Total No. of Questions – [06]

MAY 2022 - ENDSEM EXAM

B. TECH. (Computer) (SEMESTER - I)

COURSE NAME: Natural Language Processing

COURSE CODE: CSUA40182C

Question Paper Solution

Time: [1 Hour] [Max. Marks: 30]

Q.1) a) Identify correct lexical relations in the following examples.

1)Happy, joyful, glad - Synonymy

2)Happy, sad- Antonymy

3) world record, a record of conversation - Polysemy

4) book ticket, bring me a book - Homonymy

b) Classify relation between word meaning and give example for each. [6 marks]

There are similar relationships between words. Words that mean the same thing but look different are called synonyms. Their meanings are very similar (e.g., pretty/cute). An antonym is a word that has the opposite meaning of another word (e.g., pretty/ugly). A homonym is a word that sounds like another word but has a different meaning (e.g., there/their).

Synonymy: two senses of two different words (lemmas) are identical, or nearly identical, we say the two senses are synonyms. synonym Synonyms include such pairs as couch/sofa vomit/throw up filbert/hazelnut car/automobile

Antonymy: Whereas synonyms are words with identical or similar meanings, antonyms are words with an opposite meaning, like: long/short big/little fast/slow cold/hot dark/light rise/fall up/down in/out Two senses can be antonyms if they define a binary opposition or are at opposite ends of some scale. This is the case for long/short, fast/slow, or big/little, which are reversives at opposite ends of the length or size scale. Another group of antonyms, describe change or movement in opposite directions, such as rise/fall or up/down. Antonyms thus differ completely with respect to one aspect of their meaning—their position on a scale or their direction—but are otherwise very similar, sharing almost all other aspects of meaning. Thus, automatically distinguishing synonyms from antonyms can be difficult

hyponym : A word (or sense) is a hyponym of another word or sense if the first is more specific, denoting a subclass of the other. For example, car is a hyponym of vehicle, dog is a hyponym of animal, and mango is a hyponym of fruit. Conversely, we say that vehicle is a hypernym of car, and animal is a hypernym of dog. It is unfortunate that the two words (hypernym) are very similar and hence easily confused; for this reason, the word superordinate is often used instead of hypernym.

Meronymy : part-whole Another common relation is meronymy, the relation. A leg is part of a chair; a wheel is part of a car. We say that wheel is a meronym of car, and car is a holonym of wheel.

OR

Q.2) a) Distributional semantics (DS), also known as vector space semantics, is a usage-based model of meaning, based on the assumption that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behavior. Its main focus is the lexicon: DS is primarily an empirical method for the analysis of lexical meaning (but see Section 5.2 for distributional models of compositional semantics). DS offers both a model to represent meaning and computational methods to learn such representations from language data. Given the ever-increasing availability of digital texts,

distributional models can rely on huge amounts of empirical evidence to characterize the semantic properties of lexemes. Distributional representations are built from text corpora as samples of language usage and offer new ways to investigate the interplay between meaning and contexts, and to tackle the dynamicity and plasticity of meaning

The basic method of building distributional vectors consists of the following procedure:

- co-occurrences between lexical items and linguistic contexts are extracted from a corpus and counted;
- the distribution of lexical items is represented with a co-occurrence matrix, whose rows correspond to target lexical items, columns to contexts, and the entries to their co-occurrence frequency;
- raw frequencies are then usually transformed into significance weights to reflect the importance of the contexts; and
- the semantic similarity between lexemes is measured with the similarity between their row vectors in the co-occurrence matrix.

