

**AIML CAPSTONE PROJECT**

NATURAL

LANGUAGE PROCESSING

**INTERIM REPORT – MACHINE TRANSLATION**

**Prepared by: Capstone Grp9 Team**

**Supervised by: Aniket Chhabra**

# Document Identification & Approval

|  |  |
| --- | --- |
| Document Name | Capstone NLP machine translation |
| Version Number | 1.0 |
| Effective Date | 4th March 2023 |
| Document Status | Interim Report |
| Document Author | Ajay KV  Bijay  Kiran  Rajesh Kumar  Srideepa  Sushma |
| Approved Date |  |

# Document Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Version & Date | Section & Page | Name | Comments |
| 1.0 & 04/03/2023 |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Prepared by:**

|  |
| --- |
| **March 22 Capstone Grp 9** |

**Reviewed by:**

|  |
| --- |
| **Aniket Chhabra** |

# Table of Contents

Contents

[Document Identification & Approval 2](#_Toc128698063)

[Document Revision History 3](#_Toc128698064)

[Table of Contents 4](#_Toc128698065)

[PROBLEM STATEMENT 5](#_Toc128698066)

[Dataset 5](#_Toc128698067)

[PROJECT OBJECTIVE: 6](#_Toc128698068)

[Milestone 1: 6](#_Toc128698069)

[Assumptions 6](#_Toc128698070)

[About RNN and LSTM 6](#_Toc128698071)

[Exploring Dataset 7](#_Toc128698072)

[Sentence length analysis 8](#_Toc128698073)

[Europarl-v7 sentences 8](#_Toc128698074)

[Commoncrawl sentences 9](#_Toc128698075)

[News-commentary-v9 sentence 9](#_Toc128698076)

[Final Merged dataset sentences 10](#_Toc128698077)

[Wordcloud Analysis on final dataset 11](#_Toc128698078)

[English Wordcloud 12](#_Toc128698079)

[German Wordcloud 12](#_Toc128698080)

[Accuracy of models 13](#_Toc128698081)

[Simple RNN model 13](#_Toc128698082)

[Simple LSTM model 13](#_Toc128698083)

[Conclusion 15](#_Toc128698084)

# PROBLEM STATEMENT

Machine Translation is the automated translation of source material into another language without human intervention. The database comes fromACL2014 Ninth workshop on Statistical Machine Translation. This workshop mainly focuses on language translation between European language pairs. The idea behind the workshop is to provide the ability for two parties to communicate and exchange the ideas from different countries

# Dataset

The database is basically sentences in German/English of various events. Three datasets are obtained from the Statistical Machine Translation workshop. Either the dataset can be downloaded from the link or can be used from the shared files. Three datasets are,

* Europarl v7
* Common Crawl corpus
* News Commentary

Link to download the dataset: https://statmt.org/wmt14/translation-task.html

# PROJECT OBJECTIVE:

Design a Machine Translation model that can be used to translate sentences from German language to English language or vice-versa.

## Milestone 1:

Data-cleaning, Pre-processing and modal creation

* Import and merge all the three datasets.
* Data cleansing
* NLP pre processing - Dataset suitable to be used for AIML model learning
* Design, train and test simple RNN & LSTM model
* Interim Report

## Assumptions

1. Total dataset of all merged file is 4,593,132 for German and same for English. Google Colab resources are not sufficient to handle this much huge size with GPU enable. Even 10K sentences from each dataset (Total 30K for German and 30K for English) were not able to handle by google colab. Currently we are using 5K sentences for each dataset.

# About RNN and LSTM

Machine Translation (MT) is a subfield of Natural Language Processing (NLP) that aims to automatically translate text from one language to another. There are two main approaches to MT: rule-based MT and statistical MT.

Rule-based MT relies on a set of linguistic rules to translate text. These rules are typically created by linguists and language experts, and the translation process is based on these rules. However, rule-based MT systems often require a lot of manual effort to create and maintain the rules, and they can struggle with idiomatic expressions and other nuances of language.

Statistical MT, on the other hand, relies on statistical models that learn to translate text from data. The models are trained on bilingual corpora, which are pairs of sentences in the source and target languages. The models then use these corpora to learn the patterns and relationships between the two languages, and can then translate new sentences based on these patterns.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models are commonly used in statistical MT. These models are able to process sequences of data, such as sentences, and can learn to capture the dependencies between the words in the sentence. RNNs are able to maintain a "memory" of previous inputs, which makes them well-suited for sequential data like text. LSTMs are a type of RNN that are able to selectively remember and forget information, which makes them even more powerful for language tasks.

