Task 3.2

**How do you think Language Translators Apps use Machine Learning?**

Language translation apps use machine learning to translate text from one language to another. Machine learning algorithms are used to train models that can recognize patterns in language data and learn how to translate between languages. These models are trained on large datasets of translated text, which allows them to learn the rules of grammar and syntax in each language. Once the model is trained, it can be used to translate new text from one language to another. [There are many different approaches to machine learning for language translation, including rule-based systems, statistical machine translation, and neural machine translation](https://data-flair.training/blogs/language-translation-machine-learning/).

Ref: [Language Translation with Machine Learning - DataFlair (data-flair.training)](https://data-flair.training/blogs/language-translation-machine-learning/)

**Explain what preprocessing (cleaning like removal of certain words, text shortening etc.) occurs and how these apps maintain the structure and meaning of the sentence while translating.**

**Text Preprocessing**

In natural language processing, text preprocessing is the practice of cleaning and preparing text data. NLTK and re are common Python libraries used to handle many text preprocessing tasks.

**Noise Removal**

In natural language processing, noise removal is a text preprocessing task devoted to stripping text of formatting.

import re

text = "Five fantastic fish flew off to find faraway functions. Maybe find another five fantastic fish? Find my fish with a function please!"

# remove punctuation

result = re.sub(r'[\.\?\!\,\:\;\"]', '', text)

print(result)

# Five fantastic fish flew off to find faraway functions Maybe find another five fantastic fish Find my fish with a function please

**Tokenization**

In natural language processing, tokenization is the text preprocessing task of breaking up text into smaller components of text (known as tokens).

from nltk.tokenize import word\_tokenize

text = "This is a text to tokenize"

tokenized = word\_tokenize(text)

print(tokenized)

# ["This", "is", "a", "text", "to", "tokenize"]

**Text Normalization**

In natural language processing, normalization encompasses many text preprocessing tasks including stemming, lemmatization, upper or lowercasing, and stopwords removal.

**Stemming**

In natural language processing, stemming is the text preprocessing normalization task concerned with bluntly removing word affixes (prefixes and suffixes).

from nltk.stem import PorterStemmer

tokenized = ["So", "many", "squids", "are", "jumping"]

stemmer = PorterStemmer()

stemmed = [stemmer.stem(token) for token in tokenized]

print(stemmed)

# ['So', 'mani', 'squid', 'are', 'jump']

**Lemmatization**

In natural language processing, lemmatization is the text preprocessing normalization task concerned with bringing words down to their root forms.

from nltk.stem import WordNetLemmatizer

tokenized = ["So", "many", "squids", "are", "jumping"]

lemmatizer = WordNetLemmatizer()

lemmatized = [lemmatizer.lemmatize(token) for token in tokenized]

print(stemmed)

# ['So', 'many', 'squid', 'be', 'jump']

**Stopword Removal**

In natural language processing, stopword removal is the process of removing words from a string that don’t provide any information about the tone of a statement.

from nltk.corpus import stopwords

# define set of English stopwords

stop\_words = set(stopwords.words('english'))

# remove stopwords from tokens in dataset

statement\_no\_stop = [word for word in word\_tokens if word not in stop\_words]

**Part-of-Speech Tagging**

In natural language processing, part-of-speech tagging is the process of assigning a part of speech to every word in a string. Using the part of speech can improve the results of lemmatization.

Ref: codecademy

**Describe the techniques used to convert the words between languages.**

1)[**Rule-based machine translation (RBMT)** is a machine translation system based on linguistic information about source and target languages](https://en.wikipedia.org/wiki/Rule-based_machine_translation). [It is basically retrieved from (unilingual, bilingual or multilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively](https://en.wikipedia.org/wiki/Rule-based_machine_translation).

* [RBMT systems are the first commercial machine translation systems](https://www.bing.com/search?q=Neural+machine+translation+in+150+words&qs=n&form=QBRE&sp=-1&ghc=1&lq=0&pq=neural+machine+translation+in+150+words&sc=4-39&sk=&cvid=471E9C8E708B41A7B158A20E4ABA3441&ghsh=0&ghacc=0&ghpl=)[and were developed in the early 1970s](https://en.wikipedia.org/wiki/Rule-based_machine_translation). [The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning](https://en.wikipedia.org/wiki/Rule-based_machine_translation).
* [RBMT systems can be characterized as the systems opposite to Example-based Systems of Machine Translation (Example Based Machine Translation), whereas Hybrid Machine Translations Systems make use of many principles derived from RBMT](https://en.wikipedia.org/wiki/Rule-based_machine_translation).

There are three different types of rule-based machine translation systems:

1. **Direct Systems (Dictionary Based Machine Translation)** map input to output with basic rules.
2. **Transfer RBMT Systems (Transfer Based Machine Translation)** employ morphological and syntactical analysis.
3. [**Interlingual RBMT Systems (Interlingua)** use an abstract meaning](https://en.wikipedia.org/wiki/Rule-based_machine_translation).

Ref: en.wikipedia.org, omniscien.com

2) **Statistical machine translation (SMT)** is a machine translation approach that uses statistical models to translate text from one language to another. It is a subfield of natural language processing (NLP) that involves analysing large amounts of bilingual text to build models that can accurately translate between languages.

* SMT works by analysing large bilingual corpora, such as parallel texts or sentence-aligned translation pairs, to identify patterns and relationships between words and phrases in different languages. These patterns are then used to build probabilistic models that can be used to generate translations for new sentences or documents.
* [One of the key advantages of SMT is its ability to handle a wide range of language pairs and translation tasks, from simple phrase-based translations to more complex neural machine translations](https://en.wikipedia.org/wiki/Statistical_machine_translation).
* [Statistical machine translation superseded the previous, rule-based approach because it required explicit description of each and every linguistic rule, which was costly, and which often did not generalize to other languages](https://en.wikipedia.org/wiki/Statistical_machine_translation). [Since 2003, the statistical approach itself has been gradually superseded by the deep learning-based neural network approach](https://en.wikipedia.org/wiki/Statistical_machine_translation).

Ref: en.wikipedia.org, geeksforgeeks.org, machinetranslate.org

3)**Neural machine translation (NMT)** is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modelling entire sentences in a single integrated model.

* NMT is a machine translation approach that applies a large artificial neural network toward predicting the likelihood of a sequence of words, often in the form of whole sentences.
* Unlike statistical machine translation, which consumes more memory and time, NMT trains its parts end-to-end to maximize performance.
* NMT systems are quickly moving to the forefront of machine translation, recently outcompeting traditional forms of translation systems. The word sequence modelling was at first typically done using a recurrent neural network (RNN). [A bidirectional recurrent neural network, known as an encoder, is used by the neural network to encode a source sentence for a second RNN, known as a decoder, that is used to predict words in the target language](https://en.wikipedia.org/wiki/Neural_machine_translation).

Ref: en.wikipedia.org, deepai.org