Sentence Compression

By: Jonah, Gavin, Beck

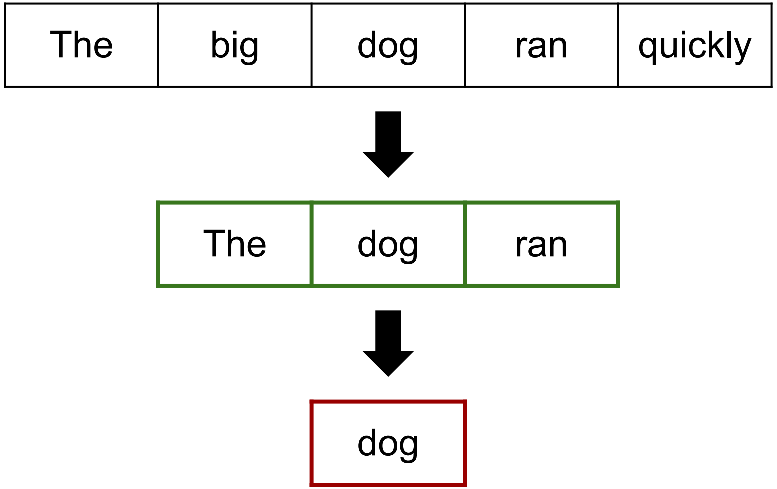
Introduction

Sentence/Text Compression - What is it?

- Shortening the length of a sentence while retaining its key meaning and essential information

Applications:

- Text Summarization
- Information Retrieval
- Machine Translation
- Subtitle Generation



The big dog ran quickly

The dog ran

dog

Approaches

Grammar Based Compression

- Split sentence into ordered subsets
- Check if subset is grammatical
- Calculate probability with a PCFG

Gated Recurrent Unit (GRU) Based Compression

- Type of RNN designed for sequential data
- Made with the Pytorch NN library
- Encoder-Decoder Framework

Transformer-Based Compression

- Pretrained Transformer
- BART from Meta AI
- Finetune with MSR dataset

Dataset: Google Sentence Compression

Data gathered from various news articles from across the Internet

- Specifically ones where the first sentence and headlines were similar
- Headlines were used as the “compressed sentence”

Problems:

- Excessive compression of sentences
- High compression ratio (Very long target sentence vs. very short compressed sentence)
- Lost much of the essential information and meaning
- Not very suitable for our desired application

Example data:

“The USHL completed an expansion draft on Monday as 10 players who were on the rosters of USHL teams during the 2009-10 season were selected by the League's two newest entries, the Muskegon Lumberjacks and Dubuque Fighting Saints.”

Becomes

“USHL completes expansion draft.”

Dataset 2: Microsoft Text Compression

Data gathered from various sources from the Open American National Corpus (OANC1).

- More variety than just news articles
- Includes business letters, journals, technical documents, etc.
- Each source text has up to **5 crowd-sourced rewrites**, which are constrained to a compression ratio
- Human reviewed
- **Multi sentence compression,** primarily two-sentence

Example data:
‘“Except for this small vocal minority, we have just not gotten a lot of groundswell against this from members,” says APA president Philip G. Zimbardo of Stanford University.’

Becomes
“APA president of Stanford has stated that except for a vocal minority they have not gotten a lot of pushback from members.”

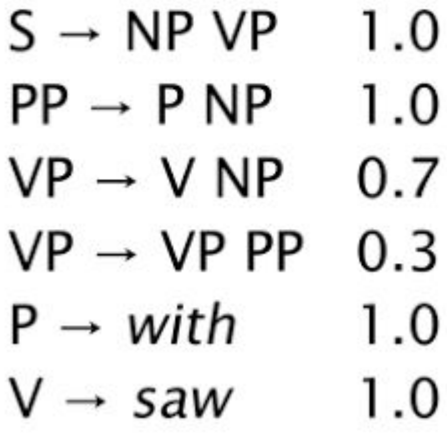
Data Preprocessing

01	Pandas Dataframe	Convert data to pandas dataframe for ease of use
02	Filter out unneeded data	Did not need extra compressions nor human review scores, also rows with missing data
03	Compressed Sentence Selection	Select the shortest compression for best results in training
04	Word Tokenization	Split text strings to tokens for training
05	Special Token Insertion	Adds tokens like <s>, </s> and <pad> to manage sentence boundaries

Grammar Based Compression

PCFG Implementation

- Shifted from rule-based to probability-based compression using PCFGs.
- Built PCFG from NLTK Treebank parse trees, replacing words with POS tags.
- Handled special cases like punctuation for grammar parser compatibility.



