Adversarial Neural Text Simplification with a Focus on the Medical Domain

Master Thesis Exposé

Lukas Abegg
lukas.abegg@hu-berlin.de

Humboldt Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik

April 07, 2020

Table of Contents

1 Introduction

The European Economic and Social Committee (2019) [1], has noted that health literacy, i.e., the ability to read and comprehend medical texts, is essential for public health. However, in a report on health literacy of the German population published by Schaeffer et al. (2018) [2] the results showed that the lack of understanding mainly stems from an inability to understand highly specialised literature written in a subject-specific manner. This is especially true for medical and pharmaceutical texts.

In order to rectify this problem and to make medical information more available to a broader audience they must be translated into a simpler equivalent. This process is called Text Simplification (TS). TS involves splitting long and syntactically complex sentences into shorter and clearer ones. However it is critical that the core message should be preserved after TS has been applied. Today, Automatic Text Simplification (ATS) is a central topic in Natural Language Processing (NLP) research not only because of its interesting challenges but also its ability to democratize specific domains such as health information [3].

Research in TS often approach the problem as Neural Machine Translation (NMT). Since the development of transformer models, they have achieved State-Of-The-Art (SOTA) performance in NMT [4, 5]. However, current SOTA NMT approaches require large parallel corpora for training, which is particularly lacking in the medical domain. Especially for the use in TS, such a biomedical specific parallel corpora is unfortunately not available. To ensure that we have enough datasets with sufficiently large parallel corpora to employ NMT on TS tasks, we evaluate our approach on a dataset with biomedical text, but also on a general TS dataset.

Pre-trained Language Models such as BERT [5] have shown to improve data-poor NLP tasks [6]. We hope similar improvements can be achieved on low-data parallel corpora. Lee et al. (2020) [6] have trained BERT on a large-scale biomedical corpora specifically for use in biomedical tasks.

NMT models traditionally are optimized with Maximum Likelihood Estimation (MLE), which results in generated text that often is too dissimilar with human translations. In this work, we want to overcome this issue by using the idea of Goodfellow et al. (2014) [7] to use a Generative Adversarial Network (GAN) to jointly train a generative Neural Text Simplification (NTS) model G that produces simplified sentences which are difficult to distinguish from human simplifications for a discriminative model D. We hope that due to the adversarial training the generative NTS model can utilize the discriminator to learn the distribution of human simplified output sentences, i.e., it learns what the golden sentences look like.

We also want to ensure that our NTS model achieves high evaluation scores

in the mainly used evaluation metrics (e.g., sentence-level BLEU [8], METEOR [9] or TER [10]). Reinforcement Learning (RL) based training objectives alone do not solve this problem [11]. Therefore, inspired by Yang et al. (2018) [12], we use a sentence-level based similarity evaluation metric (e.g., sentence-level BLEU, METEOR or TER) and a lexical simplicity based metric (SARI [13]) as objectives in addition to the naive GAN objective, for the policy gradient training based on the REINFORCE algorithm [14].

2 Problem Statement

We want to learn from SOTA approaches in general NMT and study how well they can be exploited for NTS. Therefore we want to tackle the following key challenges:

Using Transformers in NTS. Transformers have shown better performance in general NMT tasks than traditional NMT models like RNNSearch [15]. We want to study whether the same applies to NTS.

Using BERT as Word Embedding Layer in NTS. We would like to make use of two advantages of BERT: (1) We exploit the fact that BERT has contextual understanding in the hope that our model has a strong semantic preservation. (2) BioBERT [6] is pre-trained on large-scale corpora. We hope this prior knowledge will improve the results even on small datasets. We want to investigate if a domain-specific contextual language model improves the results of our NTS model.

Using Adversarial Training with a novel Reward Function. Our multi-objective based reward function aims to achieve high scores in common evaluation metrics but also to generate human-like simplifications. We experiment with different similarity evaluation metrics at sentence-level such as sentence-level BLEU, METEOR or TER in combination with SARI as additional reinforcement objectives besides the naive GAN objective. We want to investigate whether our NTS model shows a better performance in TS due to the adversarial training. We also want to find out if our reward function improves the results of our NTS model in comparison to the naive GAN objective.