In this project, we will use a combination of statistical MT and LSTM models to translate sentences from German to English (and vice versa). By training a model on bilingual corpora, we can create a system that can automatically translate sentences without human intervention. The ultimate goal of the project is to create a model that can accurately translate sentences between the two languages, and to explore ways to improve the accuracy and efficiency of the model.

# Exploring Dataset

To create a single dataset for our machine translation model, we combined the three datasets - Commoncrawl, Europarl-v7, and News-commentary-v9 - into one dataframe with two columns: "en" and "de" for English and German text, respectively. The datasets were loaded into Python and merged using the pandas library, which allowed us to easily concatenate the datasets while preserving the parallel structure of the sentences. The resulting data frame contains a total of 4,593,132 sentence pairs, with an equal number of English and German sentences.

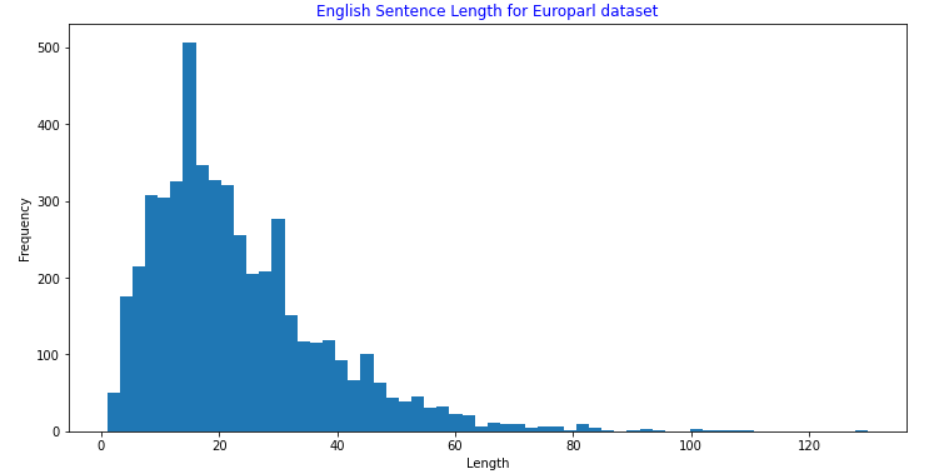
The project involves three main data sources, each of which includes two files containing text in both English and German.

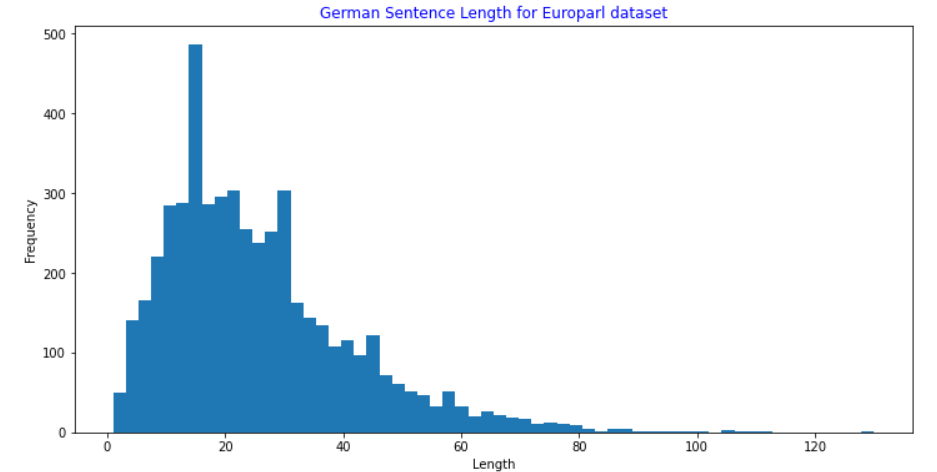
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data** | **Commoncrawl** | | **Europarl-v7** | | **News-commentary-v9** | |
| Language | English | German | English | German | English | German |
| Lines | 2399123 | 2399123 | 1920209 | 1920210 | 201995 | 201854 |

Each text file contains multiple lines of sentences in their respective languages, with corresponding sentences between the English and German files. For example, the 10th sentence in the English file will be the translation for the 10th sentence in the German file.