Suppose we have extracted and counted the co-occurrences of the targets $T = \{bike, car, dog, lion\}$ with the context lexemes $C = \{bite, buy, drive, eat, get, live, park, ride, tell\}$ in a corpus. Their distribution is represented with the following co-occurrence matrix $M_{T\times C}$, in which m_{ic} is the co-occurrence frequency of t with c:

$$(1)\quad \begin{array}{c} \\ bike \\ car \\ dog \\ lion \end{array} \begin{array}{c} \begin{array}{ccccccccc} bite & buy & drive & eat & get & live & park & ride & tell \end{array} \\ \begin{pmatrix} 0 & 9 & 0 & 0 & 12 & 0 & 8 & 6 & 0 \\ 0 & 13 & 8 & 0 & 15 & 0 & 5 & 0 & 8 \\ 0 & 0 & 0 & 9 & 10 & 7 & 0 & 0 & 1 \\ 6 & 0 & 0 & 1 & 8 & 3 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Matrix 1, with rows labeled with target lexemes and columns with context lexemes, is called a word-by-word co-occurrence matrix. A co-occurrence matrix whose columns are labeled with text regions is referred to as a word-by-region matrix.

b) Discuss Word Sense Disambiguation (WSD) for word 'bark' in following example.

"Cinnamon comes from the bark of the Cinnamon tree." and "The dog barked at the stranger." [4 marks]

Cinnamon comes from the bark of the Cinnamon tree."

"The dog barked at the stranger."

Let us now try a sentence with both words:

"The dog was scratching the bark of the tree, when the man approached the dog to make it stop, the dog barked."

Suppose, this sentence is passed to an algorithm for sentiment analysis, "bark" and "barked" might mean the same meaning.

So, we can understand that, same words can mean differently based on the usage of the word in a particular sentence. The usage of words defines a lot about their meaning. But the problem lies that, in NLP, while dealing with text data, we need some way to interpret the different words with different meanings.

Q.3) a) Model Regex for finding phone number and email [4 marks]

Finding email

"^[a-zA-Z0-9_\-]+@[a-zA-Z0-9_\-]+\.[a-zA-Z0-9_\-]"	x@y.z
"@\w+.\w+"	Domain within email

Finding telephone number

"([0-9][3]-)[2][0-9][4][^0-9]*$"	xxx-xxx-xxxx
"([0-9][3].)[2][0-9][4][^0-9]*$"	xxx.xxx.xxxx
"[0-9][10][^0-9]*$"	xxxxxxxxxx
"\\([0-9][3]\\)[0-9][3]-[0-9][4][^0-9]*$"	(xxx)xxx-xxxx

b) Write at least two features of Bootstrapping and Supervised Relation Extraction. Construct any real time example for one of the approach. [6 marks]

Bootstrapping

If you don't have enough annotated text to train on ...

- But you do have:

- some seed instances of the relation
- (or some patterns that work pretty well)
- and lots & lots of unannotated text (e.g., the web)

• Bootstrapping can be considered semi-supervised

Bootstrapping problems

Requires that we have seeds for each relation

- Sensitive to original set of seeds

• Big problem of semantic drift at each iteration
• Precision tends to be not that high
• Generally have lots of parameters to be tuned
• No probabilistic interpretation

- Hard to know how confident to be in each result

Supervised

For each pair of entities in a sentence, predict the relation type (if any) that holds between them.

The supervised approach requires:

• Defining an inventory of relation types
• Collecting labeled training data (the hard part!)
• Designing a feature representation
• Choosing a classifier: Naïve Bayes, MaxEnt, SVM, ...

Supervised approach can achieve high accuracy

- At least, for some relations
- If we have lots of hand-labeled training data

• But has significant limitations!

- Labeling 5,000 relations (+ named entities) is expensive
- Doesn't generalize to different relations, languages

• Next: beyond supervised relation extraction

- Distantly supervised relation extraction
- Unsupervised relation extraction

Q.4) a) Apply the Latent Dirichlet Allocation [LDA] on any real time application. [4 marks]

LDA has been conventionally used to find thematic word clusters or topics from in text data. Besides this, LDA has also been used as components in more sophisticated applications. Some of the applications are shown below.

Cascaded LDA for taxonomy building: An online content generation system to organize and manage a lot of contents. Details of this application are provided in Smart Online Content Generation System.

LDA based recommendation system: a recommendation engine for books based on their Wikipedia articles. Details of this application are provided in the LDA based recommendation engine.