Our focus lies on the training and evaluation of our proposed approach for NTS. Unfortunately, due to the lack of specific biomedical parallel corpora, we cannot focus only on biomedical text. However, our main goal is to investigate whether we can achieve SOTA results on parallel corpora with biomedical text that are better than those of generalized TS approaches.

3 Related Work

3.1 Neural Text Simplification

Recent work in Text Simplification utilizing neural networks has shown promising results [17, 18]. Štajner and Saggion (2018) [19] have shown that NTS approaches perform better than SOTA ATS approaches. They found that NTS

models have a higher accuracy rate in correct changes of medical terms and produce a more simplified output than any ATS approach. Van den Bercken et al. (2019) [16] have applied a similar architecture on medical text resulting in improvements, when using NTS instead of ATS. However, they pointed out that NTS algorithms trained on general parallel corpora perform well on general language simplification tasks, but poorly on medical text simplification [34, 35].

3.2 Transformers in NMT

Vaswani et al. (2016) [4] introduced transformers, a novel network architecture which achieves SOTA results in machine translation tasks such as WMT2014 English-German and WMT2014 English-French. This architecture allows faster training than RNN and CNN based architectures, since significantly more parallelism is allowed. Extensions and modified variants of transformer networks have already been successfully applied to several other NMT tasks [20, 21, 22]. Recently, the original architecture of transformers has been improved by adding longer attention spans to solve the context segmentation problem [32], and the RL training was stabilized by a *gating* mechanism [33].

3.2.1 Integration of BERT into Transformers to improve NMT. The language model BERT with its contextualised word representations is also based on a transformer [5]. Recent studies on NMT [23, 24, 25] started to focus on the integration of BERT into transformer models for NMT and emphasized the possible enhancements of NMT models through contextual embeddings like BERT.

3.3 Reinforcement Learning in NMT

Large improvements over baseline models through the use of RL have been reported when RL is combined with classic objectives (e.g., BLEU or ROUGE) as demonstrated in [26, 27]. Using RL combined with classic objectives has been also successfully applied in GAN approaches using adversarial training: Wu et al. (2018) [28] and Yang et al. (2018) [12] proposed using a policy gradient in the training objective to optimize directly towards a sentence-level reward.

Furthermore, RL has also been introduced in NTS by Zhang and Lapata (2017) [30] with a similar objective approach as [28, 12]. Inspired by the Minimum Risk Training (MRT) method [27, 29], they proposed a combination of BLEU score and MLE in their approach to train a NTS model for sentence simplification.

4 Approach

The core element of our approach is a NTS model which is trained in a GAN architecture with adversarial reinforcement learning. The discriminative model

D is used to provide a feedback in form of a reward, to challenge and teach our generative model G. Overall, our proposed approach can be described as follows:

Generator. The generator G generates word by word a simplified output sequence y based on the source input sequence x. We follow the work of Parisotto et al. (2019) [33] and implement our NTS model as an advanced version of the transformer network, called Gated Transformer-XL (GTrXL).

Discriminator. The discriminator D tries to classify a simplification sequence as machine-generated (y_g) or human-made (y_h). To distinguish between these two classes, D takes a simplified target sequence y as input and the corresponding source sequence x as an additional condition. We follow the work of Wu et al. (2018) [28] and Yang et al. (2018) [12] and employ a CNN architecture with a layer-by-layer convolution and pooling strategy to capture matching features on different abstraction layers. The CNN takes over the task to classify the simplified sequence y as machine-generated or human-made. Therefore it takes y and the corresponding source sequence x as input parameters.

Multi-Objective Reinforcement Learning. Our multi-objective based reward function consists of two static objectives and a dynamic objective. The objective $Q_{sl-metric}$, based on a sentence-level metric and the SARI based objective Q_{sari}, are static functions and are not updated during training. Together with our discriminator D, which is considered as the dynamic objective D, i.e., the naive GAN objective, they form the multi-objective based reward function.

BERT Embeddings. In order to use a language model which fits our biomedical context, we use a domain-specific version of BERT with pre-trained weights called BioBERT-Large v1.1 (+ PubMed 1M) [6]. The BERT model does not automatically serve word embeddings. Therefore we follow the proposal of Devlin et al. (2018) [5] and sum up the last four layers of the BERT model to get an individual embedding vector for each word. These vectors are then used for the embedding layer.

