Edit Distance Based Curriculum Learning for Paraphrase Generation

Sora Kadotani[†], **Tomoyuki Kajiwara**[‡], **Yuki Arase**[†], **Makoto Onizuka**[†]
[†]Graduate School of Information Science and Technology, Osaka University
[‡]Graduate School of Science and Engineering, Ehime University
[†]{kadotani.sora, arase, onizuka}@ist.osaka-u.ac.jp
[‡]kajiwara@cs.ehime-u.ac.jp

Abstract

Curriculum learning has improved the quality of neural machine translation, where only source-side features are considered in the metrics to determine the difficulty of translation. In this study, we apply curriculum learning to paraphrase generation for the first time. Different from machine translation, paraphrase generation allows a certain level of discrepancy in semantics between source and target, which results in diverse transformations from lexical substitution to reordering of clauses. Hence, the difficulty of transformations requires considering *both* source and target contexts. We propose an edit distance between a paraphrased sentence pair as a difficulty metric in curriculum learning. Experiments on formality transfer using GYAFC showed that our curriculum learning with edit distance improves the quality of paraphrase generation. Additionally, the proposed method improves the quality of difficult samples, which was not possible for previous methods.

1 Introduction

Paraphrase generation is a task that transforms expressions of an input sentence while retaining its meaning. While there are various subtasks in paraphrase generation, formality transfer (Rao and Tetreault, 2018; Niu et al., 2018; Kajiwara, 2019; Wang et al., 2019; Kajiwara et al., 2020; Zhang et al., 2020; Wang et al., 2020; Chawla and Yang, 2020) has been extensively studied. As paraphrase generation can be regarded as a machine translation task (Finch et al., 2004; Specia, 2010) within the same language, the same models (Bahdanau et al., 2015; Vaswani et al., 2017) have been applied to a monolingual parallel corpus.

Recent studies (Platanios et al., 2019; Liu et al., 2020) have shown that curriculum learning (Bengio et al., 2009) achieves faster convergence and improved translation quality on neural machine translation. Curriculum learning designs a training process starting from easy training samples and gradually proceeds to difficult training samples. In these previous studies, curriculum learning that uses source-side features, *i.e.*, sentence length and word rarity, as a metric to determine the difficulty has improved the quality of translation.

In this study, we adopt curriculum learning to the paraphrase generation task. Paraphrasing allows a certain level of semantic divergence between source and target sentences. For example, some paraphrases only require just a small number of transformations as shown in Table 1, while some others require drastic transformations as Table 2 shows. For the former, transformation is easy because the target sentence can be generated by copying almost all the input sentence's words. For the latter, transformation is difficult because the input sentence requires replacement and reordering of clauses besides lexical and phrasal paraphrasing. Because of this feature in paraphrase generation, *difficulty* in transformations requires to consider both source and target contexts.

To address this problem, we propose to use an edit distance between a paraphrased sentence pair as a difficulty metric that approximates necessary amounts of transformations. We evaluate our method on a formality transfer task using Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018). The result of paraphrase generation from informal English to formal English confirmed the effectiveness of curriculum learning based on the edit distance. The detailed analysis revealed that the proposed method contributes to performance improvement in difficult samples regardless of the difficulty metrics, while sentence length and word rarity based methods degraded the performance.

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 229–234

Source Sentence	Target Sentence
Yeah I think it would be funny.	I think it would be funny.
I have one brother and three sisters.	I have one brother and three sisters.
Do you mean which is least horrible?	Do you mean which is **the** least horrible?
Their first two albums were **pretty** good.	Their first two albums were **very** good.

