Deep Neural NLP
Project Presentation

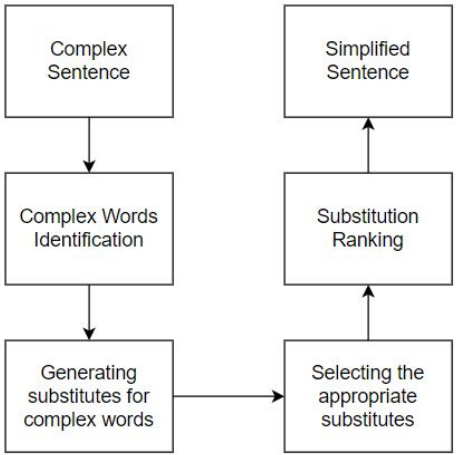
Research Paper Simplification

Presented by:
Jainisha Choksi (202211019)
Hinal Desai (202211035)
Man Desai (202211040)
Rutvik Prajapati (202211053)

01
INTRODUCTION

SIMPLIFICATION

- Simplification means modifying the content and structure of the text so as to make it easier to understand and read, but keeping the main idea and original meaning intact.
- It involves rephrasing the complex sentences into simpler sentences while retaining the original meaning.



02

Problem formulation

PROBLEM FORMULATION

- Given a research paper as an input, we aim to produce a simplified version of that research paper with a simpler vocabulary while preserving the main ideas in the original sentence.
- Our focus is on **lexical simplification.**
- Lexical simplification refers to replacing the complex words with simpler synonyms without changing the grammatical meaning of the text.
- **Input:** Research paper in natural language text format.
- **Output:** Simplified version of input text, with complex sentences rephrased.

03
Methodology

METHODOLOGY

01	Complex Word Identification	Identifying the words in the sentence that need to be simplified.
02	Substitute Generation	Getting the possible substitutes for the identified complex words.
03	Selecting the best candidates	Selecting the best substitutes based on zipf values.

METHODOLOGY

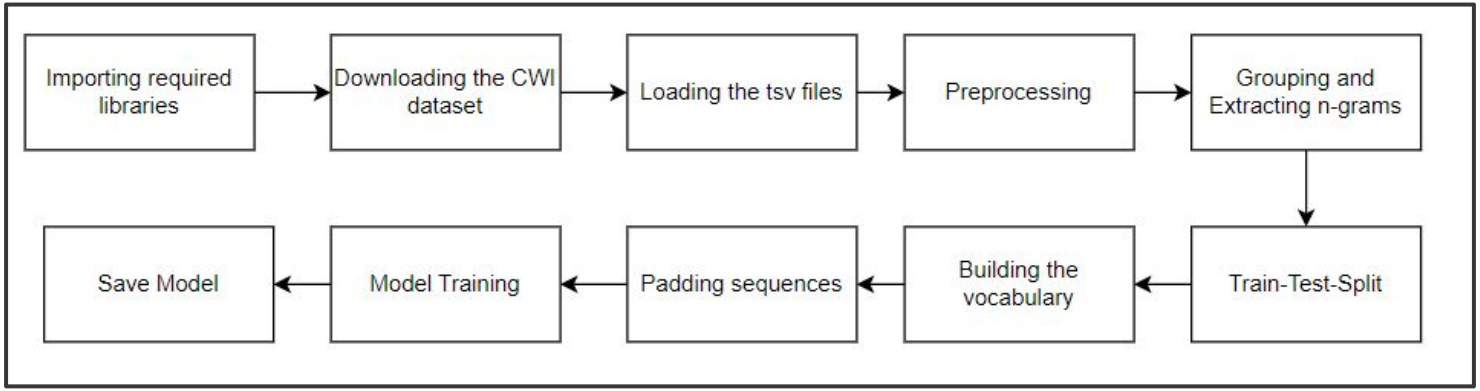


Fig: Complex Word Identification model

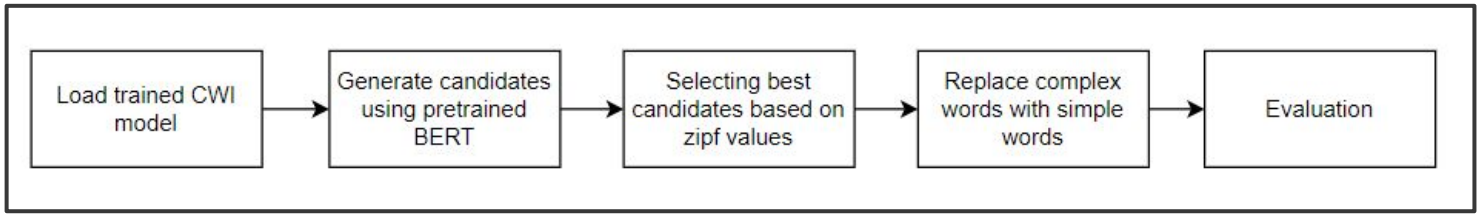


Fig: Substitute generation & Selecting best candidates

04

Model architecture

CWI MODEL

- **Complex word identification** is a subtask of lexical simplification. It identifies the difficult words or phrases in a text.
- The dataset used for complex word identification is the **Complex Word Identification (CWI) Shared Task 2018 dataset.** It contains information about complex phrases annotated with some statistics.
- Dataset consists of:
    - **id:** unique identifier
    - **sentence:** actual sentence where there exists complex phrase annotation.
    - **start, end:** start and end offsets of complex phrase annotation
    - **target:** actual complex phrase annotation.
    - **nat, non_nat, nat_marked:** number of native annotators, non-native annotators and total number of annotators who marked this complex phrase.