4.1 Evaluation

We conduct extensive experiments on a Wikipedia to Simple-Wikipedia simplification task containing biomedical text. This dataset was presented by van den Bercken et al. (2019) [16]. It contains 154,805 sentence-pairs in English including 2,267 sentences from the medical domain subset. This dataset is not split into training and evaluation sets. We follow the proposal of [16] to randomly select 500 medical sentence-pairs for validation and 350 medical sentence-pairs for testing. The remaining 153,955 sentence-pairs, including 1,417 medical sentence-pairs, are used for training. We compare our proposed transformer architecture with BERT embeddings pre-trained on biomedical text with the SOTA approach from [16] and investigate whether adversarial RL with our proposed reward function, which is based on a weighted combination of evaluation metrics and the naive GAN objective, encourages our model to achieve better results than a generalized TS approach.

In addition to our evaluation on a parallel corpora with biomedical text, we

employ our approach on Newsela [36], a parallel corpora of news articles in English. This parallel corpora consists of 1,130 news articles which are rewritten by professional editors to create high quality simplifications for children at different grade levels (every article is rewritten in four different complexity levels). The dataset is split into 94,208 sentence pairs for training, 1,129 sentence pairs for validation and 1,076 sentence pairs for testing. We compare our proposed approach with the SOTA approach of Zhang and Lapata (2017) [30] and their evaluation on this dataset. On this corpora we use generalized pre-trained BERT embeddings from a model called BERT-Large, Cased [5].

4.2 Training Strategy

We first pre-train G on the parallel training set using MLE. After that, we use G to generate simplifications of source sentences from the training set. We also pre-train D, using the generated simplifications in combination with the true parallel data. Both models are trained until they do not significantly improve any more. Then, we finally jointly train the generator and the discriminator.

To train G only with a policy gradient leads to a weak model performance [12]. Li et al. (2017) [31] introduced a strategy called teacher forcing to mitigating this problem by updating the generator with the reward of human-made references for simplifications. The teacher forcing training runs one time, once G is updated by the policy gradient training. After the generator is completely updated, we use G to generate new simplifications, which are then used again to train D. For each optimization step of G, we also run one for D. This simultaneously adversarial training continues until none of the competitors, neither G nor D, make any progress.

References

[1] European Economic and Social Committee (2019). Opinion of the European Economic and Social Committee on "Digital health literacy — for citizen-friendly healthcare in Europe in times of demographic change". In *Official Journal of the European Union (2019/C 228/01), pages 1-6.*

[2] Schaeffer, D., Vogt, D., Gille, D. and Beerens, E. M. (2018). Gesundheitskompetenz in vulnerablen Bevölkerungsgruppen. In *Monitor Versorgungsforschung, (06/2018), pages 55-59.* doi: 10.24945/mvf.0618.1866-0533.2111.

[3] Saggion, H. (2017). Automatic Text Simplification. In *Synthesis Lectures on Human Language Technologies, 10(1), pages 1-137.* doi: 10.2200/s00700ed1v01y201602hlt032.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2016). Attention is all you need. In *Advances in Neural Information Processing Systems, 2016, pages 6000-6010.*

[5] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. In *ArXiv, abs/1810.04805.*

[6] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. In *Bioinformatics, 36(4), pages 1234-1240.* doi: 10.1093/bioinformatics/btz682.

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems, 2014, pages 2672-2680.*

[8] Nakov, P., Vogel, S. and Guzman, F. J. (2012). Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING) 2012, Mumbai, India, pages 1979-1994.*

[9] Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization 2005, Ann Arbor, USA, pages 65-72.*

[10] Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L. and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas: Visions for the Future of Machine Translation (AMTA) 2006, Cambridge, USA.*

[11] Wu, L., Tian, F., Qin, T., Lai, J. and Liu, T.-Y. (2018). A Study of Reinforcement Learning for Neural Machine Translation. In *Proceedings of the 23rd conference on Empirical Methods in Natural Language Processing (EMNLP) 2018, Brussels, Belgium, pages 3612-3621.*

[12] Yang, Z., Chen, W., Wang, F. and Xu, B. (2018). Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2018, New Orleans, USA, pages 4344-4355.*

[13] Xu, W., Napoles, C., Pavlick, E., Chen, Q. and Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. In *Transactions of the Association for Computational Linguistics, 4, pages 401-415.*

[14] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning, 8, pages 229-256.* doi: 10.1007/BF00992696.