Table 1: Examples with simple transformations (bold fonts indicate words that should be rewritten)

Source Sentence	Target Sentence
whats the **name** of **the** song	**What is** the **title** of **this** song**.**
not sure thank you for the two points	**Unsure, appreciate the pair of** points**.**
no where there is no such thing	**That does not exist.**
they **just got a little agressive ;)**	**Suddenly** they **became angrier.**

Table 2: Examples with drastic transformations (bold fonts indicate words that should be rewritten)

2 Preliminary: Curriculum Learning for Neural Machine Translation

Initial curriculum learning methods for neural machine translation considered only the difficulty of the training sample (Kocmi and Bojar, 2017; Zhang et al., 2018). These methods achieved faster convergence; however, they could not improve machine translation quality after convergence. Following these studies, Platanios et al. (2019) and Liu et al. (2020) proposed a method that considers both the difficulty of the training samples and the model competence, which achieved both of faster convergence and improvement in the translation quality.

This study bases on the model proposed by Platanios et al. (2019), who introduced the model competence in machine translation. Their method defines $\bar{d}_i \in [0, 1]$ that is the difficulty score of the i-th training sample, and $c(t) \in [0, 1]$ that is the model competence at the training step t. The method trains the model using only easier training samples than the model competency at each training step. In other words, the number of training samples increases as the training proceeds. Their method improved the translation quality while reduced the training time.

Platanios et al. (2019) defined the difficulty $d(s_i)$ based on sentence length and word rarity. Here, an input sentence s_i consists of a word string $\{w_1, ..., w_{N_i}\}$. Considering translation of a long sentence is more difficult than a shorter one, the sentence length is adopted as one of the metrics:

$$d_{\text{length}}(s_i) \triangleq N_i. \tag{1}$$

Besides, they considered words that infrequently appear in a training corpus are also difficult to translate because these words have fewer learning opportunities. Therefore, Platanios et al. (2019) also adopted word rarity:

$$d_{\text{rarity}}(s_i) \triangleq -\sum_{j=1}^{N_i} \log \hat{p}(w_j), \tag{2}$$

where $\hat{p}(w_j)$ is the unigram probability of word w_j in the training corpus. The final difficulty score $\bar{d}_i$ is computed using the cumulative distribution functions of $d(s_i)$ values.

Platanios et al. (2019) defined the model competence $c(t)$ at the training step t:

$$c(t) \triangleq \min(1, \sqrt{t\frac{1 - c_0^2}{T} + c_0^2}), \tag{3}$$

where c_0 is the initial competence and T is the number of training steps estimated as necessary for convergence. They assumed that the competence is small at the beginning of training and increases monotonically as the training proceeds, which reaches the maximum value 1 when $t = T$.

3 Proposed Method

We approximate the difficulty of transformation in paraphrase generation as edit distance between a paraphrased sentence pair:

$$d_{\text{distance}}(s_i, t_i) \triangleq \text{LevenshteinDistance}(s_i, t_i), \tag{4}$$

where $\text{LevenshteinDistance}(\cdot, \cdot)$ computes the Levenshtein distance between the source sentence and the target sentence t_i. The edit distance between sentences with simple transformations like Table 1 is small, and the edit distance between sentences with drastic rewriting like Table 2 is large. Hence, our curriculum learning starts training with paraphrases with a small number of transformations and gradually learns more dynamic transformations.

Algorithm 1 Edit-distance based curriculum learning

Input: Dataset $D = \{(s_i, t_i)\}_{i=1}^{M}$, consisting of M samples, neural machine translation model θ.

Output: Trained neural machine translation model θ.

List of difficulty values $L \leftarrow \emptyset$
for $i = 1, ..., M$ **do**:
    $L \leftarrow L \cup \{d_{\text{distance}}(s_i, t_i)\}$.
end for
Compute a cumulative distribution function from difficulty values in L
for $i = 1, ..., M$ **do**:
    Compute the difficulty score $\bar{d}_i$
end for
for $t = 1, ..., T$ **do**: ▷ Curriculum learning
    Compute the model competence $c(t)$.
    Sample a data batch B_t uniformly from all $s_i \in D$, such that $\bar{d}_i \leq c(t)$.
    Train neural machine translation model θ using B_t as input.
end for

We apply the edit-distance based difficulty metric to the competence-based curriculum learning (Platanios et al., 2019) framework. The entire algorithm is shown in Algorithm 1.

