Bad2Worse: A Neural Translation Framework for Text Un-simplification

Aditya Vaidya and **Antony Yun**
Department of Computer Science
University of Texas at Austin
Austin, TX 78712
{avaidya, antony.yun}@utexas.edu

Abstract

While text simplification is a popular field of NLP research, we see a gap in the literature for a related, but novel take on this task: text un-simplification. In this work, we implement and propose several models that attempt to perform this task. We use train on and evaluate with the Newsela dataset, an expert-crafted corpus of simple and complex sentences. Our models use sequence-to-sequence ("Seq2Seq") architectures to increase the complexity of their input. We finally discuss how measuring performance in our task is different from the traditional problem. While this task may have little practical use outside of memes, we used this toy problem to learn more about recent work in text simplification.

1 Introduction

Text simplification is a common task and an area of active research within natural language processing. This task involves decreasing the complexity of the input by deleting extraneous clauses, paraphrasing complicated phrases, and resolving complicated word order. State-of-the-art solutions model text simplification as a machine translation (MT) problem, where the source language is "complicated" English and the target is "simple" English (Xu et al., 2015; Narayan and Gardent, 2014). This enables the use of classic MT paradigms, from phrased-based translation systems (Wubben et al., 2012) to neural networks, like variants of the sequence-to-sequence model (Sutskever et al., 2014).

In this paper, we build systems to increase the verbosity of a phrase or sentence, which we formulate as the inverse of text simplification. While this doesn't have as many downstream applications, we used this task to motivate us to learn more about simplification techniques by applying them to a non-standard problem. To the extent of our literature review, this are the first systems of their kind.

This task was inspired by a meme format that become popular in 2016,[1] in which there are two columns, with an image on the left and text on the right. As one reads from top to bottom, the left column becomes more abstract, and the right columns becomes increasingly more verbose. Figure 1 has an example in this format.

Our ultimate goal is to build a model that, given a starting phrase or sentence, can produce the increasingly verbose language that would appear in the second column. A system that performs well on this task would ideally add unnecessary clauses and phrases, and increase lexical complexity while maintaining grammaticality and the input's semantics. Thus, unlike text simplification, we are generating "complex" text from "simple" text. This means our task involves *adding* more information than is currently present, which may require the model to conjure up words based on what it already has. We did expect this to be the primary limitation in the performance of any models.

2 Prior Work

As stated before, we found no previous literature that attempts un-simplification. Thus, we reviewed corpora, models, and evaluation for text simplification, and discuss how they may be adapted for use in our task.

2.1 Corpora

Nearly all models require parallel ("aligned") datasets, where each simple sentence corresponds to at least one complex sentence and vice versa.

[1] Source: https://knowyourmeme.com/memes/increasingly-verbose-memes

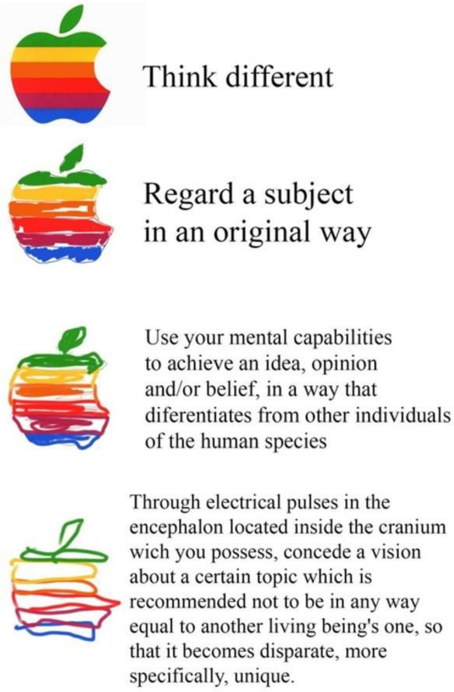
Think different
Regard a subject in an original way
Use your mental capabilities to achieve an idea, opinion and/or belief, in a way that diferentiates from other individuals of the human species
Through electrical pulses in the encephalon located inside the cranium wich you possess, concede a vision about a certain topic which is recommended not to be in any way equal to another living being's one, so that it becomes disparate, more specifically, unique.

