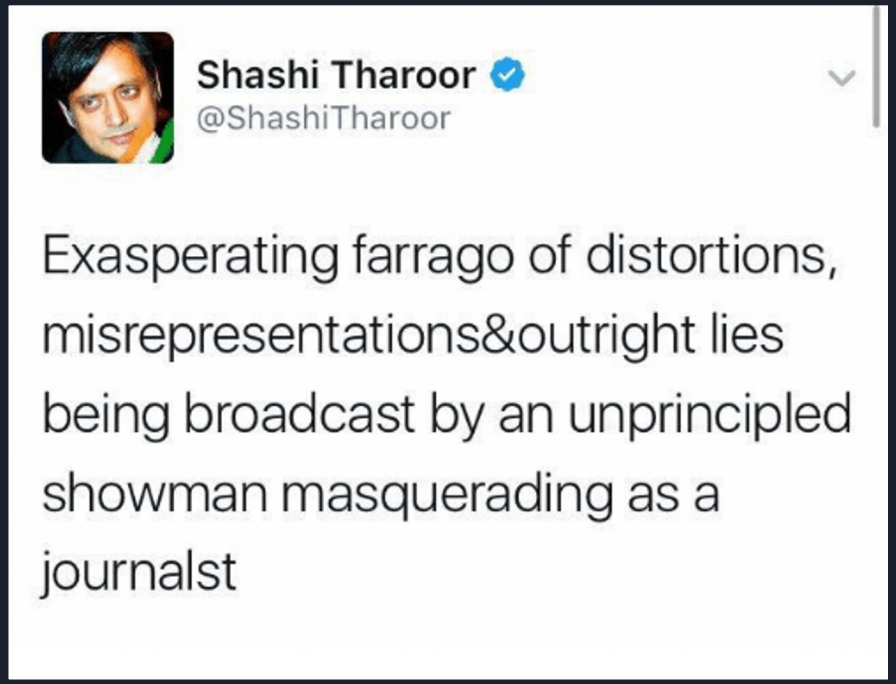
The Text Simplifier

Team Members :

- Venkata Vuppuluri
- Rohith Kadivendi
- Manisha Bachu
- Harika Gumudavally
- Vamsi Tallam
- Revanth Reddy Male
- Raghu Sannapareddy

PROBLEM



Simplified sentence-

Extremely irritating lies being broadcast by a law-breaking showman pretending to be a journalist

PROBLEM STATEMENT

“Text Simplification is a process of replacement of long complicated sentences to short simple sentences with minimal loss of context”

MOTIVATION

“English is one of the most spoken languages in the world, hence there is a need to learn the language even for non-native speakers. “

“Often times literature is abstruse and word-to-word interpretations of sentences may not help in most cases, and it makes reading harder and difficult to comprehend.”

USE CASES

- **Simplification** : Long and complicated sentences prove to be a stumbling block for current systems relying on NL input. Ex: Language Translator, sentiment analysis system.

- **Parsing** : Syntactically complex sentences are likely to generate a large number of parses and may cause parsers to fail altogether. (simpler sentences lead to faster and less ambiguous parsing)

- **Machine Translation** : It is a sub-field of computational linguistics that deals with use of software to translate text to speech from one language to another(will lead to improvement in the quality)

- **Summarization :** Simplification can be used to remove irrelevant text with greater precision, and thus aid in summarization.

- **Clarity of Text** : For a layman, all the esoteric information in areas such as business, legal, entertainment, technical, science and even literature can be made available in an understandable language.

Our Approach

Supervised Training methodology using Neural Machine Translation

Note: A transformer based model with attention is built from scratch but disregarded due to large training time, prediction time and memory usage

TOOLS

Embeddings : GloVe, Word2Vec, fastText

Machine Translation framework: OpenNMT
(alternative frameworks in appendix)

Evaluation : BLEU, SARI
Computation Device: Google Colaboratory

Utilities : Huggingface, SentencePiece

Data Sets

Training data:

- Input sentence is in standard english.
- Output sentence is in simplified english.
- The Dataset was created using a web scraping script from with data taken from Wikipedia and simple.Wikipedia
- http://www.cs.pomona.edu/~dkauchak/simplification/

