Simplifying Text Through Neural Networks and a Sentence Fusion Graph

Tina Giorgadze Elliot Harris Nadia Mehjabin Advisor: Sven Anderson. CS Department, BSRI 2022

Abstract

Sentence-based text simplification reduces the lexical, semantic, and syntactic complexity of a sentence while maintaining most of its original meaning. It has numerous applications and benefits, such as helping young and non-native speakers and making medical or legal documents more readable. Our research builds upon previous work by combining expert knowledge with existing sentence simplification pipelines based on neural machine translation. Specifically, we take existing pairs of original and simplified sentences and simplify the original by identification of complex words, generation of alternatives, and then a selection and ranking of those alternative words. We simplify sentence structure using a set of rules that break down each sentence into multiple sentences each containing only one main idea. The transformed data is used to train neural networks that generate further simplifications, primarily via lexical simplification. These sentences are fused to generate new simplified sentences, which are then ranked, yielding a final optimal simplification. Simplifications will be evaluated using human workers on Amazon Mechanical Turk.

Lexical Simplification

Type	Potential(%)	Precision(%)	Recall	Fscore
Paetzold Generator	.567	.09	.13	.11
Kauchak Generator	.609	**.189**	.153	**.169**
Wordnet Generator	.649	.13	.155	.14
All 3	**.89**	.10	**.33**	.161

Figure 4: *The most successful aspect of our Lexical Simplification Pipeline was our generation of target replacement words. From the chart above, we saw strong results from the Kauchak Generator implementation from the LEXenstein framework (Paetzold, G., & Specia, L. (2015)).*

LSTM and Transformer Model

The LSTM and Transformer Models are Neural Network Models that are heavily used in the field of Natural Language Processing and text Simplification. The LSTM (or Long Short-Term Memory) was one of the first neural nets to fully utilize the idea of training attention (which words, structures are more important to remember when processing text). The Transformer Model surfaced as an improvement to the LSTM, in the paper "Attention is All You Need". While Both models are known to be the best in store for Text Simplification, we tried to improve their already promising results by placing them inside an expert system pipeline (Figure 3).

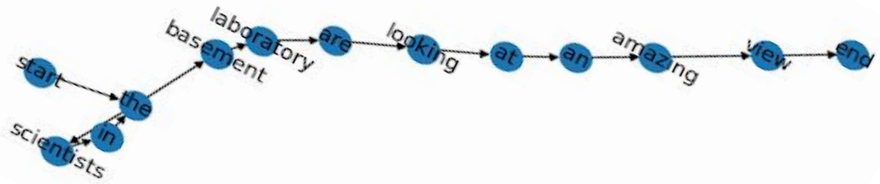

Figure 1: an example of sentence fusion. Sentences fused: The scientists in the basement laboratory are looking at an amazing view. (original sentence)

1. The view from the basement laboratory is breathtaking.
2. The scientists are in a basement laboratory, but they have an amazing view.
3. The researchers are in an underground lab, but they have an astonishing view.
4. The scientists in this basement laboratory are all looking at a smartphone.
5. The young scientists in this basement laboratory are all gathered around a smartphone.

Sentence Fusion

The last step in our research was the construction of a sentence fusion graph. The program takes in a sentence and its original simplifications as inputs, and constructs a graph where each node is a word from given sentences, and an edge occurs between every two adjacent words in the given simplifications. Then, we add weights to the graph, which are proportional to how far two words occur from each other. The weight is also inversely proportional to the frequencies of the words, allowing us to find common words that occur closely together. (figure 2). Next, we use Djikstra's short path algorithm to look for new simplifications. Finally, to rank our new simplifications, we trained a binary classifier to look at vector representations of simplified sentences and decide which corresponds with a simpler sentence.

$$w_{e_{i,j}} = \frac{f(i)+f(j)}{f(i)\times f(j)\times \sum_{s\in S} diff(s,i,j)^{-1}}$$

Figure 2: Edge weight formula between words i and j

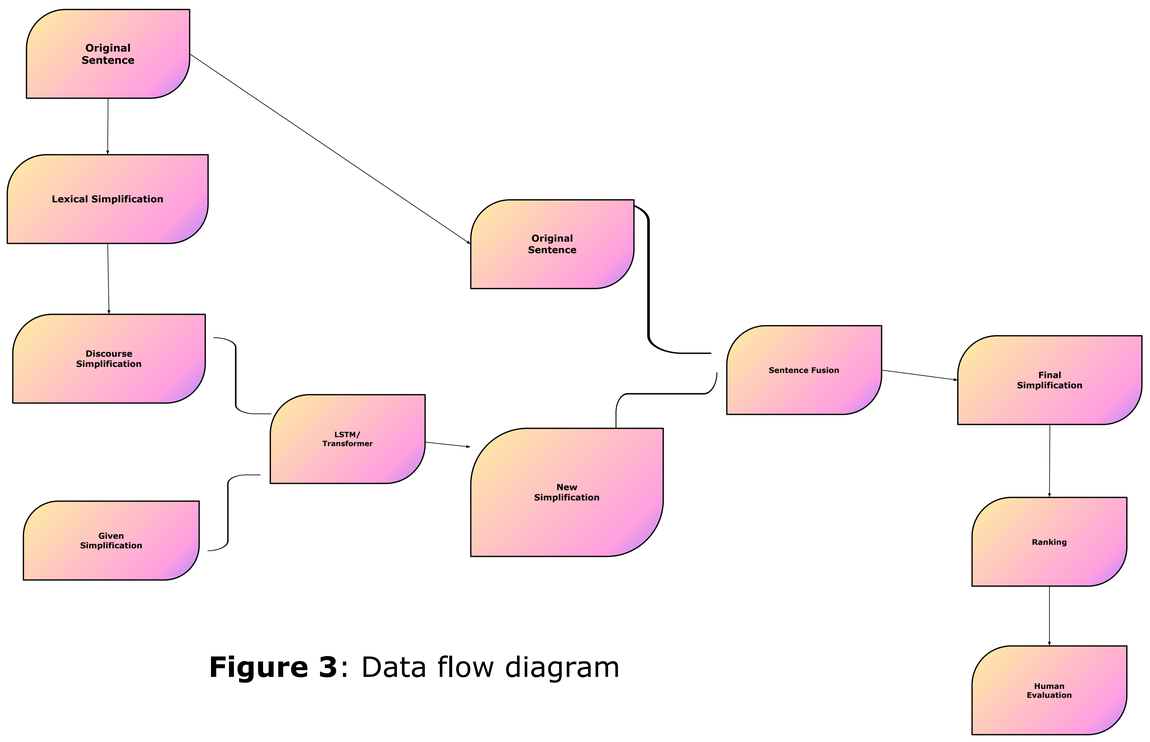

Figure 3: Data flow diagram