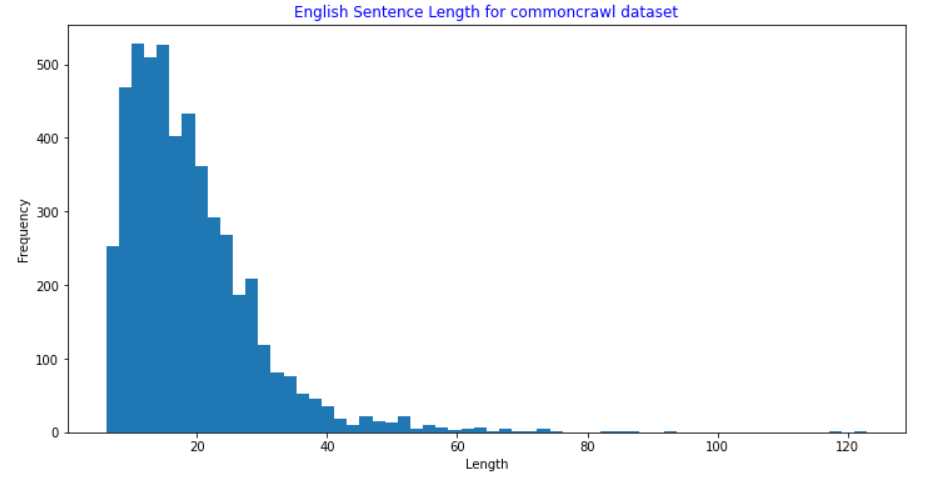
## Sentence length analysis

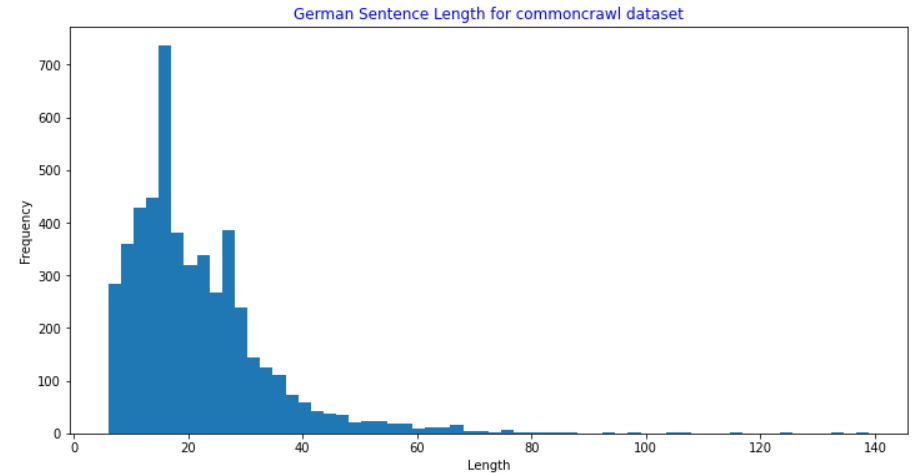
### Europarl-v7 sentences



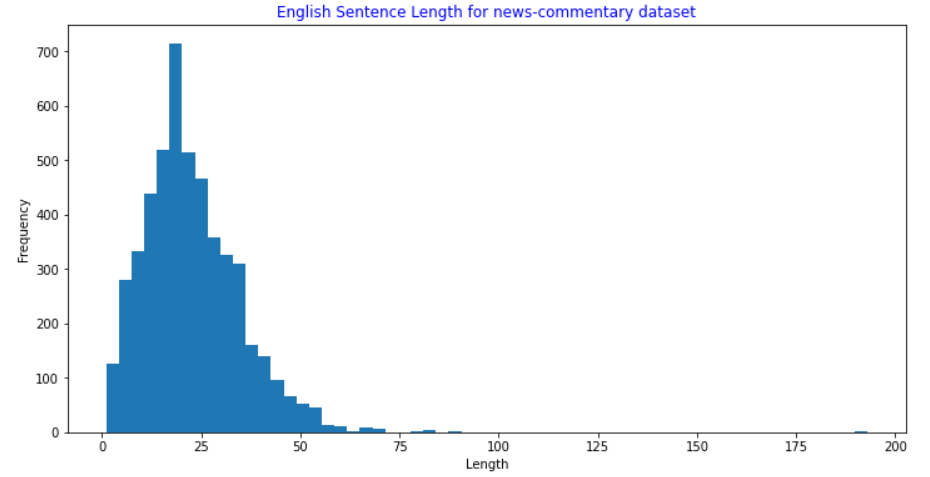


