Sequence-to-Sequence Models for Automated Text Simplification

Robert-Mihai Botarleanu[1], Mihai Dascalu[1,2], Scott Andrew Crossley[3], and

Danielle S. McNamara[4]

[1] University Politehnica of Bucharest

[2] Academy of Romanian Scientists

[3] Georgia State University

[4] Arizona State University

Sequence-to-Sequence Models for Automated Text Simplification

Abstract

A key writing skill is the capability to clearly convey desired meaning using available linguistic knowledge. Consequently, writers must select from a large array of idioms, vocabulary terms that are semantically equivalent, and discourse features that simultaneously reflect content and allow readers to grasp meaning. In many cases, a simplified version of a text is needed to ensure comprehension on the part of a targeted audience (e.g., second language learners). To address this need, we propose an automated method to simplify texts based on paraphrasing. Specifically, we explore the potential for a deep learning model, previously used for machine translation, to learn a simplified version of the English language within the context of short phrases. The best model, based on an Universal Transformer architecture, achieved a BLEU score of 66.01. We also evaluated this model's capability to perform similar transformation to texts that were simplified by human experts at different levels.

Keywords: Natural language processing, Text simplification, Paraphrasing, Sequence-to-sequence model

Sequence-to-Sequence Models for Automated Text Simplification

Robert-Mihai Botarleanu[1], Mihai Dascalu[1,2(✉)],
Scott Andrew Crossley[3], and Danielle S. McNamara[4]

[1] University Politehnica of Bucharest, 313 Splaiul Independentei,
060042 Bucharest, Romania
`robert.botarleanu@stud.acs.pub.ro`,
`mihai.dascalu@cs.pub.ro`
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3,
050044 Bucharest, Romania
[3] Department of Applied Linguistics/ESL, Georgia State University,
Atlanta, GA 30303, USA
`scrossley@gsu.edu`
[4] Department of Psychology, Arizona State University, PO Box 871104, Tempe,
AZ 85287, USA
`dsmcnama@asu.edu`

Abstract. A key writing skill is the capability to clearly convey desired meaning using available linguistic knowledge. Consequently, writers must select from a large array of idioms, vocabulary terms that are semantically equivalent, and discourse features that simultaneously reflect content and allow readers to grasp meaning. In many cases, a simplified version of a text is needed to ensure comprehension on the part of a targeted audience (e.g., second language learners). To address this need, we propose an automated method to simplify texts based on paraphrasing. Specifically, we explore the potential for a deep learning model, previously used for machine translation, to learn a simplified version of the English language within the context of short phrases. The best model, based on an Universal Transformer architecture, achieved a BLEU score of 66.01. We also evaluated this model's capability to perform similar transformation to texts that were simplified by human experts at different levels.

Keywords: Natural language processing · Text simplification · Paraphrasing · Sequence-to-sequence model

1 Introduction

The process of simplifying texts affords better comprehension on the part of struggling readers. Text simplification generally involves manipulation at the syntactic, lexical, and discourse level. All simplified texts share the same goal: reducing a reader's cognitive load and increasing text comprehensibility on the part of the L2 reader [1, 2]. The basis for text simplification is the notion that if written content is accessible, then beginning level readers, such as second language (L2) readers, can use the input to better test and confirm language hypotheses [3]. In general, much of the language to

which beginning level readers are exposed has been simplified to make it easier to comprehend. For instance, most readings provided to L2 students contain less sophisticated words, fewer rare words, greater syntactic complexity, and more explicit cohesive devices such as connectives or lexical overlap between text segments [1, 2]. However, in almost all cases, a human has to manually simplify the text at the grammatical, syntactic, morphological, or lexical levels [4].

The aim of this paper is to propose a novel method of automatically simplifying texts using sequence-to-sequence Machine Learning models in order to paraphrase certain expressions into easier to understand, equivalent forms. Such an approach has strong potential to aid practitioners, teachers, and textbook writers to better meet the needs of students with lower reading skills.

