Saint-Petersburg State University

System programing department

# Comparative Analysis of NLP Models for Automatic Grammar Correction

Karimi Hurmatullah

**Scientific supervisor:** D.V.Lutsiv, Associate Professor, System Programming Department

Saint-Petersburg
2023

# Introduction

- Grammar correction is the task of identifying and correcting grammatical errors in text.
- One of the problems - ‘select the most suitable model for grammar correction‘, we need to conduct research.

# Objective

To develop and implement an NLP-based system for the automatic replacement of terms in text with their correct forms.

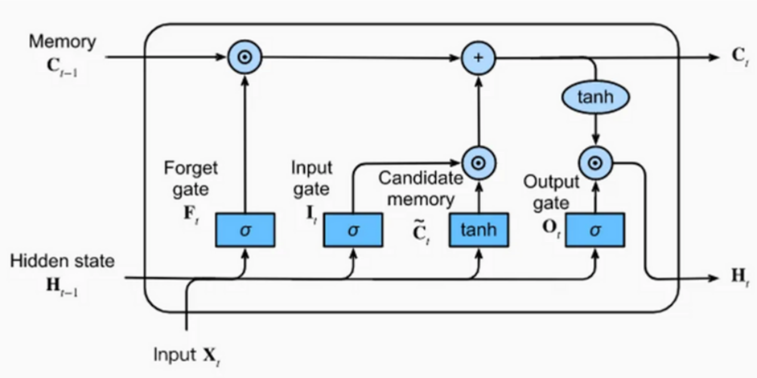
- Investigate NLP models capable of syntactic analysis of sentences.
- Utilize NLP models trained on grammatically correct text corpora to identify and rectify grammatical errors in text fragments.
- Evaluate the performance of these models through experiments and identify the most suitable one.
- Develop a prototype of a tool that can perform grammar correction on texts, ensuring proper case and number usage.

# NLP Models

- Long-Short Term Memory (LSTM)
- Attention Mechanism (Bahdanau)
- Text-to-Text Transfer Transformer (T5)

# Long Short-Term Memory (LSTM)[1]

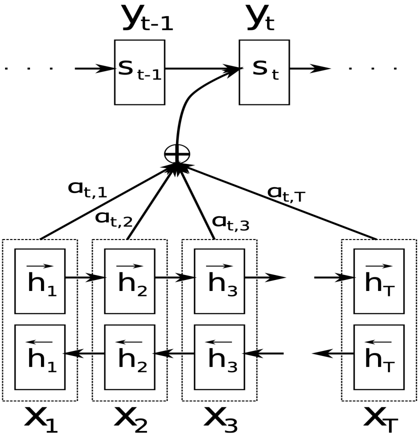
LSTM excels at capturing long-term dependencies, making it a powerful tool for sequence prediction in Deep Learning.



[1] Ottavio Calzone, "An Intuitive Explanation of LSTM - 2022"

# Attention Mechanism [2]

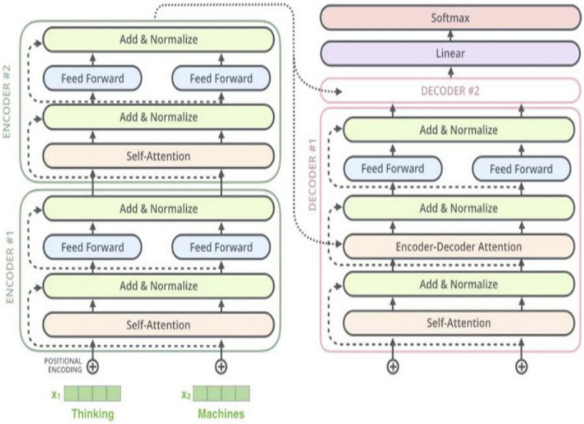
Attention mechanisms enable deep learning models to focus on specific part of data by assigning weights.



[2] Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua, "Neural machine translation by jointly learning to align and translate - 2014"

# Text-to-Text Transfer Transformer (T5) [3]

T5 leverages the Transformer architecture to effectively capture and process hierarchical representations of input and output sequences for text-to-text tasks.



[3] Qiurui Chen, "T5: a detailed explanation - 2020"

# Implementation

In implementation, the following things were executed:

- Data Collection
  - The Lang-8 Learner Corpora were utilized to collect data for this study.
- Data Preprocessing
  - The dataset was thoroughly cleaned to ensure its accuracy and consistency before analysis.
  - The dataset was purged of duplicate records to ensure the integrity and reliability of the data.
- Data Partitioning
  - The prepared data was systematically divided into three distinct subsets: training, validation, and testing sets.
- Model Selection
  - Three algorithms, LSTM, Attention Mechanism (Bahdanau), and T5, were chosen for training.
- Model Evaluation
  - To train the models, Adam and AdamW optimizers were employed.
  - To evaluate the quality of the generated text against the actual data, the GLEU score was used.

# Experiments

GLEU[4] score is simply the minimum of recall and precision. This GLEU score's range is always between 0 (no matches) and 1 (all match) and it is symmetrical when switching output and target.

| GLEU Score | |
|---|---|
| Models | Measure |
| LSTM | 0.217 |
| Attention Mechanism | 0.319 |
| T5 | 0.418 |

[4] NLTK, "GLEU Score Module"

# Examples

LSTM excels at capturing long-term dependencies, making it a powerful tool for sequence prediction in Deep Learning.

- LSTM
  - Input: Yes , I have finally got my own one .
  - Actual: Yes, I have finally gotten my own one.
  - Predicted: Sometimes , I have been one of this month .
- Attention Mechanism (Bahdanau)
  - Input: Yes , I have finally got my own one .
  - Actual: Yes, I have finally gotten my own one.
  - Predicted: Yes , I finally finally got my own one .
- Text-to-Text Transfer Transformer (T5)
  - Input: Yes , I have finally got my own one .
  - Actual: Yes, I have finally gotten my own one.
  - Predicted: Yes, I have finally gotten my own one.

# Additional Feature - Language Tool[5]

- LanguageTool is an open-source spelling, grammar, punctuation, and style tool that can correct mistakes in your writing.
- It's ideal for both native and non-native English speakers. Plus, it works with over 25 other languages.

[5] Language Tool - https://languagetool.org/

# Result

- Investigated various NLP models capable of analyzing sentences, including LSTM, Attention Mechanism (Bahadanu), and T5.
- Embraced T5 as a prominent model for grammar correction in sentences.
- Evaluated three models: LSTM, Attention Mechanism, and T5, through experiments.
- Integrated the tool with the LanguageTool library for additional sentence correction capabilities.
- Implemented a CI pipeline using GitHub Actions.
- Designed a prototype grammar correction model that effectively identifies and corrects grammatical errors in sentences.
- Source code available: https://github.com/Hurmatullah/English-Grammar-Corrector.git