Indian literature. A small minority of Indians are members of a community that has IL as a native language.

1. Pronunciation: Indian English has retroflex consonants and dental ones too. There are only two alveolars which are both fricatives. The dental fricatives become stops; the

voiceless one is aspirated in the North. Aspiration of stops is lost. There is a long traditional cot-caught merger where the latter's pronunciation had survived. Both rhotic and non-rhotic accents exist and sometimes co-exist in the same dialect. The prepositions of and off are pronounced the same. Some diphthongs are shortened. The consonant v and the approximant w are allophones.

2. Grammar: Indian English is more conservative about the grammar than most other dialects, though some standard sentences may be ungrammatical elsewhere.
3. Vocabulary: Indian English has seen a recent shift in preference of American vocabulary over British, though the change is gradual and sometimes resisted. Some Indian vocabulary is drawn from local languages and usually non-standard.
4. Culture: Indian English is usually accompanied by multilingualism and is therefore more versatile. There are honorifics and some words like uncle and aunt(y) are not usually familiar.
5. Phonology: In comparison with English Hindi has approximately half as many vowels and twice as many consonants. This leads to several problems of pronunciation. One difficulty is distinguishing phonemes in words such as said / sad; par / paw; vet / wet, etc. Words containing the letters th (this, thing, months) will cause Hindi learners the same kind of problems that they cause most other learners of English. The phoneme / ʒ / as exemplified by the s in pleasure is missing in Hindi and so pronunciation of such words is difficult. Consonants clusters at the beginning or end of words are more common in English than Hindi. This leads to errors in the pronunciation of words such as straight (istraight), fly (faly), film (filam).