Figure 1: Example "increasingly verbose", or "deconstruction", meme using an Apple advertising slogan. Source: https://www.pinterest.com/pin/720998221564412691/

The following are common aligned datasets that we have identified:

- Parallel Wikipedia (PWKP, Zhu et al. (2010)) – This corpus contains aligned sentence pairs from English Wikipedia and Simple Wikipedia.
- English Wikipedia and Simple English Wikipedia (EW-SEW, Hwang et al. (2015)) – This corpus is similar to PWKP, but also rates the quality of the alignment in terms of semantics.
- Newsela (Xu et al., 2015) – The Newsela dataset consists of news articles that have been rewritten by experts at five different grade levels for children.

These corpora all include a pairing of simple and complex texts, with similar meanings. In traditional text simplification papers, researchers train a model to generate the simple text from complex text. For our purpose, we plan to reverse the order, and instead produce complex text from simple text.

We decided against using PWKP and EW-SEW due to possible alignment errors, which are artifacts of Simple and traditional English Wikipedias not written to be aligned. These datasets were programmatically aligned, and as such, some aligned pairs do not have the same semantics (Amancio and Specia, 2014). Some have claimed that the prevalence of this corpus, despite all its flaws, has stifled progress in the whole field of simplification (Xu et al., 2016).

The Newsela dataset, however, has pairs that are constructed to be roughly aligned in content, so the semantics of parallel simple and complex sentences are nearly the same. From our inspection of the data, the complex sentence always has *at least* as much information as the simple sentence. In some cases, there is information added, but in other cases, the only change is structural in nature. (See Table 3.)

We ultimately used the Newsela dataset. Though its data quantity is lower ($\approx$ 70k vs $\approx$ 108k), we believe its greater consistency and overall quality will result in better models over using the Wikipedia-based datasets.

2.2 Models

Recent text simplification literature seems to be neural network-based, but its foundations were non-neural. Woodsend and Lapata (2011) state that most previous work used hand-crafted rules. These authors in particular learn a grammar from a corpus, and then use integer linear programming to order all possible rewrites that were generated from the grammar.

Nisioi et al. (2017) claim to be the first application of sequence-to-sequence ("Seq2Seq") neural networks to text simplification, and report state-of-the-art scores in human evaluation. They split their model's activity into two parts: lexical simplification and content reduction; they claim to be the first neural translation system for both. Zhang and Lapata (2017) use deep reinforcement learning for text simplification, by treating an encoder-decoder model as a network; and using simplicity, relevance, and fluency as its reward functions.

We decide to implement a vanilla Seq2Seq neural network for the sake of simplicity. We discuss the final models we use in Sections 3 and 4.

2.3 Performance Metrics

One of the earliest metrics for determining readability was the Flesch-Kincaid grade level test, which outputs the grade level required to read given sentences – higher scores indicate more "complex" sentences, and lower indicate simple sentences. It is a function of the input sentences' average length (in words) and average number of syllables per word. For simplification systems, the Flesh-Kincaid metric is often gamed by just splitting sentences. Wubben et al. (2012) even show that the metric is *negatively* correlated with "simplicity", and instead suggest using BLEU (Papineni et al., 2002).

BLEU is a metric that we previously discussed in class. It was originally created for machine translation, but it is commonly used in text simplification. Wubben et al. (2012) show that the metric is correlated with preserving grammaticality and semantics.

SARI is a metric that explicitly evaluates the *quality* of the words that are kept, added, or deleted by the system (Xu et al., 2016). The metric requires the dataset to have multiple valid simplifications. Though this metric is increasingly popular, some claim that ungrammatical sentences may still yield high SARI scores (Vu et al., 2018). The dataset used in Xu et al. (2016) does indeed have multiple reference simplifications. Since our problem involves reversing the translation direction of standard datasets, we instead need multiple vaid *complex* sentences for each simple sentence. Unfortunately, no datasets we found met this requirement.