CWI MODEL ARCHITECTURE

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 103)]             0

 embedding (Embedding)       (None, 103, 300)          15736500

 dropout (Dropout)           (None, 103, 300)          0

 bidirectional (Bidirection  (None, 103, 300)          541200
 al)

 time_distributed (TimeDist  (None, 103, 2)            602
 ributed)

=================================================================
Total params: 16278302 (62.10 MB)
Trainable params: 16278302 (62.10 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Loss = categorical_crossentropy

Optimizer = adam

Metric = accuracy

RATIONALE BEHIND ARCHITECTURE

- **Input layer:** used to feed the data into the neural network. Input is a sequence of words.
- **Embedding layer:** used to convert words or tokens into dense vectors of fixed size. This step helps model capture semantic relationships between words and enables it to understand contextual meaning.
- **Dropout layer:** used for regularization to prevent overfitting.
- **Bidirectional layer:** processes the input sequence in both forward and backward directions, allowing the model to capture both past and future context.
- **Time distributed layer:** often used in sequence labelling tasks. It allows the model to apply the same set of parameters to every time step in the sequence independently. It helps the model make predictions for each word in the sequence.

SIMPLIFICATION MODEL

- Using BERT masked language model to get the possible candidates
- Computing the zipf frequency values of each candidates to rank and select the simplest one.
- **Zipf's Law:** It is an empirical law that describes the statistical distribution of word frequencies in natural language. It suggests that the frequency of a word is inversely proportional to its rank.
- We used the zipf_frequency function from wordfreq library to calculate the zipf_frequency.

SIMPLIFICATION MODEL

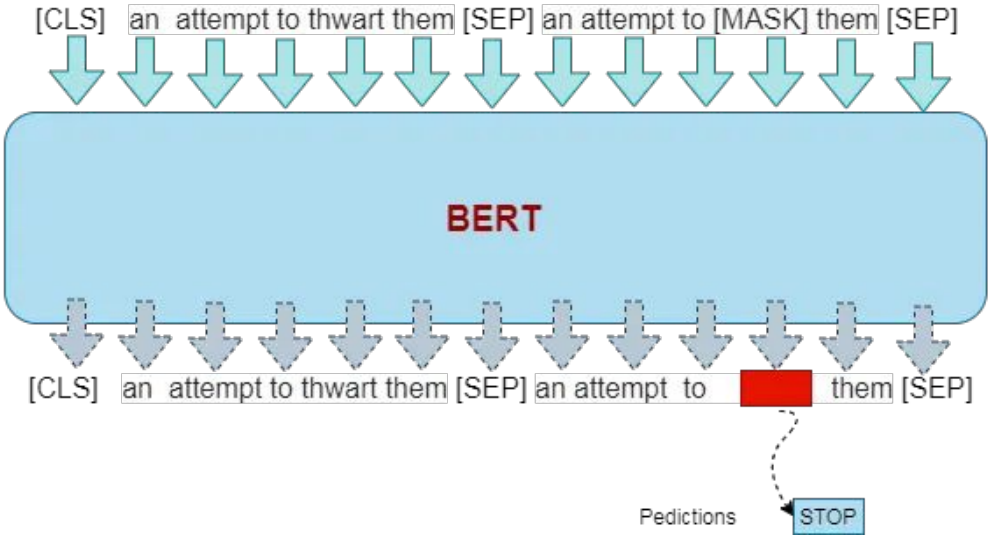


Image source: https://medium.com/@armandj.olivares/how-to-use-bert-for-lexical-simplification-6edbf5a4d15e

05
RESULTS

EVALUATION

Original text: "Event-Triggered Adaptive Output Feedback Control for Stochastic Nonlinear Systems With Time-Varying Full-State Constraints"

Simplified text: "Event-Triggered Adaptive Output Feedback system for Stochastic complex Systems With Time-Varying Full-State problems"

sentence_bleu score: 0.42502

Original text: "The developed algorithm allows the finding of solutions for a wide range of robots by using a geometric approach, representing points in a polar coordinate system"

Simplified text: "The developed system allowed the finding of solutions for a wide family of problems by using a linear approach, representing points in a polar coordinate system "

sentence_bleu score: 0.58288

THANKS!

REFERENCES

[1] Qiang, J., Li, Y., Zhu, Y., Yuan, Y. and Wu, X., 2020. LSBert: a simple framework for lexical simplification. arXiv preprint arXiv:2006.14939.

[2] Kriz, R., Miltsakaki, E., Apidianaki, M. and Callison-Burch, C., 2018, June. Simplification using paraphrases and context-based lexical substitution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 207-217).

[3] Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G.H., Specia, L., Štajner, S., Tack, A. and Zampieri, M., 2018. A report on the complex word identification shared task 2018. arXiv preprint arXiv:1804.09132.