S → NP VP	1.0
PP → P NP	1.0
VP → V NP	0.7
VP → VP PP	0.3
P → *with*	1.0
V → *saw*	1.0

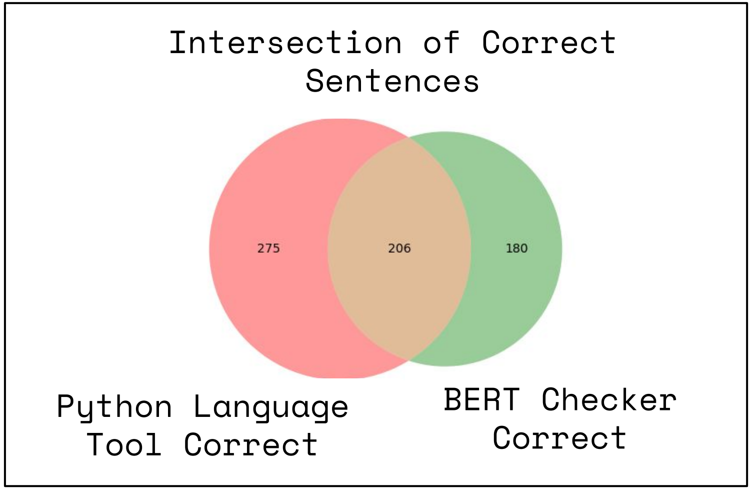
Grammar Based Compression

Grammar Checking Implementation

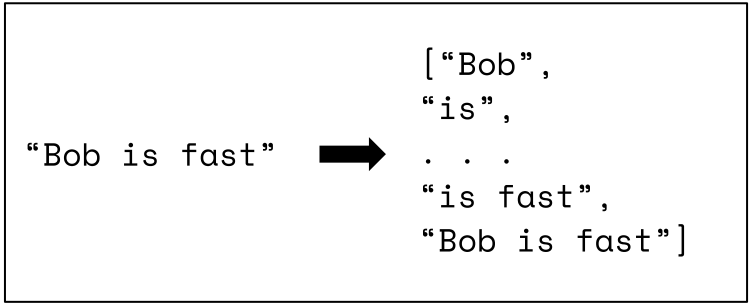
- language_tool_python library
- BERT-based grammar checker

Method

- Split sentence into ordered subsets
- Tag each word with POS
- Iterate through each subset, rank best new sentences



Intersection of Correct Sentences

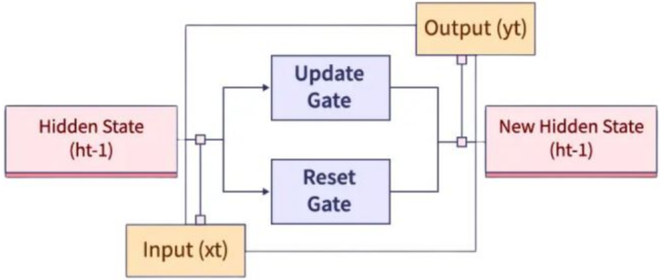


“Bob is fast” ➡ [“Bob”,
“is”,
. . .
“is fast”,
“Bob is fast”]

Gated Recurrent Unit (GRU)

- A fancy RNN
- Designed to handle sequential data
- A better alternative to LSTMs: simpler architecture so less gates & less expensive
- Update gate: controls how much info passed to the future
- Reset gate: controls what to forget
- As comparison, LSTM has 3 gates (input, forget, output)

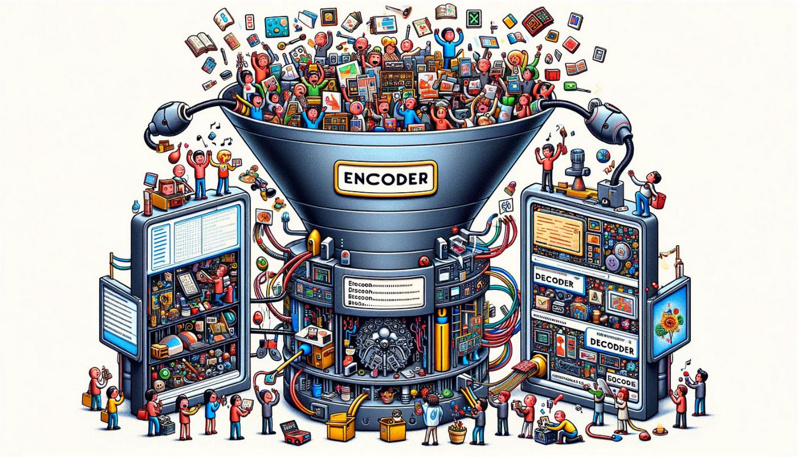
What is Gated recurrent units?