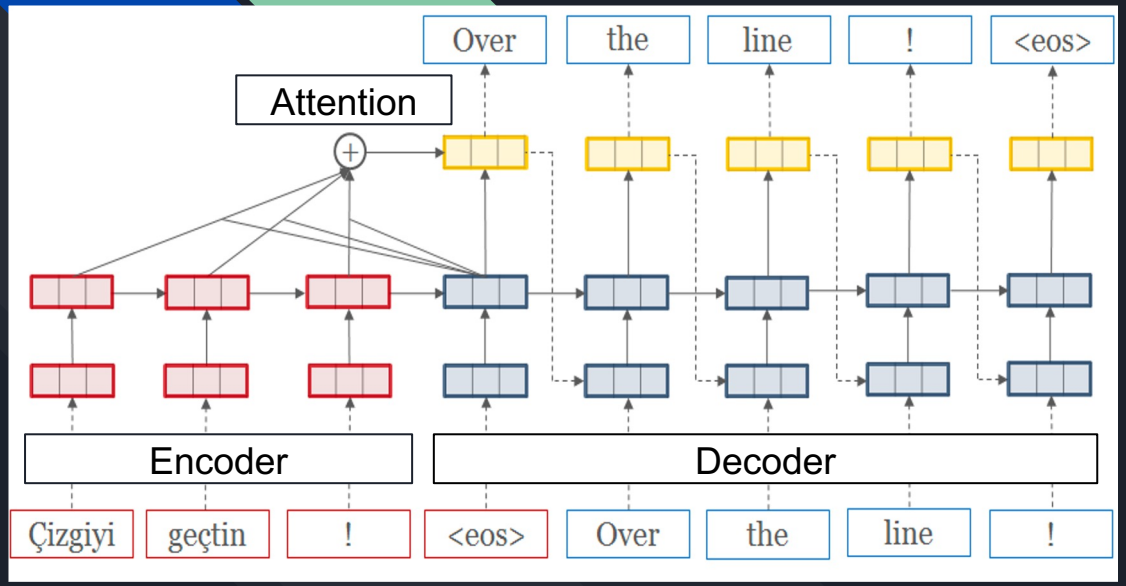
Evaluation:

- Turk Corpus "https://github.com/cocoxu/simplification/blob/master/data/turkcorpus"
  - Each input sentence has 8 reference sentences as output
- ASSET Data "https://github.com/facebookresearch/asset"
  - Each input sentence has 10 reference sentences as output

Steps :

1. **Preprocessing** (tokenization, glove embeddings)
1. **Training** (transformer+attention)
1. **Translate** (give sentences to predict)
1. **Evaluate** (performance of model based on metrics : BLEU/SARI)

Transformer



Encoder decoder architecture with attention mechanism.

```
# Optimization
model_dtype: "fp32"
optim: "adam"
learning_rate: 2
warmup_steps: 8000
decay_method: "noam"
adam_beta2: 0.998
max_grad_norm: 0
label_smoothing: 0.1
param_init: 0
param_init_glorot: true
normalization: "tokens"

# Model
encoder_type: transformer
decoder_type: transformer
enc_layers: 6
dec_layers: 6
heads: 8
hidden_size: 512
word_vec_size: 512
transformer_ff: 2048
dropout_steps: [0]
dropout: [0.1]
attention_dropout: [0.1]
```

RESULTS

Notably absent from the city are fortifications and military structures.
The city is absent from the military structures and buildings are absent .

After she has finished her wrestling career, James plans to own a farm and be an equine trainer.
After her own career , James has finished wrestling career and plans to be a farm trainer .

Thalassa is irregularly shaped and shows no sign of any geological modification.
Thalassa is shaped with no sign of any geological change and shows .

Helene transitioned into a "hybrid" storm with both tropical and extratropical characteristics that afternoon, with both a deep warm core and an asymmetric, frontal-like appearance.
Helene turned into a tropical storm , an extratropical storm , with both warm characteristics that afternoon and a deep warm core .

Metrics

Dataset	SARI	SacreBLEU
Asset	32.20	39.3
Turk	26.3	36.73

Different types of metrics to evaluate the simplified text:

1) BLEU (Bilingual Evaluation Understudy)

1) SARI (Simplification Automatic evaluation Measure through Semantic Annotation)

   SARI is a more relevant metric than BLEU for sentence simplification task

Limitations :

1. Supervised : It’s expensive to create data
1. Subjective : Human evaluation is still necessary to validate the text
1. Data : Unavailability of ideal target sentences
1. Resources : Computationally expensive

Appendix

Alternative Machine Translation frameworks

Name	Language	Framework	Status
Tensor2Tensor	Python	TensorFlow	Deprecated
FairSeq	Python	PyTorch	Active
Nmt	Python	TensorFlow	Deprecated
OpenNMT	Python/C++	PyTorch/TensorFlow	Active
Sockeye	Python	MXNet	Active
Nematus	Python	Tensorflow	Active
Marian	C++	–	Active
THUMT	Python	PyTorch/TensorFlow	Active
NMT-Keras	Python	Keras	Active
Neural Monkey	Python	TensorFlow	Active

Thank you