LINGUISTICS 250 – COMPUTATIONAL ANALYSIS OF TEXT

This course introduces machine learning and natural data language processing. At its core, it provides explanations behind computational text analysis and how language is interpreted by machines.

This page will break down several of the components and will keep more of a focus on the linguistics aspect and the first half of the course.

TABLE OF CONTENTS

Linguistics and computers

Linguistics is the study of language and its structure. It explains how language works within societies and communities. What LING 250 largely looks at is the computational process of language and by using corpus linguistics, we can analysis collections of spoken and written text to describe the environments around us. NATURAL LANGUAGE PROCESSING (NLP) is a subfield of linguistics and a focus of this course.

Corpus linguistics

An approach to linguistics that relies on the use of corpora and needs data-processing software to make sense of the information provided. A CORPUS is a collection of written text, such as a book or a speech. It is then turned into machine-readable corpus where a computer can then search for information and words within a text to find the context.

A key component of corpus linguistics is then being able to analysis tone to determine intention behind text (is it positive or negative?). A component of this is sentiment analysis, which will be discussed further on.

Computational linguistics

Computational linguistics uses computers to perform tasks involving human language, which as conducting an annotated corpus. When corpus is loaded with information, a computer can then tag and parse the information. This labels the data to give a full insight into what is being understood. Since a large portion of human communication is mediated by computers, it is important for computers to be able to understand tone and context behind what humans are saying. By using SEMANTICS, words and phrases are interpretated and context is discovered.

Regular Expressions

Regular expressions are a sequence of characters that will specify a search pattern in text [1]. They provide a compilation of regular expressions when the script is loaded and are used to classify and capture generalization. They aid in NLP because they can look at how many words there are to then examine the lemma and word form.

LEMMA means a word is a part of the same stem and speech, but it differs slightly. For example, ‘tree’ and ‘trees’ come from the same stem but if you google searched a photo you would see different results. WORDFORM refers to the full inflected surface form so ‘tree’ and ‘trees’ are different wordforms because they are not identical.

Text processing – Tokenization

Tasks conducted using NLP must be able to do text normalization. TEXT NORMALIZATION involves segmenting and tokenizing words in a running text. It must also normalize formats and still be able to segment sentences. A part of text normalization refers to lemma and wordform. How many words are within a given sentence? What if the words are repeated? We can use type to tell us the element of the vocabulary and token to tell us the instance of that type in a running text.

Tokenization comes with issues, such as if multiple words are used to describe one place (e.g., New York – would this count as one token or two?) and language issues. In language, not all words are segmented, and many languages use contractions to make their sentences shorter. One thing than computers can do, is use morphemes to examine the small units that make up words. STEMS are the core meaning units while AFFIXES are bits and pieces that adhere to stems. Again, the example of 'tree' and 'trees' applies. Stemming aids in reducing terms to just their stems to help in information retrieval. It removes the affixes to look at complex morpheme segmentation.

Sentence Segmentation

Typically relies on end of sentence punctuation as most forms of punctuation are unambiguous, but '.' can be interpreted differently based off its context. A period does not always mean the end of a sentence but can be used in abbreviations or numbers. In computer programming for NLP, computers are implemented with a. code that has them determine if a word Is end-of-sentence. It uses past processed information to look at new information to see what the context is. The structures used are called DECISION TREES, which are learned by machine learning from a training corpus. It's an if-then-else statement that works through a process to determine context.

N-grams

N-grams are a continuous sequence of words and symbols in a document. They can be used for sequences of words and once knowledge of counts of N-grams are given, the next words in a sequence can be guessed. Related to word prediction, we can use sources of knowledge and the preceding words to keep track of what may appear next. They can assess probability of a sequence of words and is useful in spelling correction, machine translation and more.

You may already use N-grams without even knowing. When texting on a phone, your phone will suggest the next word or phrase in an expression. Your phone has used your preceding words to take a probabilistic approach to what you may say next.

A problem with N-grams is the distance between words. When a sentence is longer, it becomes harder for the computer to follow and then make a prediction.

Language models

Computers can use conditional probability to assess. This is called a language model. It computes the probability of a sentence or sequence of words and computers the grammar. It can look for counts but it is not a perfect model. With most sentences, there are a lot of possible outcomes. Most NLP systems will use assumptions to calculate a smaller probability of what the next word could possibly be.

