NLP Lecture 2

Tokenization

- *Tokenization is breaking a text chunk in smaller parts. Whether it is breaking Paragraph in sentences, sentence into words or word in characters.*
- *https://medium.com/data-science-in-your-pocket/tokenization-algorithms-in-natural-language-processing-nlp-1fceab8454af*
- *https://analyticsindiamag.com/hands-on-guide-to-different-tokenization-methods-in-nlp/*

Stemming

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers.

A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”. Stemming is an important part of the pipelining process in Natural language processing. The input to the stemmer is tokenized words.

https://www.geeksforgeeks.org/introduction-to-stemming/

http://snowball.tartarus.org/algorithms/porter/stemmer.html

Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

Examples of lemmatization:

-> rocks : rock,corpora : corpus,better : good

Diff b/w stemming and lemmatization

https://towardsdatascience.com/lemmatization-in-natural-language-processing-nlp-and-machine-learning-a4416f69a7b6

POS Tagging

- Part of Speech (hereby referred to as POS) Tags are useful for building parse trees, which are used in building NERs (most named entities are Nouns) and extracting relations between words.
- POS Tagging is also essential for building lemmatizers which are used to reduce a word to its root form.
- POS tagging is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition.
- For example: In the sentence “Give me your answer”, answer is a Noun, but in the sentence “Answer the question”, answer is a verb.

Different POS Tagging Techniques

1. **Lexical Based Methods** — Assigns the POS tag the most frequently occurring with a word in the training corpus.

2. **Rule-Based Methods** — Assigns POS tags based on rules. For example, we can have a rule that says, words ending with “ed” or “ing” must be assigned to a verb. Rule-Based Techniques can be used along with Lexical Based approaches to allow POS Tagging of words that are not present in the training corpus but are there in the testing data.

Different POS Tagging Techniques(contd)

3. **Probabilistic Methods** — This method assigns the POS tags based on the probability of a particular tag sequence occurring. Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) are probabilistic approaches to assign a POS Tag.

4. **Deep Learning Methods** — Recurrent Neural Networks can also be used for POS tagging.

Conditional Random Fields

A CRF is a Discriminative Probabilistic Classifiers which models conditional probability distribution, i.e., P(y|x).

Logistic Regression, SVM, CRF are Discriminative Classifiers. Naive Bayes, HMMs are Generative Classifiers.

In CRFs, the input is a set of features (real numbers) derived from the input sequence using feature functions, the weights associated with the features (that are learned) and the previous label and the task is to predict the current label. The weights of different feature functions will be determined such that the likelihood of the labels in the training data will be maximised.

In CRF, a set of feature functions are defined to extract features for each word in a sentence. Some examples of feature functions are: is the first letter of the word capitalised, what the suffix and prefix of the word, what is the previous word, is it the first or the last word of the sentence, is it a number etc. These set of features are called **State Features**.

In CRF, we also pass the label of the previous word and the label of the current word to learn the weights. CRF will try to determine the weights of different feature functions that will maximise the likelihood of the labels in the training data. The feature function dependent on the label of the previous word is **Transition Features.**

https://sklearn-crfsuite.readthedocs.io/en/latest/api.html#module-sklearn_crfsuite

https://github.com/AiswaryaSrinivas/DataScienceWithPython/blob/master/CRF%20POS%20Tagging.ipynb

Wordnet

- WordNet is the lexical database i.e. dictionary for the English language, specifically designed for natural language processing.
- Synset is a special kind of a simple interface that is present in NLTK to look up words in WordNet. Synset instances are the groupings of synonymous words that express the same concept.
- Example -

```
from nltk.corpus import wordnet
for syn in wordnet.synsets("good"):

    for l in syn.lemmas():

          synonyms.append(l.name())
```

```
{'beneficial', 'just', 'upright', 'thoroughly', 'in_force', 'well', 'skilful',
'skillful', 'sound', 'unspoiled', 'expert', 'proficient', 'in_effect',
'honorable', 'adept', 'secure', 'commodity', 'estimable', 'soundly',
'right', 'respectable', 'good', 'serious', 'ripe', 'salutary', 'dear',
'practiced', 'goodness', 'safe', 'effective', 'unspoilt', 'dependable',
'undecomposed', 'honest', 'full', 'near', 'trade_good'}
```

Word Sense Disambiguation (WSD)

For example, consider the two sentences.

“The bank will not be accepting cash on Saturdays. ”

“The river overflowed the bank.”

The word bank in the first sentence refers to the commercial (finance) banks, while in second sentence, it refers to the river bank.

https://towardsdatascience.com/a-simple-word-sense-disambiguation-application-3ca645c56357

Query Expansion

- Query expansion (QE) is the process of reformulating a given query to improve retrieval performance in information retrieval operations, particularly in the context of query understanding.

https://github.com/ellisa1419/Wordnet-Query-Expansion/blob/master/wordnet.py