Due to resource constraints, we did consider any metrics that required large-scale human evaluation. In this paper, we use BLEU since it is the standard in text simplification, and is correlated well with important linguistic qualities. Despite its issues, we also use Flesch-Kincaid reading level as an "absolute" measure of sentence complexity, since BLEU is a relative measure that has no notion of "simplicity": it only tells the number of n-grams used by the output in the test set.

3 Our Seq2Seq Implementation

First, in PyTorch we implemented our own version of the Seq2Seq model, first presented in Sutskever et al. (2014). The model consists of two RNNs connected by a single hidden state: an encoder and a decoder RNN. The encoder takes the input tokens and outputs to the hidden state; the decoder takes the hidden state as input and produces probability distributions over tokens. As in the original paper, we use a 2-layer LSTM for both the encoder and decoder networks. For decoding the output of the decoder network, we iterated over the probabilities of outputting each token, and greedily selected the most likely token until we hit the end-of-sequence token, `<eos>`. We initialized the model with pretrained GloVe embeddings (Pennington et al., 2014) and fine-tuned during training.

4 OpenNMT

In the interest of time, we only implemented a basic Seq2Seq model from scratch. We used Harvard's OpenNMT (Klein et al.) neural machine translation implementation to test more complicated model architectures.

4.1 Vanilla Seq2Seq

We trained a corresponding Seq2Seq model using OpenNMT analogous to our own implementation. We used a word embedding size of 500, 2 layer bidirectional LSTM encoder, 2 layer unidirectional LSTM decoder, LSTM hidden size of 500, batch size of 64, batch normalization, dropout of 0.3, and SGD learning rate of 1.0.

4.2 Global Attention

One of the main issues we noticed with the Seq2Seq models was the loss of topical information. The generated sentences generally fit the same theme as the input text but didn't always maintain named entities or details.

In an attempt to address this problem, we also trained a model using the global attention mechanism described in Luong et al. (2015). A traditional RNN decoder takes the previous cell's hidden state and the word embedding of the previous word as input. The hidden state contains some notion of history that remembers previous words in the input and generated output, but empirically RNNs perform poorly beyond a few time steps.

The attention unit attempts to learn an alignment probability distribution across each of the input tokens. An alignment vector is generated by a softmax over the dot product of the current decoder hidden state and each of the encoder hidden states. The attended context vector is created by taking a weighted average of encoder hidden states based on their alignment values. This context vector is then concatenated with the decoder hidden state and fed into a softmax layer to obtain a probability distribution for the output token.

Our Seq2Seq with Global Attention model uses the mechanism described above while keeping the rest of the model architecture and hyperparameters constant.

4.3 Global and Copy Attention

While attention helps incorporate more of the source text into the target prediction, these models still exclude specific factual information. Copy attention attempts to solve this issue by providing a mechanism to copy words from the source text. The Pointer-Gen + Coverage architecture (See et al., 2017) uses a generation probability to combine the probabilities of copying a word from the source and generating a word through the sequence model. It also penalizes repetition of outputs through the coverage technique, which adds a loss term that compares how much attention a word has received thus far with the current attention distribution.

Our model incorporates copy attention into the previous section's architecture.

5 Results

For each model, we measured BLEU score compared to the ground truth target complex sentences, BLEU score compared to the source simple sentences, Flesch-Kinkaid grade level, and sentence length, averaged across the test dataset. The "Source" model, which we include as a baseline, returns the source text as the prediction. The "Source" and "Target" columns also include statistics about the data for comparison with the actual models.

BLEU score already incorporates a pairwise comparison with the source and target, but the other metrics do not. We ran pairwise t-tests comparing each model's reading level and sentence length with those of both the source and target tests. We found that $p \ll 0.01$ for all of the tests, so the differences shown are statistically significant.

