Sign Language Translation with Gloss Pair Encoding

Taro Miyazaki[1], **Sihan Tan**[1,2], **Tsubasa Uchida**[1], **Hiroyuki Kaneko**[1]
[1] NHK Science and Technology Research Laboratories
[2] Tokyo Institute of Technology
{miyazaki.t-jw, uchida.t-fi, kaneko.h-dk}@nhk.or.jp,
tan.s.ae@ra.sc.e.titech.ac.jp

Abstract

Because sign languages are the first language for those who are born deaf or who lost their hearing in early childhood, it is better to use sign languages rather than transcribed spoken language to provide important information to these people. We have been developing a sign language computer graphics generation system to provide information to deaf people, and in this paper, we present a translation method from spoken language to sign language that can be used in the system. In general, since the number of glosses used when transcribing sign language is limited, a single meaning is often expressed by a combination of multiple sign words, i.e., the word "library" is expressed in Japanese Sign Language with two words: "book" and "building." To merge these expressions into one token, we propose gloss pair encoding (GPE), which is inspired by bite pair encoding (BPE). This technique is expected to enable more accurate handling of expressions that have a single meaning in multiple sign words. We also show that it is effective as data augmentation on the sign language side in sign language translation, which has not been done much so far.

Keywords: Sign Language Translation, Machine Translation, Byte Pair Encoding, Gloss Pair Encoding

1. Introduction

Sign languages are typically the first language for those who are born deaf or who lose their hearing in early childhood. To provide information for these individuals, it is better to use sign language than to transcribe spoken language, as reading transcriptions of a spoken language, which is their second language, places an unnecessary burden on them.

We have been developing a sign language computer graphics (CG) generation system to provide information to deaf people. This system consists of two parts: a machine translation part and a CG generation part. The first part translates the input sentence of spoken language into a sign language gloss sequence, and the next part generates sign language CG based on the gloss sequence by referring to the motion data of each sign word. In this paper, we focus on improving the performance of the machine translation part.

In general, the number of glosses used for transcribing sign language is smaller than vocabulary size of spoken languages. For example, in our corpus, the gloss-based vocabulary size of Japanese Sign Language is approximately 4,000, while the word-based vocabulary size of Japanese is 27,000. Therefore, a single meaning is often expressed by a combination of multiple sign words, i.e., the word "library" is expressed in Japanese Sign Language with two words: "book" and "building." In this case, the glosses "book" and "building" play the role of subwords. Also, some glosses play the role of a letter, as in fingerspelling. In other words, glosses are sometimes used as a word, sometimes as a subword, and sometimes even as a letter. Since the granularity of glosses themselves can differ significantly, we believe that using them as they are in machine translation may cause degradation of the translation performance. Therefore, we propose a method to combine multiple glosses into one merged-gloss, and match the granularities.

The proposed method is named gloss pair encoding (GPE), which is inspired by byte pair encoding (BPE) (Sennrich et al., 2016). BPE is often used in machine translation to merge byte pairs that appear consecutively with high frequency into one merged subword. Our GPE is also merge gloss pairs that appear consecutively with high frequently into a merged-gloss. This allows multiple sign words that express one meaning to be treated as a single token. For example, the two glosses "book" and "building," which express the meaning of library, can be treated as a single merged-gloss "book+building." This is expected to enable more accurate handling of glosses with multiple meanings. Furthermore, through experiments, we demonstrate that it is also effective as data augmentation on the sign language side in sign language translation, which has not been done much so far.

Our contributions are summarized as follows. (1) We propose gloss pair encoding (GPE) to treat a gloss sequence that appears consecutively with high frequency as one token. (2) We show that by setting an appropriate number of vocabulary words, using merged-gloss with GPE can improve translation performance. (3) We experimentally demonstrate that by using a corpus with and without GPE in combination, translation performance can be improved due to the effect of data augmentation.

2. Related Work

2.1. Tokenization in Machine Translation

Machine translation utilizing neural networks originally treated words as the smallest unit (Cho et al., 2014; Bahdanau et al., 2015). Later on, in order to take advantage of the fact that each part of a word has something in common (e.g., the "person" part is common between personal and person, and there are also similarities in meaning) subwords have come to be used.