4 Experiment

We evaluate the performance of edit-distance based curriculum learning on a style transfer task: paraphrase generation from informal English to formal English using GYAFC[1] (Rao and Tetreault, 2018).

4.1 Corpus and Evaluation Metric

GYAFC provides parallel sentences from two domains, Entertainment & Music (E&M) and Family & Relationships (F&R). Following Niu et al. (2018), we expand the training set by combining sentences of each domain and add the label `2formal` or `2informal` at the beginning of an input sentence. Statistics of GYAFC corpus are shown in Table 3.

As preprocessing, we used Moses toolkit[2] (Koehn et al., 2007) for tokenization and normalize-punctuation. We also used byte-pair encoding[3] (Sennrich et al., 2016) to limit the number of token types to $16,000$.

	Train	Train*	Dev	Test
E&M	$52,595$	$209,124$	$2,877$	$1,416$
F&R	$51,967$	$209,124$	$2,788$	$1,332$

Table 3: Statistics of GYAFC (Train* indicates the training set after expansion.)

On GYAFC, Rao and Tetreault (2018) reported that a correlation exists between manual annotation and BLEU (Papineni et al., 2002) scores for the task of informal to formal English transfer. Hence, we used BLEU as an evaluation metric.

4.2 Setup

As a paraphrase generation model, we implemented transformer (Vaswani et al., 2017) model using Joey NMT[4] (Kreutzer et al., 2019). Our transformer model has four-layers with a hidden size of 512 and a four attention heads for both the encoder and decoder. We used word embeddings of 512 dimensions tying the source, target, and the output layer's weight matrix. We also added dropout to the embeddings and hidden layers with a probability of 0.2. We trained using the Adam optimizer (Kingma and Ba, 2015) with the learning rate of 0.0002. The batch size was $4,096$ tokens. We saved the model every 800 updates applying early stopping with patience of five.

To evaluate the effectiveness of the edit distance[5] on curriculum learning (denoted as CL-ED), we compared to curriculum learning with sentence length (denoted as CL-SL) and word rarity (denoted as CL-WR). To compute the model competency with Equation (3), we need to set two hyperparameters of c_0 and T. We set c_0 to 0.01 and T to the number of training steps necessary for the transformer model with ordinary training reaches the 95% of the maximum BLEU score on the development set.

4.3 Results

The experimental results are shown in Table 4, where 'Baseline' is the transformer model trained without curriculum learning. In the E&M domain,

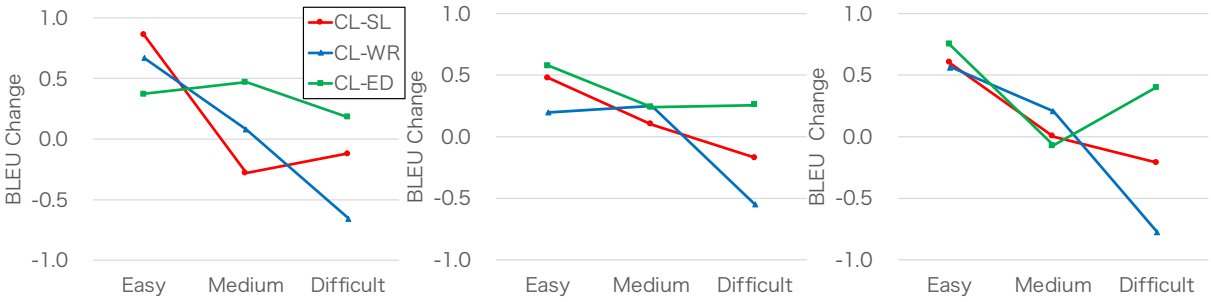
[1] https://github.com/raosudha89/GYAFC-corpus
[2] https://github.com/moses-smt/mosesdecoder
[3] https://github.com/rsennrich/subword-nmt
[4] https://github.com/joeynmt/joeynmt
[5] https://github.com/roy-ht/editdistance