2 Method

2.1 Corpora

Three datasets were used in the simplification algorithm. First, phrases and paraphrases were collected from the ParaPhrase DataBase (PPDB) [5], which consists of English pairs of phrases and paraphrases, with their associated alignment and entailment properties, with three types of paraphrases: lexical, phrasal and syntactic. For the purpose of this project, the PPDB XXXL English pack was filtered such that only those pairs of source-target phrases that correspond to equivalence entailments remained, with the target text being chosen as the one to maximize the Dale-Chall readability formula [6].

The second source of simplified data came from WordNet synonym sets. The WordNet lexical database [7] contains synsets (i.e., sets of synonyms) which can be used to generate synonym pairs by intersecting the synsets of various dictionary terms. Using these, we supplemented our paraphrasing data with additional pairs of synonyms to expand the number and range of potential rephrases. Age of acquisition (AoA) scores were used for establishing a simplification criterion (i.e., we selected which words in the synonym set were easier to understand based on AoA scores).

Another dataset integrated into the corpus consists of sentence aligned pairs between the Simple English Wikipedia entries and their corresponding English Wikipedia entries [8]. This corpus has been previously used for textual simplification and presents a good diversity of simplified sentence pairs.

The three simplified paraphrase sources in our corpus have significant differences when it comes to the scope and nature of the simplifications they provide, allowing for more robust model development. Synonyms from Wordnet tend to be only one word long, while PPDB typically has phrases of 6 to 8 words in length and the Simple Wikipedia aligned dataset uses entire phrases.

2.2 Model Architectures

The Transformer we used [9] followed an encoder-decoder architecture. The inputs consisted of sequences of word embeddings, which were then modified by adding a

positional encoding that uniquely identifies each position in the text. The resulting embeddings were processed by a multi-head attention layer that consists of a self-attention distributed across a number of heads. Attention computes the compatibility function of a query Q given a set of corresponding key-value pairs (K-V). These relationships modeled by self-attention do not necessarily correspond to those typically understood in natural language (e.g., syntactic structure, coreferences etc.), but are rather some latent dependencies that arise from the text.

A variation of the Transformer is the Universal Transformer [10], an extension of the original architecture that is Turing complete. The Universal Transformer uses for recurrence either a separable convolutional or a neural network with a rectified linear unit activation and two affine transformations [10].

3   Results

BLEU scores [11], one of the frequently employed metrics for machine translation, were used to evaluate the models. BLEU scores range from 0 to 100, where 100 indicates that the translation is identical to the reference translation. The BLEU score is usually formed as a geometric mean of the individual n-gram precision scores combined with a brevity penalty, assigned so as to discourage shorter translations. In addition to the deep learning models described previously, the BLEU scores for a "Repeater" provide an estimate of the similarity between the normal and simplified phrases. Both the evaluation and the model training were conducted using the tensor2tensor library [12] (Table 1).

Table 1. BLEU scores for the tested models.

Model name	Train set	Development set	Test set
Repeater	59.72 (baseline)	60.24 (baseline)	60.24 (baseline)
Transformer	**78.76 (+19.04)**	64.92 (+4.68)	64.71 (+4.47)
Universal Transformer	69.99 (+10.27)	**66.00 (+5.76)**	**66.01 (+5.77)**

Transformer-based models attain BLEU scores that indicate good generalization, with the Universal Transformer model presenting less overfitting. Simplification is only performed on phrases instead of paragraphs or the whole text because the data present in the corpus is, at most, limited to sentences. Table 2 presents examples of paraphrase suggestions generated by the Transformer model.

Table 2. Sample paraphrases generated for an input essay in ascending order of BLEU scores.