Byte pair encoding (BPE) is one of the most commonly utilized subword extracting methods. BPE was first proposed by Gage (1994) for encoding strings of text into tabular form, and Sennrich et al. (2016) then applied it to natural language processing methods including machine translation.

To avoid creating subwords that cross word boundaries, it is necessary to provide a separation for each word in advance. This is simple enough for languages that already have white spaces to separate words, such as English and German, but for other languages, such as Chinese and Japanese, it is necessary to separate words in advance. In response to this challenge, sentencepiece (Kudo and Richardson, 2018) was proposed as a method that allows end-to-end tokenization even in languages without word breaks.

2.2. Machine Translation of Sign Language

Several methods for translating spoken language into sign language have been proposed. Zhang and Duh (2021) regarded sign language translation as a low-resource machine translation task, and applied some of the techniques that are often used in low-resource language translation such as hyperparameter search and back translation. Zhu et al. (2023) applied techniques common to low-resource machine translation to sign language machine translation and showed that these techniques can also improve sign language translation. All of these methods use gloss sequence as sign language transcription.

The disadvantage of using gloss is that it causes important information in sign language, such as facial expressions and finger movement speed, to be lost. Therefore, gloss-free translation methods have recently proposed. Lin et al. (2023) proposed an end-to-end gloss-free translation method. Zhou et al. (2023) developed a novel pre-trained paradigm that combines masked self-supervised learning with visual language supervision learning, and they reported that this approach can deliver good translation. While gloss-free translation methods are currently used for translating sign language video into spoken language, there are few examples of its application for translating spoken language into sign language. This is because it is very challenging to generate motion data of sign language directly, which is necessary for gloss-free translation from spoken language to sign language,

The Conference on Machine Translation (WMT), a well-known workshop series of machine translation, initiated a shared task on sign language translation in 2022 (Müller et al., 2022). We hope this will lead to even more active research into sign language translation.

3. Proposed Method

As mentioned in the Introduction, the granularity of glosses can differs significantly, which is one of the reasons machine translation of sign language is difficult. Therefore, in our approach, we merge frequently occurring gloss sequences into one token by using gloss pair encoding (GPE), which is based on byte pair encoding (BPE) (Sennrich et al., 2016) and modified for sign language, to match the granularity of tokens.

First, we explain the original BPE, and next we present our proposed GPE.

3.1. Byte Pair Encoding (BPE)

BPE first initializes the vocabulary while covering all the characters in the training data, and regarding input data as sequences of characters with a special end-of-word symbol "·", which is utilized to restore the subword segmentation sentence to the original sentence. It then counts the frequencies of all symbol pairs and replaces the most frequently used pair ("A", "B") into one merged-character 'AB,' and adds it to the vocabulary. BPE applies this process repeatedly until finally it outputs the vocabulary including all the characters and merged-characters. The final vocabulary size can be controlled as a hyperparameter of the number times to repeat merge operations.

BPE makes it possible to achieve subword segmentation, where sequences of characters with meaning becomes a single vocabulary.

3.2. Gloss Pair Encoding (GPE)

In our GPE, the operation is almost the same as with BPE but differs in that is compresses frequent pairs of glosses instead of frequent pairs of bytes.

GPE first initializes the vocabulary while covering all the glosses in the training data. Unlike BPE, GPE does not use a special end-of-word symbol "·". It then counts the frequencies of gloss pair, and replace the most frequently used pair ("gloss$_A$", "gloss$_B$") into one merged-gloss "gloss$_A$+gloss$_B$," and adds it to the vocabulary. GPE applies this

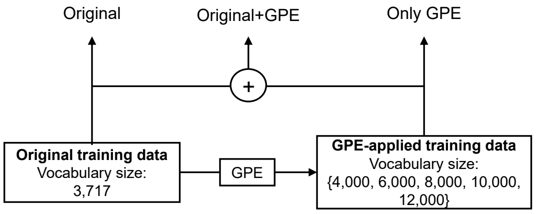

Figure 1: Three types of training data used in experiments.

process repeatedly until finally it outputs the vocabulary including all the glosses and merged-glosses The final merged-gloss vocabulary size can be controlled as a hyperparameter of the number times to repeat merge operations.