	E&M	F&R
Source	49.19	50.94
Baseline	69.81	75.02
CL-SL	69.83	74.90
CL-WR	70.05	74.62
CL-ED	**70.34**	**75.41**

Table 4: BLEU scores on the GYAFC test set

Source	dead on arrival... there relationship is dead on arrival
Reference	Their relationship is dead on arrival.
Baseline	Dead on arrival, there relationship is dead on arrival.
CL-SL	Dead on arrival is dead on arrival.
CL-WR	Dead on arrival is dead on arrival.
CL-ED	The relationship is dead on arrival.

Table 5: Examples of generated sentences by each model



(a) Sentence length (b) Word rarity (c) Edit distance

Figure 1: Changes in BLEU scores compared to Baseline for each difficulty metric

CL-ED and CL-WR improved BLEU score of Baseline. In the F&R domain, only CL-ED outperformed Baseline. These results indicate that existing curriculum learning based on sentence length and word rarity is not effective in paraphrase generation. In contrast, curriculum learning with the edit distance was effective on both domains.

4.4 Discussion

We investigated which type of sentences that the curriculum learning improved their paraphrase quality. We divided all the test sets into three classes: Easy, Medium, and Difficult, of the same size (916 sentences each) using difficulty metrics of sentence length, word rarity, and edit distance, respectively. We then computed a BLEU score of each class and calculated improvements over Baseline.

Figure 1 shows the BLEU score differences of CL-SL, CL-WR, and CL-ED, compared to Baseline, respectively. Overall, the performance improvement on the Easy class is significant across the methods, which is intuitive as such sentences are easy to learn and used for training throughout curriculum learning. CL-SL and CL-WR degraded the BLEU scores on Medium class, and even deteriorated the baseline transformer on the Difficult class. In contrast, CL-ED improved the BLEU scores of Baseline even on the Difficult class, regardless of the metric of difficulty.

Table 5 shows output examples. The Baseline output almost the same sentence as the input without necessary transformations. While CL-SL and CL-WR output a sentence that does not make sense, CL-ED, which is our method, successfully paraphrases the source sentence.

5 Summary and Future Work

In this study, we applied the edit distance to curriculum learning for paraphrase generation. Experiment results on an informal to formal style transfer task confirmed the effectiveness of our method, particularly for paraphrasing difficult sentences.

Curriculum learning can be applied to any task when reasonable metrics for task difficulty are available. Transfer learning using a pre-trained model (Devlin et al., 2019; Lewis et al., 2020) has significantly improved the performance of various natural language processing tasks. In transfer learning, fine-tuning samples similar to the ones in the pre-training corpus should be easier to learn. We plan to apply our edit-distance based curriculum learning to transfer learning.

Acknowledgments

This work was supported by JST, ACT-X Grant Number JPMJAX1907, Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization. *Findings of the Association for Computational Linguistics*, pages 2340–2354.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Andrew Finch, Taro Watanabe, Yasuhiro Akiba, and Eiichiro Sumita. 2004. Paraphrasing as Machine Translation. *Journal of Natural Language Processing*, 11(5):87–111.

Tomoyuki Kajiwara. 2019. Negative Lexically Constrained Decoding for Paraphrase Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052.

Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. 2020. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8042–8049.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, pages 1–15.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of the 11th International Conference Recent Advances in Natural Language Processing*, pages 379–386.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. 2019. Joey NMT: A Minimalist NMT Toolkit for Novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 109–114.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-Based Curriculum Learning for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. 2019. Competence-based Curriculum Learning for Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1162–1172.

Sudha Rao and Joel Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 129–140.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language*, pages 30–39.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pages 5998–6008.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing Pre-Trained Neural Networks with Rules for Formality Style Transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3573–3578.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality Style Transfer with Shared Latent Space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An Empirical Exploration of Curriculum Learning for Neural Machine Translation. *arXiv:1811.00739*, pages 1–16.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel Data Augmentation for Formality Style Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228.