Phrase	Reference simplification	Paraphrase choices	BLEU score
In problems	To the issues	In trouble, Because of problems	30.32
Represents the only	Was the sole	Is the only, Are the only	38.00
Errors that	The mistakes that	Mistake that, Mistakes that	42.88
And a violation of	A breach of the	And a breach of, And the breach of	59.46
Still underway	Still in progress	In progress, Still running	60.65
Provision of access	Give access	For access, Terms of access	70.71
Relevant provisions of the charter of	The provisions of the charter of	Provisions of the charter of, of provisions of the charter of	81.7

As a post-hoc analysis, we used a corpus of 100 texts [4] which were each simplified to three levels (advanced, intermediate, and elementary) to better assess the performance of the model on real world texts. We measure the uncased BLEU score for the Transformer model paraphrases generated on the advanced texts and compare them to their intermediate and elementary forms. We also try various probability thresholds which indicate the minimum joint probability of a candidate simplification. All evaluations are performed using the Transformer model. The results from Table 3 indicate that the more alterations the model is allowed to make (lower thresholds), the worse it performs. One reason for this may be the manner in which the human experts perform alterations in these texts, such as the use of sentence fusion, phrase splitting, phrase reordering. and the elimination of certain sequences of text wholesale. These alterations are beyond the capabilities of what our model has been trained to perform, although they provide insight into future directions for analysis.

Table 3. BLEU scores for the Transformer model's translations on the real-life testing corpus.

Threshold	Intermediate	Elementary
0.0	59.17	39.01
0.05	65.96	42.09
0.15	67.80	42.64
0.70	69.76	43.88

4   Conclusions

In this paper, we analyzed the capabilities of modern Neural Machine Translation models in the context of text simplification, via paraphrasing. By expanding on previous work done by Kauchak [8], we generate a text simplification dataset that includes samples of varying scopes: synonyms, few word idioms, and entire phrases. We set up our learning problem such that the models are trained to transform an English sequence into another, equivalent, sequence with higher readability. We then train Machine

Translation architectures consisting of encoder-decoder Neural Networks in order to evaluate how well they can transduce text written in English into a simpler form.

Our results suggest that human modifications to the text diverge from those found in the textual simplification corpora we used. The reference simplifications tended to include stylistic and structural alterations, such as fusing or breaking up phrases, eliminating portions of the text, and changing the structure of the document.

Our constructed dataset expands on those commonly used in text simplification and we show that the neural models examined in this study are indeed capable of generalizing on these data. A future avenue of research for this topic is the construction of a dataset that is better aligned with the kind of alterations humans make during essay simplification. This might require the addition of syntactic parsers, part of speech taggers, and tools that can measure elements of text cohesion including vectors of connectives and semantic representations across texts. This work and future endeavors of this kind have strong potential to make crucial contributions to students' capacity to understand and learn from text - a concern of a broad range of practitioners and researchers.

References

1. Crossley, S.A., McNamara, D.S.: Assessing L2 reading texts at the intermediate level: an approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). Lang. Teach. **41**(3), 409–429 (2008)
2. Crossley, S.A., Louwerse, M.M., McCarthy, P.M., McNamara, D.S.: A linguistic analysis of simplified and authentic texts. Modern Lang. J. **91**(1), 15–30 (2007)
3. Hatch, E.M.: Second Language Acquisition: A Book of Readings. Newbury House Pub, Rowley (1978)
4. Allen, D.: A study of the role of relative clauses in the simplification of news texts for learners of English. System **37**(4), 585–599 (2009)
5. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB: the paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 758–764 (2013)
6. Chall, J.S., Dale, E.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books, Northampton (1995)
7. Miller, G.A.: WordNet: A lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
8. Kauchak, D.: Improving text simplification language modeling using unsimplified text data. In: 51st Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long papers, pp. 1537–1546. ACl, Sofia, Bulgaria (2013)

9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 5998–6008 (2017)
10. Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., Kaiser, Ł.: Universal Transformers. arXiv preprint, arXiv:1807.03819 (2018)
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. ACL, Philadelphia, PA, USA (2002)
12. Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N.: Tensor2tensor for neural machine translation. arXiv preprint, arXiv:1803.07416 (2018)