NLP(No DEEP) - Lab 4 - Theory Report

Eliot LECLAIR, Alex POIRON, Tom THIL, Aurelien VISENTIN

1 First principles of NLP

1.1 Phonetics and phonology

Firstly, phonetics deals with the physical properties of speech, that is, how we make and perceive sounds with our mouths and ears.

Then, phonology is concerned with how we combine these basic sounds produced by our mouths into sound units from which words are formed, depending on the language.

1.2 Morphology and syntax

First and foremost, morphology is the study of form, in linguistics it means the form of words. It is therefore the study of the internal structure of words, how they are composed and the rules by which they are formed in certain languages. Above all, it identifies morphemes, the simplest units of language that can carry lexical meaning.

Syntax is more comprehensive and examines the external relationships between words to construct sentences. It deals with the grammatical rules of languages, such as word order, which must be followed to form valid sentences from words.

1.3 Semantics and pragmatics

Semantics is the study of the meaning of languages. It identifies, for example, antonym, a word that means the opposite of another word

Pragmatics is similar in that it too studies the meaning of language, but focuses on its dependence on context. Language can have a different meaning depending on the context in which sentences are placed, and some sentences can even lose their meaning completely when taken out of context.

2 What is the difference between stemming and lemmatization ?

Lemmatization replaces words with their lemma, in other words. the original form of the word in the dictionary. Stemming, however, merely tries to recognise and cut affixes out of words.

2.1 How do they both work ?

Lemmatization uses more complex methods and requires knowledge of the language under study to work. These methods vary, but also mainly include external dictionaries or machine learning methods such as seq2seq.

Stemming uses various fixed rules and dictionaries to recognise and shorten affixes, or machine learning may be used.

2.2 What are the pros and cons of both methods ?

We can use both in the text's simplification for analysis. This is due to the fact that related words with similar lexical meanings will be substituted into a single term, making it easier to train models on text.

Because stemming methods are typically easier to implement and faster to execute than lemmatization methods, stemming has a distinct advantage over lemmatization. However, stemming is not always accurate, and in some cases lemmatization is the only way to address a scenario. In fact, lemmatization will be able to handle words that have a similar lexical meaning but different forms.

3 On logistic regression :

3.1 How does stochastic gradient descent work ?

An iterative technique called stochastic gradient descent is used to optimize loss functions based on a dataset. To minimize the output of the loss function, variables are iteratively moved in the opposite direction as the gradient determined on the current point.

On stochastic gradient descent, this gradient is only computed using one (or a "batch" of") data points from the full dataset, with the algorithm iteratively going through all the data points in the dataset. On conventional gradient descent, this gradient is computed using the complete dataset.

3.2 What is the role of the learning rate ?

When removing the gradient from the loss function's variables in the **Stochastic Gradient Descent**, the learning rate is a variable that is added to the value. Accordingly, a large learning rate indicates that each gradient descent step will move the current point significantly, but a small learning rate means that the current point will only move marginally.

3.3 Will it always find the global minimum ?

Because the loss function used for optimization, the logistic loss, is a convex function, the logistic regression optimization issue is formulated as a convex problem. Since the global minimum must always be discovered (or approximated) on logistic regression, any local minimum in a logistic regression problem must be the global minimum.

4 What problems does TF-iDF try to solve ?

The TF-iDF test determines how significant a word is in a particular document. The main issue with measurements like these is that we can assume that words that appear frequently in a document are of high importance in this document.

4.1 What the is the TF part for ?

The term "term frequency" (abbreviated "TF") refers to the frequency of a certain word in a particular document. The weights of a word that occurs frequently in a manuscript are increased since, according to theory, words with high TF are significant in the document.

4.2 What is the iDF part for ?

Inverse document frequency, or iDF, is a measure that is inversely proportional to the frequency of a given term throughout the whole document set. To put it another way, a term with a high iDF is not very common in the document set, whereas a word with a low iDF is present in large quantities.

5 Summarize how the skip-gram method of Word2Vec works using a couple of paragraphs.

The goal of Word2Vec is to suggest an effective method for projecting words onto vectors. This indicates that we search similarity among word vectors that are frequently seen together.

To do this, we'll utilize a corpus of text and attempt to learn the ideal embeddings : embeddings that maximize similarity between vectors of words that frequently appear in the same context (in our corpus) and minimize similarity between vectors of words that do not.

This can be accomplished using a classifier that calculates the likelihood that a target word will appear in a given context given a set of context terms. When we use the vectors of the target word and the context words as weights in our classifier to calculate this probability, the vectors will be learned in such a way that the target word vectors become progressively more similar to the context word vectors when the target word does indeed occur frequently with the context words.

5.1 How does it uses the fact that two words appearing in similar contexts are likely to have similar meanings ?

When two words are used in the same context, their target vector embeddings converge to those of the same context vectors as learning proceeds. These two vectors are similar to the same context vectors at the conclusion of learning, thus they will likely also be similar.

6 What are the differences between an RNN and an LSTM ?

Recurrent neural networks, or RNNs, are neural networks that have short-term memory by influencing their inputs with their output. They are employed in sequential contexts, such as NLP, where, for instance, the output of an RNN when it considers one word in a sentence will influence the RNN when it considers the following word in the phrase.

The phrase "long short-term memory" (LSTM) refers to a particular application of RNN in which additional special units are added to store and reuse the RNN outputs, extending their short-term memory. These special's job is to make judgments on when to remember, use, and store facts that are learned through the algorithm.

6.1 What problem is an LSTM trying to solve compared to a basic RNN ?

The issue with gradient descent, which affects RNNs frequently, is when the gradient used in backpropagation to update the weights "evaporates" (approximate 0). In RNNs, this is brought on by repeated multiplications. The learning process can then sluggish or perhaps cease altogether.

Due to the new special cells in LSTMs, the backpropagated error is also retained in "memory," allowing the model to overcome the vanishing gradient issue.

7 What would you expect if we use one of our classifiers trained on IMDB on Twitter data, and why ?

Given how significantly different the vocabulary used in internet chats and IMDB ratings is, we believe it would likely not work correctly. Due to the rarity of tweets that are exclusively favorable or exclusively negative, categorisation may also be challenging.