GPE merges gloss-pairs that appear continuously and frequently into a single merged-token. As a result, glosses that have the role of a subword are merged into merged-gloss, and we can expect the granularity of glosses to be ensured.

Also, by using GPE, the vocabulary size can be freely set using a hyperparameter, so the difference in vocabulary between spoken language and sign language can be reduced. This may also improve the translation performance.

3.3. Applying GPE to NMMs

Sign language utilizes non-manual movements (NMMs) such as head-nods and pointing. Note that although pointing is often treated as a sign word, we treated it as one of NNMs in this paper. Pointing does not express any meaning by itself, but it expresses meaning when combined with other words. Since this is the same as a head-nod, we decided to treat pointing as one of NMMs as a head-nod.

Head-nods often serve as function words, such as indicating the beginning or end of a sentence and expressing breaks in sentences or parallel relationships. Pointings is used to express meaning by referencing a previous word to emphasize the subject (Liddell, 2003).

In sign languages, NMMs are used much more frequently than sign words. Therefore, if NMMs are included in the merge target of GPE, it can be expected that many merged-glosses containing NMMs will be created. To evaluate this effect, we compare the performance when merging NMMs with GPE (with-NMM-GPE) and when not (without-NMM-GPE). In without-NMM-GPE, we first divide sentences by NMMs, then remove the NMMs, and finally apply GPE.

4. Experiments

We conducted an experiment to evaluate the text-to-gloss translation performance utilizing data with different sizes of vocabulary using GPE. In the experiment, we prepared one setting that did not apply GPE (Original), one that used only training data that applied GPE (OnlyGPE), and one that both applied and did not apply GPE are merged as training data (Original+GPE), as shown in Figure 1. In Original+GPE, training data with two different vocabularies are mixed and shuffled for learning. Since the vocabulary expanded by applying GPE always includes the same vocabulary as Original, it is possible to learn with a single encoder-decoder model without having to separate the translation models. We did not apply GPE to the development and test data, and the merged-gloss of the translation result was restored to the original gloss for evaluation.

In the experiment, we prepared training data with a total of five patterns of vocabulary size by applying GPE: 4,000, 6,000, 8,000, 10,000, and 12,000 for both with-NMM-GPE and without-NMM-GPE. As a baseline, we also conducted an experiment in which the vocabulary size of 3,717 that appeared three or more times in the corpus was used without GPE. The number of Japanese vocabularies was set to 8,000 in all experiments. We describe the experiments in detail below.

4.1. Our Corpus

We used an in-house corpus called the Japanese-JSL sign language news corpus for our experiment. This corpus is created from daily NHK sign language news programs, which are broadcast on NHK TV with Japanese narrations and JSL signings. The corpus includes around 160,000 Japanese transcriptions, JSL transcriptions, and JSL videos. Japanese is transcribed by revising the results of applying speech recognition on news programs. JSL is transcribed by native signers who manually transcribe each sign motion into sign language gloss. Note that, Japanese and JSL sentence pairs are not literal translations, so there are many subject complements, omissions, and so on. We transcribed all of the manual and some of the NMMs (e.g. head-nods and pointing) in linear transcription. In most cases, these type of manual and non-manual features are not expressed at the same time, so this transcription simplifies the JSL expressions while simultaneously retaining most of the necessary information.

We selected 129,950 sentence pairs that do not include *classifier*, which is hard to be transcribed into gloss. This is because *classifier* has a large vocabulary and no fixed hand or finger expressions, so our sign language CG generation system cannot convert them into sign language CG. We randomly split the corpus into 127,950 for training, 1,000 for development, and 1,000 for testing.