Sentiment Analysis

When a corpus is inputted, a numeric value is returned that expresses the subjective content. It is the sentiment of a text and will tell us the language tone used that determines how positive or negative a text is. This can be used to search for positive and negative reviews for things such as products and movies. Sentiment analysis extracts the subjectivity of a text and sees what aspects are being evaluated. When looking online at a restaurants review, computers have taken the review and analysed it to find the most common words to determine how consumers view the restaurant. This does not necessarily mean it refers to emotions, but rather polarity. Emotions are not sentiment and fall under a different classification.

A difficulty of sentiment analysis is IRONY AND SARCASM. When speaking out loud, one may say ‘the food here is sooo great.’ in a sarcastic tone. But when written out, tone is not given and computers cannot detect sarcasm. Even though when spoken the review is not great, the computer may interpret it as such. This is an interesting concept because it reminds us of the linguistics aspect. It is not easy for computers to always understand what is being said, so it is important to examine context and continue to develop NLP systems that can detect irony and sarcasm.

There are different approaches to sentiment analysis. One is LEXICON-based, where a set of dictionaries is created with words that express opinion. This categorizes words based on positivity or negativity and nuances of meanings. It can also use the word’s semantic orientation to determine its sentiment. Another approach is STATISTICAl where classifiers can be built to determine features. Machine learning will take the data, out it into a classifier. Lexicon base will take a dictionary, apply it to new data and then extract words and apply the rules. The building of dictionaries can come from intuition or use existing resources. A problem with the lexicon approach is that dictionaries are static and new domain involves creating new dictionaries. And once again, the idea of context comes back. Words can have different meanings and therefore hard to categorize. A problem with the statistical models is that the feature is often too specific and what is useful in one domain, is not useful in another. This again makes categorization hard and creates problems.

Text classification

An aspect of text classification that computers do without thinking twice is classification of spam. All email boxes have a spam folder where computers send suspicious emails, and your computer warns you before opening it. Computers also run through articles to find the subject matter and assign subject categories and identify the language and authorship of a piece.

To conduct text classification, input is put in and then the output is a predicted class. It bases its measures on combinations of words and features. Accuracy is often high as it receives consumer feedback (example in emails when a email is incorrectly sent to spam, a user will mark that it is not spam and the computer will then register this

Naive Bayes

This is a principle, simple classification method that uses a BAG OF WORDS to determine the representation of a document. It takes a corpus and places the most common words together in a 'bag' to see what appears the most. Then text can be read and classified to determine the nature of the corpus.

Parts of Speech (POS) Tagging

Parts of speech refers to traditional parts of speech such as nouns, adjectives, verbs, adverbs, pronoun, etc. What POS tagging goes is assign a part of the speech to a class marker. So it will take a sentence and determine what category each words belong to. It will state if it's a noun, verb, adjective or other. It is helpful in extracting information and providing translations. If we wanted to look at movie reviews, adjectives can be examined to see what kind of genre the movie fits into. If it's a comedy, it will use words that are more related to that genre compared to a horror movie. The reviews will be tagged differently and then represent what we could expect to see in the future.

Rule-based tagging

This is a relatively straight-forward method. By taking a dictionary, all possible words are then assigned. Rules are written to remove tags and then the correct tag is left for each word. Rules are written when using prediction. So if a work has multiple meanings, the current form of the word can be predicted by looking at the previous works and the words after. We can consider all possible sequences of tags. By using BAYES rule, the likely hood can be determined.

Kinds of probabilities

The first kind is TAG TRANSITION PROBABILITIES. This using determines that are likely to precede adjectives and nouns. The second is WORD LIKELY HOOD PROBABILITIES which looks at the most likely probability to follow.

Ethics in data analysis

Societal Bias

Language data represents society's thoughts and biases. It represents the society we have; not necessarily the society we want. Machines can contain biases because they look at current data and discover patters. Then, they can accept a spectrum of known biases to then try to address the current biases in society and technology. Machines are not biases free as they unknowingly take in biased information. Bias in data produced bias models can be discriminatory and harmful. Data does not include variables that properly capture the phenomenon we want to predict, and data may unknowingly have bias which is then reproduced.

One unknowing way machines provide bases is in sentiment classifiers. One study found that machines applied lower sentiment and more negative emotion to sentences that have African American names [2]. What this then does is further push this bias out into society where then African American names are associated with negative emotions and stereotypes.

Response Bias

Movie reviews are a popular form of computational analysis that is regularly used. When consumers give their feedback online, it is then analysed to determine the sentiment. Other people online can then read the review and determine whether they think they would like the movie. However, only a small portion of people contribute to online reviews. Their opinions and preferences are unlikely to reflect the opinions of the entire population. [3]