[15] Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In *ArXiv, abs/1409.0473.*

[16] van den Bercken, L., Sips, R.-J. and Lofi, C. (2019). Evaluating Neural Text Simplification in the Medical Domain. In *Proceedings of the 30th World Wide Web Conference (WWW) 2019, San Francisco, USA.*

[17] Nisioi, S., Štajner, S., Ponzetto, S. P. and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) 2017, Vancouver, Canada, pages 85-91.*

[18] Sulem, E., Abend, O. and Rappoport, A. (2018). Simple and Effective Text Simplification Using Semantic and Neural Methods. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) 2018, Melbourne, Australia, pages 162-173.*

[19] Štajner, S. and Saggion, H. (2018). Data-driven Text Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING) 2018, Santa Fe, USA.*

[20] Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M. and Liu, Y. P. (2018). Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 23rd conference on Empirical Methods in Natural Language Processing (EMNLP) 2018, Brussels, Belgium.*

[21] Kreutzer, J., Bastings, J. and Riezler, S. (2019). Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019, Hong Kong, China.*

[22] Popel, M. (2018). CUNI Transformer Neural MT System for WMT18. In *Proceedings of the 23rd conference on Empirical Methods in Natural Language Processing (EMNLP) 2018, Brussels, Belgium.*

[23] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H. and Liu, T. (2020). Incorporating BERT into Neural Machine Translation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR) 2020, Addis Ababa,*

Ethiopia.

[24] Yang, J., Wang, M., Zhou, H., Zhao, C., Yu, Y., Zhang, W. and Li, L. (2019). Towards Making the Most of BERT in Neural Machine Translation. In *ArXiv, abs/1908.05672*.

[25] Clinchant, S., Jung, K. and Nikoulina, V. (2019). On the use of BERT for Neural Machine Translation. In *ArXiv, abs/1909.12744*.

[26] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Machereyet, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H. and Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In *ArXiv, abs/1609.08144*.

[27] Ranzato, M. A., Chopra, S., Auli, M. and Zaremba, W. (2015). Sequence Level Training with Recurrent Neural Networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR) 2016, San Juan, Puerto Rico*.

[28] Wu, J., Xia, Y., Tian, F., Zhao, L., Qin, T., Lai, J. and Liu, T.-Y. (2018). Adversarial Neural Machine Translation. In *Proceedings of the 10th Asian Conference on Machine Learning (ACML) 2018, Beijing, China, pages 534-549*.

[29] Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M. and Liu, Y. (2016). Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) 2016, Berlin, Germany, pages 1683-1692*.

[30] Zhang, X. and Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 22nd conference on Empirical Methods in Natural Language Processing (EMNLP) 2017, Copenhagen, Denmark, pages 584-594*.

[31] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A. and Jurafsky, D. (2017). Adversarial Learning for Neural Dialogue Generation. In *Proceedings of the 22nd conference on Empirical Methods in Natural Language Processing (EMNLP) 2017, Copenhagen, Denmark, pages 2157-2169*.

[32] Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V. and Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) 2019, Florence, Italy, pages 2978-2988*.

[33] Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gülçehre, Ç., Jayakumar, S. M., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M. M., Heess, N. M. and Hadsell, R. (2019). Stabilizing Transformers for Reinforcement Learning. In *ArXiv, abs/1910.06764*.

[34] Adduru, V., Hasan, S. A., Liu, J., Ling, Y., Datla, V., Qadir, A. and Farri, O. (2018). Towards Dataset Creation And Establishing Baselines for Sentence-level

Neural Clinical Paraphrase Generation and Simplification. In *Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data (KDH) 2018, Stockholm, Sweden.*

[35] Kushalnagar, P., Smith, S., Hopper, M., Ryan, C., Rinkevich, M. and Kushalnagar, R. (2018). Making Cancer Health Text on the Internet Easier to Read for Deaf People who use American Sign Language. In *Journal of Cancer Education (2018), 33(1), pages 134-140.* doi: 10.1007/s13187-016-1059-5.

[36] Xu, W., Callison-Burch, C. and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. In *Transactions of the Association for Computational Linguistics, 3, pages 283-297.*