Data	No. of vocab.	without-NMM-GPE		with-NMM-GPE	
		Median	Average & std. deviation	Median	Average & std. deviation
Original	3,717	24.27	24.46 ± 0.40	–	–
Only GPE	4,000	24.75	24.73 ± 0.20	24.53	24.40 ± 0.47
	6,000	24.69	24.81 ± 0.14	21.75	21.74 ± 0.15
	8,000	24.09	24.05 ± 0.07	21.49	21.42 ± 0.17
	10,000	23.61	23.72 ± 0.35	20.83	20.72 ± 0.26
	12,000	23.45	23.46 ± 0.29	21.00	20.87 ± 0.23
Original + GPE	4,000	24.36	23.74 ± 0.60	24.37	24.37 ± 0.05
	6,000	**25.03**	25.03 ± 0.05	24.94	24.74 ± 0.47
	8,000	24.79	24.81 ± 0.09	24.75	24.66 ± 0.26
	10,000	25.02	**25.05** ± 0.09	24.64	24.59 ± 0.16
	12,000	24.61	24.79 ± 0.35	24.75	24.58 ± 0.51

Table 1: Experimental results. We show the median, average, and standard deviation of BLEU after three attempts with different random seeds. **Bold** indicates the best result in the table, and underline indicates a result that outperformed original.

4.2. Experimental Setting

We utilized a 6-layer transformer encoder-decoder model (Vaswani et al., 2017) with the *norm-first* setting (Nguyen and Salazar, 2019) for the translation model. We utilized PyTorch (Paszke et al., 2019) for implementing the model and RAdam (Liu et al., 2020) for optimization with the learning rate of 1.0×10^{-3}. We utilized cross-entropy loss for calculating loss in training. The dropout ratio for the transformer encoder and decoder was 0.1, and that for the output layer of the feed-forward neural network was 0.3. We applied sentencepiece (Kudo and Richardson, 2018) for input Japanese sentences with a vocabulary size of 8,000. We trained the models with the batch size of 256 and the number of training epoch of 50. We evaluated the model in each epoch using the development data, and chose the model with the best BLEU score on development set. We trained the models three times with different random seeds.

4.2.1. Results

Experimental results are provided in Table 1. As shown, Original+GPE performed better than Original in a wide range of vocabulary sizes for both without-NMM-GPE and with-NMM-GPE settings. In contrast, OnlyGPE performed better only in small vocabulary size, and its performance was worse than Original+GPE. In particular, OnlyGPE using the with-NMM-GPE setting underperformed the baseline in most cases. Overall, Original+GPE using the without-NMM-GPE setting performed the best.

4.3. Discussion

4.3.1. with- and without-NMM-GPE

The without-NMM-GPE setting outperformed the with-NMM-GPE setting for almost all vocabulary sizes. NMMs are very often used in sign language, so if GPE merges NMMs, most of the merged-glosses contain NMMs, and sign words are less often merged. Table 2 gives the percentage of merged-glosses that contains at least one NMMs, and Table 3 shows the top-5 frequently appearing merged-glosses in training data for the WITH and without-NMM-GPE settings. More than half of the merged-glosses in with-NMM-GPE contain NMMs, and many merged-glosses are combined with a head-nod representing the end of a sentence.

No. of vocab.	%
4,000	78.5
6,000	66.1
8,000	60.6
10,000	58.4
12,000	57.5

Table 2: Percentage of merged-gloss including NMMs in with-NMM-GPE setting.

Merged-gloss	Meanings
without-NMM-GPE	
Explanation + *Disappear*	I have given an explanation
Decide + *Disappear*	It has been decided
Place + *Place*	In various places
People + *Everyone*	Everyone
High + *Temperature*	Highest temperature
with-NMM-GPE	
pointing + head-nod	–
Exist + head-nod	Existing (EOS)
Disappear + head-nod	Finished (EOS)
head-nod + pointing	–
In + head-nod	Still (EOS)

Table 3: Top-5 merged-gloss of with-NMM-GPE and without-NMM-GPE. Glosses are in *italic*, NMMs are underbar, and merge is denoted by "+". (EOS) indicates that the head-nod in merged-gloss marks the end of a sentence.

Data	# vocab.	% of marged-gloss	
		Train	Output
Original	3,717	0	0
Only GPE	4,000	0.35	3.1
	6,000	12.91	11.65
	8,000	17.52	14.60
	10,000	20.00	15.82
	12,000	21.66	16.66
Original + GPE	4,000	0.18	0.01
	6,000	6.01	0.50
	8,000	7.94	0.55
	10,000	8.92	0.86
	12,000	9.56	0.88

Table 4: Percentage of merged-gloss among all gloss.