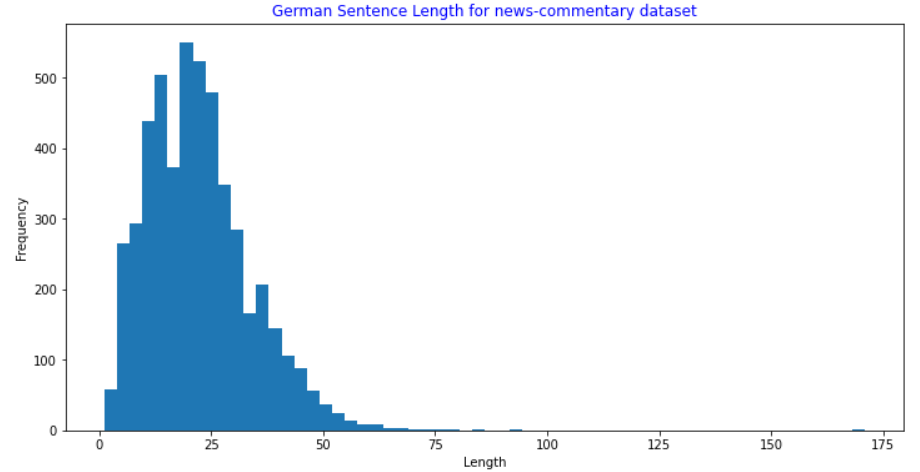
### Commoncrawl sentences



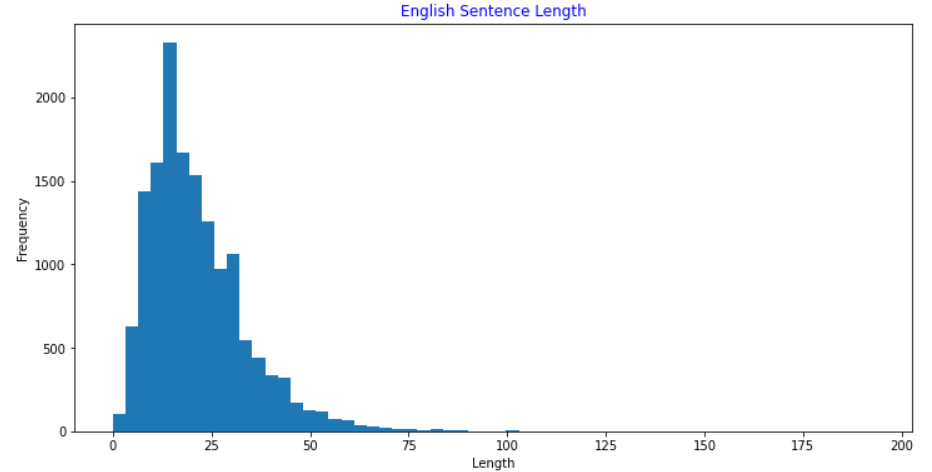


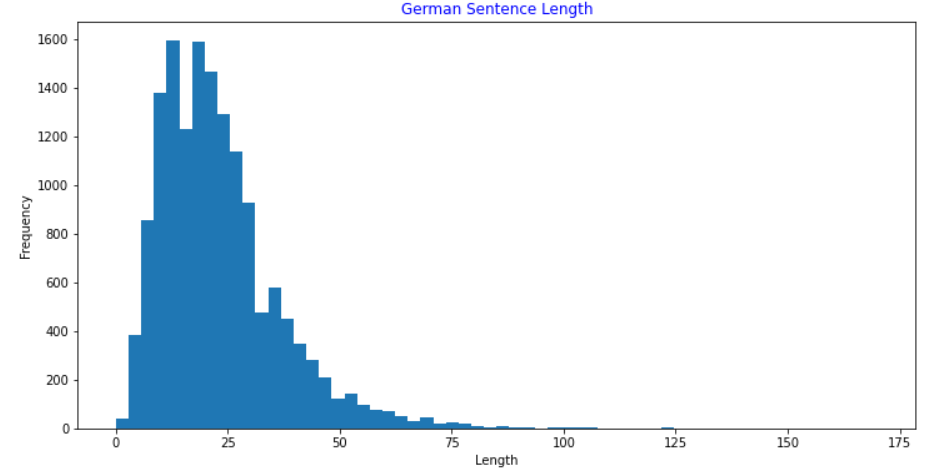
News-commentary-v9 sentence