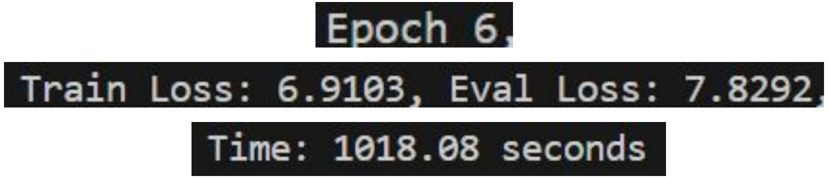


@AIonlinecourse

GRU Based Encoder-Decoder

- Sequence 2 sequence architecture
- Input -> Embedding layer -> Encoder GRU -> Decoder GRU -> Fully connected layer -> Output
- Slow training time 20min-1hour for 10 epochs
- Issues with Overfitting
- Limited vocabulary
- Struggles with Long Range Dependencies
- There’s already better suited models (like a transformer) for this problem





```
Epoch 6,
Train Loss: 6.9103, Eval Loss: 7.8292,
Time: 1018.08 seconds
```

Transformer Based Compression

- Model: BART (facebook/bart-base) - Transformer-based bidirectional encoder-decoder
- Combines bidirectional attention mechanism of BERT with decoder structure of GPT
- Less extreme deletion allows us to have multiple levels of compression
- Relatively small trained model sizes of around 600MB
- Relatively low computational costs during inference

Example input:
"In recent years, the importance of mental health awareness has grown significantly, as people around the world begin to understand that mental well-being is just as crucial as physical health."

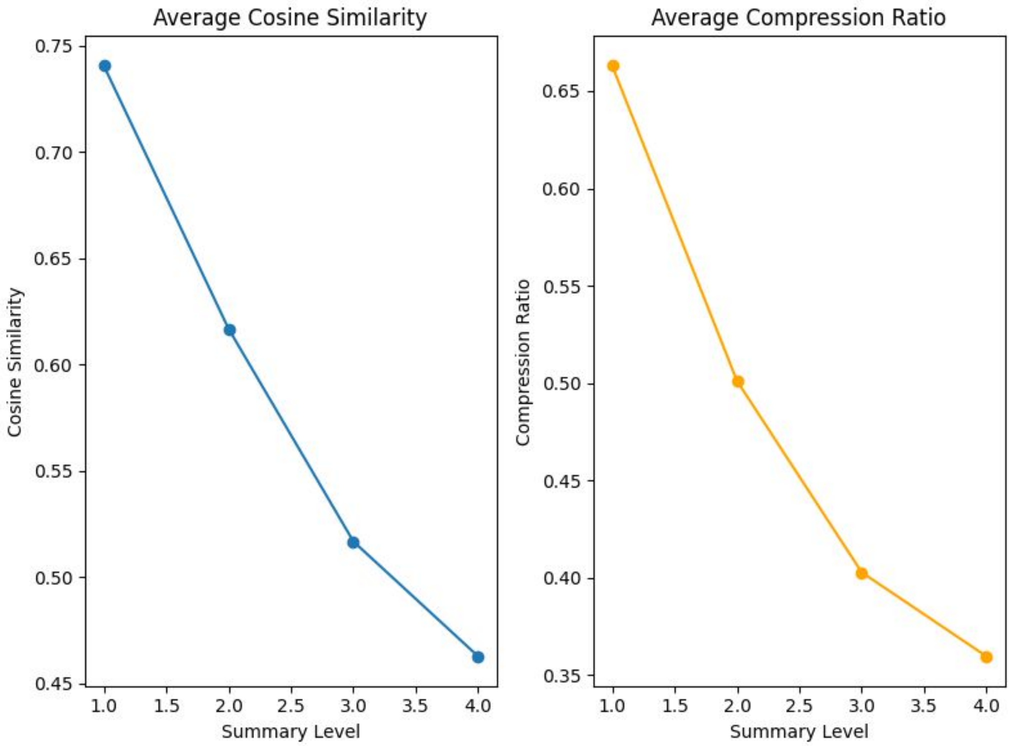
Level 2 Compression:
"In recent years, people realize that mental well-being is just as crucial as physical health."

Level 4 Compression:
"People realize that mental well-being is important."

DEMO

Results and Analysis

- Consistently able to shorten sentences while maintaining relatively high similarity scores
- Evaluation Loss was able to decrease to around 0.19
- The Transformer showed the greatest reliability, and interpretability of output



Conclusion

Best Implementation

- Microsoft Dataset
- BART Model
- 0.19 Evaluation Loss

Future Work and Improvements

- Reinforcement Learning
- Paragraph and document summarization