We evaluate the OpenNMT models and our own model separately, since the data was split differently. While the scores across these two frameworks aren't directly comparable, they provide a rough estimate. Table 1 shows our model's results, and Table 2 shows the OpenNMT model results.

Our sequence model doesn't achieve a very high BLEU score relative to the baseline, but it does significantly increase the reading level and sentence length to about that of the target. The OpenNMT sequence model has similar BLEU score, but the attention and copy attention models' BLEU scores are close to the baseline. However, all three OpenNMT models have reading levels and sentence lengths less than those of both the source and target texts.

We also include a few sample translations from the test set for our Seq2Seq model in Table 3. Each example is annotated with our observations on how the model's output relates to the simple and complex text. We identified a few common themes - some examples were coherent and added arbitrary extra details, some examples were slight paraphrases, some were thematically similar but didn't preserve any content, and some were just nonsensical or ungrammatical.

6 Analysis and Discussion

Even though the two attention models have the highest BLEU scores, we don't believe that they are learning anything useful un-simplification. In fact, they are likely just approximating the identity function, as their high source BLEU scores of almost 0.9 show very high similarity to the source. Most importantly, since their complexity metrics are slightly below that of the source, these models were rewarded during training for making slight modifications to the input and thus don't serve as a good text un-simplification system.

While our implementation had a significantly lower BLEU score, it did succeed in increasing the complexity of the input to match the complexity of the complex dataset. This improvement did come at the expense of preserving semantic meaning, as evidenced by the low BLEU score and often nonsensical outputs. The translated sentences were often similar in theme to the input but sometimes failed to include any consistent details from the input sentences. However, obtaining a BLEU score above the baseline isn't feasible since this would require randomly guessing the right details to add, so we believe BLEU as a metric doesn't accurately convey a model's performance on this task.

After qualitatively inspecting the results of our model, even though the results aren't particularly meaningful, we found them quite entertaining. We estimate that the majority of translations were at least thematically similar, and often semantically similar, to the source. Sometimes the model does pull details out of thin air, which is not ideal for a practical system but does accomplish our goal of

	Source	Target	Seq2Seq
Target BLEU	0.455	–	0.110
Source BLEU	–	–	0.101
FK reading level	8.565	10.761	10.630
Sentence length	106.715	136.602	139.360

Table 1: Evaluation for our Seq2Seq model. BLEU is evaluated on a 0-1 scale. Flesch-Kincaid grade level is usually between 0 and 12 but may have outliers beyond that range.

	Source	Target	Seq2Seq	Seq2Seq + Attn	Seq2Seq + Attn + Copy
Target BLEU	0.426	–	0.089	0.394	0.386
Source BLEU	–	–	0.187	0.892	0.876
FK reading level	7.847	10.148	6.537	7.628	7.424
Sentence length	94.142	123.774	84.358	90.020	87.661

Table 2: Evaluation for OpenNMT models. BLEU is evaluated on a 0-1 scale. Flesch-Kincaid grade level is usually between 0 and 12 but may have outliers beyond that range.

creating amusing content. This model is not robust enough to consistently generate increasingly verbose memes, but on rare occasion it produces surprisingly good output.

We're not exactly sure why our implementation increased the complexity but OpenNMT's Seq2Seq did not. Compared to our model, the OpenNMT model had a larger hidden state, had extra layers like dropout and batch normalization, and used beam search when translating. Our intuition is that mainly beam search (but possibly some combination of the rest of the differences) encouraged the OpenNMT model to pick shorter sentences.

7 Conclusion

Most of our project was spent implementing Seq2Seq in PyTorch.[2] When we realized we might not be getting the results we wanted, we wanted to implement new features into our models, like attention. To avoid re-inventing the wheel, we switched to using OpenNMT, which did allow us to experiment with different features, namely attention.

In this work, we focused on systems that worked only with English. In our meme literature review, we only found English memes, but we are interested in seeing our models' performance on other languages in the future.

