EXPERIMENT NO – 8

STUDY AND PREPROCESSING OF PARALLEL CORPUS FOR NATURAL LANGUAGE PROCESSING (NLP)

- ***Need of parallel corpus:***

1. *A parallel corpus plays an important role in machine translation (MT) systems*
2. *A parallel corpus consists of two or more monolingual corpora. The corpora are the translations of each other.*
3. *For example, a novel and its translation*

- ***Importance of parallel corpus:***

1. *Both languages need to be aligned, i.e., corresponding segments, usually sentences or paragraphs, need to be matched.*
2. *The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language.*
3. *The user can then observe how the search word or phrase is translated.*

- ***Handling of Low-Resource Language in Neural Machine Translation:*** *there are many methods for handling the low resource e language:*

1. ***Byte Pair Encoding (BPE):***

   - *To perform sub word tokenization, BPE is slightly modified in its implementation such that the frequently occurring sub word pairs are merged together instead of being replaced by another byte to enable compression.*
   - *This would basically lead the rare word athazagoraphobia to be split up into more frequent sub words such as ['▁ath', 'az', 'agor', 'aphobia'].*

2. ***Transformer models:***

   - *The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease.*
   - *The idea behind Transformer is to handle the dependencies between input and output with attention and recurrence completely*

- ***Source of corpus for experiment:***

1. *The source of the corpus is from the IIT Bombay English-Hindi Parallel Corpus.*
2. *The corpus is a compilation of parallel corpora previously available in the public domain as well as new parallel corpora we collected.*
3. *The corpus contains 1.49 million parallel segments*
4. *The corpus has been pre-processed for machine translation, and we report baseline phrase-based SMT and NMT translation results on this corpus.*

❖ ***Corpus details:***

1. *The parallel corpus has been compiled from a variety of existing sources (primarily OPUS (Tiedemann, 2012), HindEn (Bojar et al., 2014b) and TED (Abdelali et al., 2014)) as well as corpora developed at the Center for Indian Language Technology2 (CFILT), IIT Bombay over the years.*
2. *The training corpus consists of sentences, phrases as well as dictionary entries, spanning many applications and domains.*

❖ ***Corpus Statistics:***

1. *The test and dev (validation) corpora consist of newswire sentences.*

	Language	Train	Test	Dev
#Sentences		1,492,827	2,507	520
#Tokens	eng	20,667,259	57,803	10,656
	hin	22,171,543	63,853	10,174
#Types	eng	250,782	8,957	2,569
	hin	343,601	8,489	2,625

2. *The training, dev and test corpora consist of 1,492,827 and 520 and 2507 segments respectively.*
3. *Detailed Statistics are shown:*

❖ ***Preprocessing requirements:***

1. ***Tokenization:***

- *Tokenization is the process of demarcating and possibly classifying sections of a string of input characters. The resulting tokens are then passed on to some other form of processing.*
- *For example, in the text string:*

The quick brown fox jumps over the lazy dog

- *the string isn't implicitly segmented on spaces, as a natural language speaker would do.*
- *The raw input, the 43 characters, must be explicitly split into the 9 tokens with a given space delimiter as:*

```
(sentence
  (word The)
  (word quick)
  (word brown)
  (word fox)
  (word jumps)
  (word over)
  (word the)
  (word lazy)
  (word dog))
```

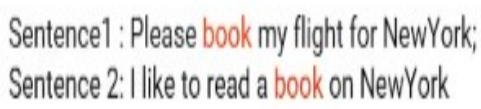
2. ***Stop word removal:***

   - *All stop words, for example, common words, such as a and the, are removed from multiple word queries to increase search performance.*
   - *The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms*

   .

3. ***Parts of speech tagging:***

   - *In corpus linguistics, **part-of-speech tagging** (**POS**), also called **grammatical tagging** is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.*
   - *For example, consider these sentences:*



   - *In both sentences, the keyword book is used, but in sentence one, it is used as a verb. While in sentence two it is used as a noun.*

4. ***Stemming:***

   - *Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.*

- *Stemming is important in natural language understanding (NLU) and natural language processing (NLP).*
- *we have a set of words — send, sent and sending. All three words are different tenses of the same root word send. So after we stem the words, we'll have just the one word — send.*

5. Lemmatization:

- *Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.*
- *If confronted with the token saw, stemming might return just s, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun.*