### Final Merged dataset sentences

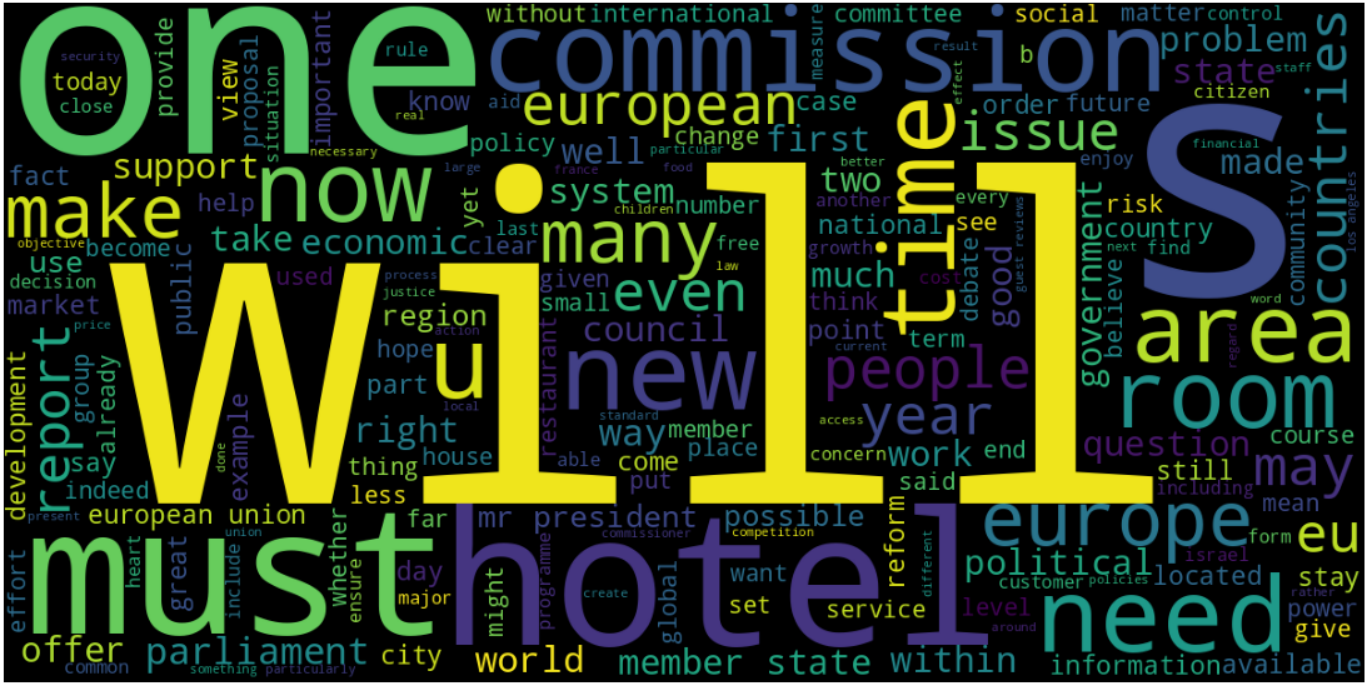




As seen from these plots, 99% of sentences have less than 50 words in both English and German text.