Through this project, we have realized the reason why nobody has attempted this task: This task is almost impossible to do well. We also weren't satisfied with any of the datasets we had, due to either lack of quantity (Newsela) or quality (PWKP and EW-SEW), and feel that it did end up hurting our models. Text simplification generally maps from a higher-dimensional space (the complex sentence) to a lower-dimensional space (the simple sentence). However, our un-simplification task maps from a lower-dimensional space to a higher one; this means it's theoretically impossible to express all elements of the higher space with the lower. More concretely, for our task, this means it's difficult for the model to pull the ground truth's label out of thin air. As a result, the model overfits to details from the training data.

If we formulate the problem as the inverse of text simplification by using neural MT systems to translate from simple to complex text, the resulting models don't do anything particularly interesting. Using state of the art techniques like attention and copy attentifon may actually hurt performance, since they discourage generating arbitrary text. We also found that BLEU score doesn't accurately measure the performance of a text un-simplification system. Average sentence length and reading level are good starting metrics, but future work could further investigate evaluation techniques. We could also consider rethinking our training loop to combine some notion of both complexity and content, since our current loss function emphasizes the content but encourages conservative copying.

In sum, this paper explores applying state of the art machine translation techniques to a previously unexplored domain. It combines both

[2] Our GitHub repository: https://github.com/a-yun/nlp-text-complication

hands-on implementation to understand the underlying model code, as well as higher level experimentation with various architectures. While the results aren't very conclusive, they provide a good baseline for the task and set the stage for future research, however unlikely it may be. Even if this domain isn't further pursued by ourselves or others, we found it to be a unique and humorous problem to work on.

This work would gain from human evaluation, especially since our original use case was for a human-dominated task: creating memes. We also never evaluate the "memey-ness" of our results; doing so would tie this work back to its origins. Finally, our model *does* show why meme creation may be dominated by humans for the foreseeable future.

Acknowledgments

This paper is the report for our final project for CS 378: Natural Language Processing, taught at UT Austin in Spring 2019 by Greg Durrett. We thank the teaching staff for their guidance throughout the semester.

References

Marcelo Amancio and Lucia Specia. 2014. An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 123–130, Gothenburg, Sweden. Association for Computational Linguistics.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard Wikipedia to simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, New Orleans, Louisiana. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

	Observations	Simple	Complex	Our Translation
1	Thematically similar	Government agencies have inspectors .	Federal agencies have inspector generals and Congress routinely investigates such claims .	Military agencies have agreed that the state ' s train has been fired last year using federal residents and restaurants .
2	Extra details	21 poison gas attack .	21 chemical weapons attack .	21 chemical weapons attack on a rebel - held Damascus suburb .
3	Extra details	Jeffrey Flocken strongly disagrees .	Jeffrey Flocken disagrees wholeheartedly .	Jeffrey Flocken , the North American regional director of the International Fund for Animal Welfare , disagrees wholeheartedly .
4	Grammatical issues	Many share that feeling .	Many share those sentiments .	Some echoed share , saying , even if there ' s a little more challenging issues .
5	Nonsense	Ki calls that morally wrong .	Ki finds that morally wrong .	Ki practices takes great for us .
6	Nonsense	And interest is growing .	Interest is growing : Supporters hope to start four more programs this year .	And sometimes the country gets less and less fat .
7	Nonsense	He said these people are wrong .	The museum ' s architect , Jean Nouvel , said these people are wrong .	He said some people are part of what ' s not going to do .
8	Paraphrase	Most of the women are single .	Most of them have long been single or were recently in failed relationships , an important detail because many insist their businesses are succeeding in part because of their independence on the home front .	Most of them are single .
9	Paraphrase	California is thinking about taxing soda .	California is considering a penny - per - ounce tax on sodas .	California is considering a soda tax .
10	Thematically similar	Our country is obsessed with weight .	Americans are obsessed with weight , with many people trying to lose their extra pounds .	Our student is adopted by critical change , a reflection of the profound message to be intense .

Table 3: Sample translation by our Seq2Seq model on the test set. Each example includes the simple text, the true complex text, our translation, and our observations about how the translation relates to the source text.