Classification of Gene expression: Understand the role of differential gene expression in cancer etiology and cellular process Details of this application is provided in A Novel Approach for Classifying Gene Expression Data using.

b) Suppose you have various photographs (documents) with captions (words). You want to display them in a gallery so you decide to categorize the photographs on various themes (topics) based on which you will create different sections in your gallery. Assume category as k=2 (nature and city).Find out category of sentence "The tree is in front of the building and behind a car". Assume suitable data wherever required [6 marks]
Naturally, the classification isn't so clear as some photographs with city have trees and flowers while the nature ones might have some buildings in it. You, as a start, decide to assign the photographs which only have nature or city elements in them into their respective categories while you randomly assigned the rest.

A lot of photographs in nature have the word tree in their captions. So it can be concluded that the word tree and topic nature must be closely related.
Next, pick the word building and check how many photographs are in nature because they have the word building in their caption. There is no many and now are less sure about building belonging to the topic nature and associate it more strongly with the topic city.
Pick a photograph which has the caption "The tree is in front of the building and behind a car" and see that it is in the category nature currently.
Chose the word tree, and calculate the first probability p(topic t | document d): other words in the caption are building and car, most photographs having captions with building or car in it are in city, so you get a low probability.
Now the second probability p(word w| topic t): we know that a lot of photographs in nature have the word trees in it. So you get a high score here.
Update the probability of tree belonging in nature by multiplying the two. Get a lower value than before for tree in topic nature because now there is no that tree and words such as building/car in the same caption, implying that trees can also be found in cities. For the same reason, when update the probability for tree belonging in topic city, it will be noticed that it will be greater than what it was before.

Q.5) a) Analyze the impact of NLP in Education Sector [4 marks]

First of all, developments in NLP can help students learn to write better essays by providing formative feedback (i.e., actionable feedback on specific essay parts) that can be used during the revising process to improve more than just grammar and mechanics.

For instance, NLP can help identify the presence of absence of important discourse elements like claims, arguments, and evidence. In addition, NLP can provide feedback to learners about the organization of an essay. These NLP solutions can be combined into automatic writing evaluation (AWE) systems that can provide low level feedback (e.g., tips about vocabulary) or higher level feedback (e.g., advice about the cohesion of discourse).

Another limitation of NLP writing studies is that they have historically focused on independent, persuasive essays that require no background knowledge (i.e., five-paragraph essays written on general topics). However, these types of essays do not accurately reflect writing expectations in the classroom.

Thus, researchers have begun to expand the writing tasks to which NLP can be applied to include source-based essays where writers are expected to integrate external documents, narratives, summaries, and expository texts.

Beyond writing, NLP can also go a long way toward helping struggling readers in the classroom. NLP algorithms can provide automatic feedback to students about the strength of self-explanations and summaries of reading samples, both of which are key elements of reading comprehension. However, teachers often don't have the time to provide students with detailed and individualized feedback on these tasks.

Newer readability formulas based on NLP can also help educators better match texts to students to ensure reading assignments are suitably challenging and productive. NLP readability formulas can calculate more accurate readability scores that outperform traditional formulas such as Flesch-Kincaid Grade Level.

These new metrics provide information about the complexity of language in terms of vocabulary, cohesion, and syntactic complexity. Not only are they better predictors of reading speed and comprehension, but they better match cognitive models of text processing. They also work for a variety of different readers and genres. NLP techniques are even being used in text simplification algorithms that can automatically modify texts to make them better fit with the reading skills of students.

Finally, NLP can be especially useful for English Language Learners (ELLs). By providing feedback on both form and function of language components, NLP can improve the rate and quality of language acquisition.

NLP can also allow ELLs more opportunities for practice in their second language outside the classroom. For instance, online language tutors can provide feedback to learners about grammatical, syntactic, and lexical errors as well proficiency assessment. NLP can also help assess proficiency level for ELLs and track their progress over time.

b) Cross Lingual Information Retrieval represent information in same way though are in different languages. Justify. [6 marks]

Cross-lingual Information Retrieval is the task of retrieving relevant information when the document collection is written in a different language from the user query. There are many situations where CLIR becomes essential because the information is not in the user's native language.