Since this combination does not extend the meaning, it is presumably not very effective. In contrast, many merged-glosses of without-NMM-GPE take on new meaning by merging multiple glosses.

Head-nods, which are a type of NMMs, often serve as function words and do not express meaning when combined with other words. Therefore, merging head-nods using GPE does not seem very effective. In contrast, pointing expresses meaning by combining with the previous word. However, our GPE often merged a pointing with the word following the pointing, which is meaningless since the pointing always points in the direction of the previous word. Also, as reported in our previous research, machine translation that merges a pointing and the previous word does not improve the performance (Miyazaki et al., 2020), thus, demonstrating that merging NMMs did not contribute to improving the translation quality.

We found that a better performance could be obtained by excluding NMMs from the merge target in GPE. In the following discussion, we examine the case of using without-NMM-GPE.

4.3.2. Effects of Merged-gloss

In the OnlyGPE setting, when the number of vocabularies was set appropriately (i.e., vocabulary size of 4,000 or 6,000 in this experiment) the performance improved. This indicates that with an appropriate vocabulary size, expressions that express one meaning by using multiple glosses can be combined into a merged-gloss, which makes it easier for translation models to learn and thereby improves the performance.

In contrast, OnlyGPE does not improve when the vocabulary size is increased to the same level as Japanese. This shows that it is not necessary to match the number of vocabularies in the source and target languages. The performance deterioration of OnlyGPE when the vocabulary size is large is presumably due to the fact that gloss-pairs that are not very frequent in the training data were also merged. The percentage of merged-glosses among all glosses for training data and translation output is shown in Table 4. With OnlyGPE, as the number of vocabularies increases, the percentage of merged-glosses in the output becomes considerably smaller compared to the training data. This suggests that GPE can create merged-glosses that are actually useful in translation only when the number of vocabulary words is around 6,000 in this dataset, and that for larger vocabulary sizes, it was mostly noise during learning. With OnlyGPE, merged glosses are not learned as a single gloss, so the influence of noise will be greater. On the other hand, with Original+GPE, merged glosses can be learned as merged-gloss as well as each single gloss by using original data, so the influence of noise can be reduced. This is why Original+GPE performed better than OnlyGPE especially for large vocabulary size in the experiments.

4.3.3. Effects of Data Augmentation

As shown in Table 4, the outputs of Original+GPE include not so many merged-glosses compared with the percentage of merged-glosses in the training data. This suggests that the increase in the amount of training data—that is, the effect of data augmentation—was large in Original+GPE and had a greater influence than the effects of the merged-glosses. Data augmentation in gloss-based sign language translation has been reported by (Zhu et al., 2023) and while they demonstrated data augmentation due to differences of the pre-processing on the spoken language side, data augmentation for sign language side was not examined. Our experiments indicate that data augmentation on the sign language side is also effective.

5. Conclusion

In this paper, we presented a translation method using gloss pair encoding (GPE), which merges multiple consecutive sign words that frequently appear in a corpus. We experimentally demonstrated that the translation performance improved when applying GPE with an appropriate number of vocabularies. We also found that by learning together with a corpus to which GPE is not applied, the effects of data augmentation can be obtained and translation performance can be further improved. When applying GPE it is better not to merge NMMs such as head-nod and pointing.

We did not perform an experiment in combination with data augmentation on the spoken language side. Many data augmentation methods for spoken language have been proposed, so considering how to combine them will be left as our future work.

6. Limitation

Also, this time we only used the training dataset of our in-house Japanese-Japanese Sign Language corpus. We hile we are confident that performance will improve regardless of the language pair, but we have not yet conducted experiments with other languages.

7. Bibliographical References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Scott K Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.

Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. Machine translation from spoken language to sign language using pre-trained language model as encoder. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural*

Information Processing Systems, volume 30. Curran Associates, Inc.

Xuan Zhang and Kevin Duh. 2021. Approaching sign language gloss translation as a low-resource machine translation task. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.