## Wordcloud Analysis on final dataset

### English Wordcloud



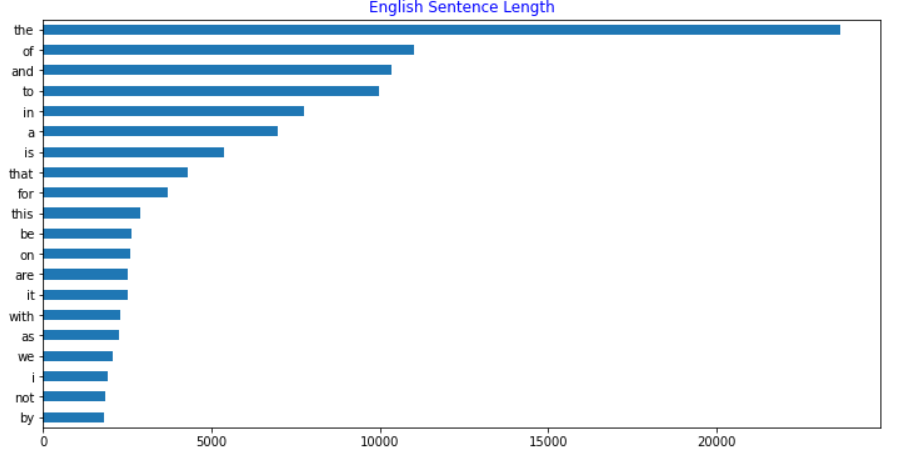
### German Wordcloud



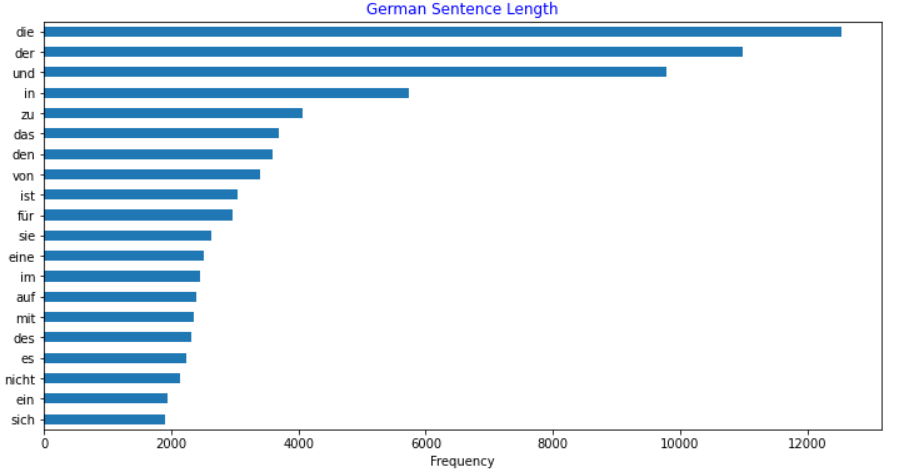
## ****Word frequency analysis****

Another way to see which words are the most common in both English and German text