Translation Approaches

CLIR requires the ability to represent and match information in the same representation space even if the query and the document collection are in different languages. The fundamental problem in CLIR is to match terms in different languages that describe the same or a similar meaning. The strategy of mapping between different language representations is usually machine translation. In CLIR, this translation process can be in several ways.

Document translation is to map the document representation into the query representation space.

Query translation is to map the query representation into the document representation space. Pivot language or Interlingua is to map both document and query representations to a third space. Query translation is generally considered the most appropriate approach. The query is short and thus fast to translate than the document, and it is more flexible and allows more interaction with users if the user understands the translation. However, query translation can suffer from translation ambiguity, and this problem is even more obvious for the short query text due to the limited context. By contrast, document translation can provide more accurate translation thanks to richer contexts. Document translation also has the advantage that once the translated document is retrieved, the user can directly understand it, while the query translation still needs a post-retrieval translation. However, several experiments show that there is no clear evidence of one approach or the other using the same machine translation system, and the effectiveness is more dependent on the translation direction.

OR

Q.6) a) for each application -1 mark (4 application -4 marks) [4 marks]

b) Semantic web describes the relations among data (i.e., resources) on the Web. Justify. [6 marks]

The Semantic Web is a Web of data. There is a lot of data we all use every day, and it's not part of the Web. The vision of the Semantic Web is to extend principles of the Web from documents to data. Data should be accessed using the general Web architecture using, e.g., URI-s; data should be related to one another just as documents (or portions of documents) are already. This also means creation of a common framework that allows data to be shared and reused across application, enterprise, and community boundaries, to be processed automatically by tools as well as manually, including revealing possible new relationships among pieces of data.

Semantic Web technologies can be used in a variety of application areas; for example: in data integration, whereby data in various locations and various formats can be integrated in one, seamless application; in resource discovery and classification to provide better,

domain specific search engine capabilities; in cataloging for describing the content and content relationships available at a particular Web site, page, or digital library; by intelligent software agents to facilitate knowledge sharing and exchange; in content rating; in describing collections of pages that represent a single logical "document"; for describing intellectual property rights of Web pages (see, eg, the Creative Commons), and in many others.

In order to achieve the goals described above, the most important is to be able to define and describe the relations among data (i.e., resources) on the Web. This is not unlike the usage of hyperlinks on the current Web that connect the current page with another one: the hyperlinks defines a relationship between the current page and the target. One major difference is that, on the Semantic Web, such relationships can be established between any two resources, there is no notion of "current" page. Another major difference is that the relationship (i.e, the link) itself is named, whereas the link used by a human on the (traditional) Web is not and their role is deduced by the human reader. The definition of those relations allow for a better and automatic interchange of data. RDF, which is one of the fundamental building blocks of the Semantic Web, gives a formal definition for that interchange.

On that basis, additional building blocks are built around this central notion. Some examples are:

Tools to query information described through such relationships (eg, SPARQL)

Tools to have a finer and more detailed classification and characterization of those relationships as well as the resources being characterized. This ensures interoperability, more complex automatic behaviors. For example, a community can agree what name to use for a relationship connecting a page to one's calendar; this name can then be used by a large number of users and applications without the necessity to redefine such names every time. (E.g., RDF Schemas, OWL, SKOS)

For more complex cases, tools are available to define logical relationships among resources and their relationships (for example, if a relationships binds a person to his/her email address, it is feasible to declare that the email address is unique, ie, the address is not shared by several persons). Tools based on this level (e.g., OWL, Rules) can ensure more interoperability, can reveal inconsistencies and find new relationships.

Tools to extract from, and to bind to traditional data sources to ensure their interchange with data from other sources. (E.g., GRDDL, RDFa, POWDER)