### English Sentences Length

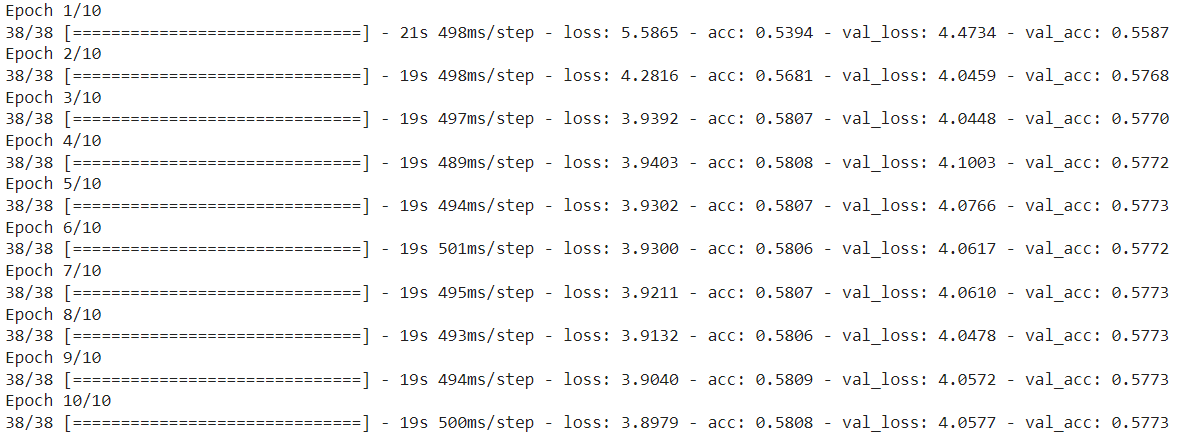


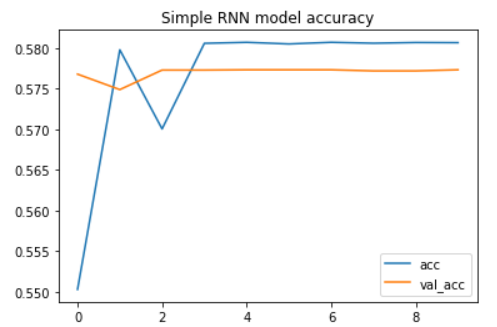
### German Sentences Length



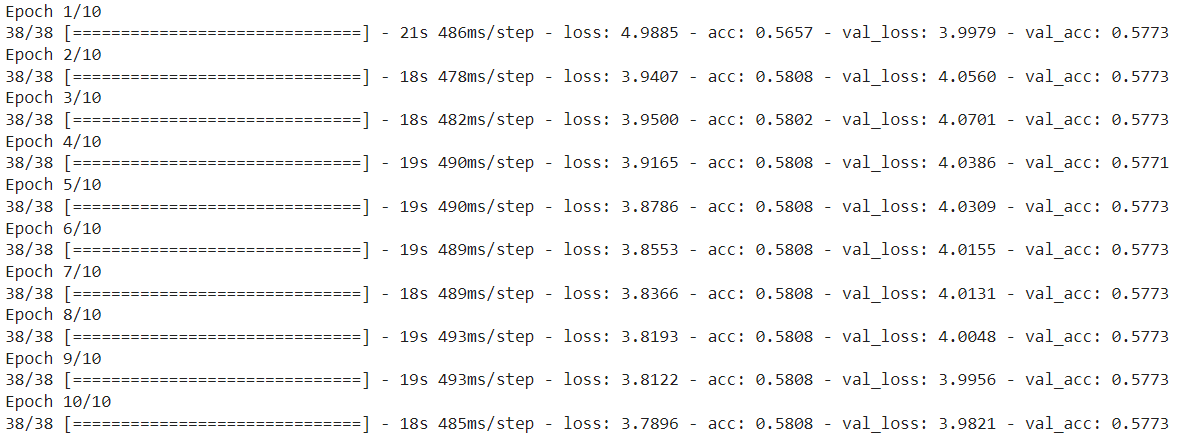
# Accuracy of models

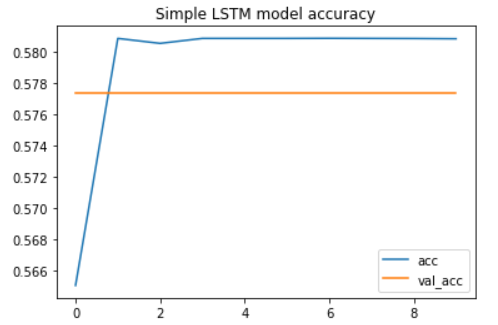
## Simple RNN model





## Simple LSTM model



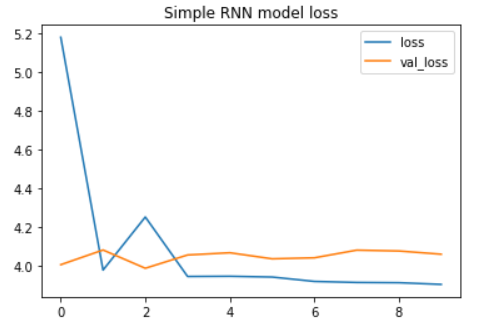


As per above diagram we can see that accuracy of LSTM is slight better than RNN model. Both model accuracy is less than 60%. Following are the consideration for model building.

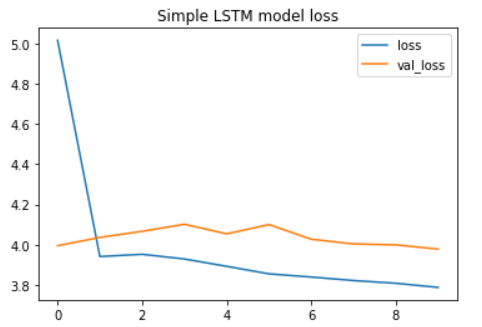
1. Optimizer used as **rms**.
2. **0.001** learning rate used for rms.
3. Loss used as **sparse\_categorical\_crossentropy**
4. Metrics used as **acc**
5. Epochs size used as 10. Accuracy was not getting improved with Epochs size.
6. Batch\_size used as 256

# Loss of models

## Simple RNN model



## Simple LSTM model



# Conclusion

Simple RNN and LSTM model accuracy is less than 60%. Further we need to explore to get more accuracy and translation

1. Glove Embedding
2. Bidirectional RNN and LSTM model
3. Encode Decoder to achieve better accuracy and translate English to German and vice versa.