Controlling the Voice of a Sentence in Japanese-to-English Neural Machine Translation

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi

Tokyo Metropolitan University
{yamagishi-hayahide, kanouchi-shin, sato-takayuki} at ed.tmu.ac.jp,
komachi at tmu.ac.jp

Abstract

In machine translation, we must consider the difference in expression between languages. For example, the active/passive voice may change in Japanese-English translation. The same verb in Japanese may be translated into different voices at each translation because the voice of a generated sentence cannot be determined using only the information of the Japanese sentence. Machine translation systems should consider the information structure to improve the coherence of the output by using several topicalization techniques such as passivization.

Therefore, this paper reports on our attempt to control the voice of the sentence generated by an encoder-decoder model. To control the voice of the generated sentence, we added the voice information of the target sentence to the source sentence during the training. We then generated sentences with a specified voice by appending the voice information to the source sentence. We observed experimentally whether the voice could be controlled. The results showed that, we could control the voice of the generated sentence with 85.0% accuracy on average. In the evaluation of Japanese-English translation, we obtained a 0.73-point improvement in BLEU score by using gold voice labels.

1 Introduction

In a distant language pair such as Japanese-English, verbs between the source language and the target language are often used differently. In particular, the voices of the source and target sentences are sometimes different in a fluent translation when considering the discourse structure of the target side because Japanese is a pro-drop language and does not the use passive voice for object topicalization.

In Table 1, we show the number of occurrences of each voice in high-frequency verbs in Asian Scientific Paper Expert Corpus (ASPEC; Nakazawa et al. (2016b)). In the top seven high frequency verbs, "show" tended to be used in active voice, whereas "examine," "find," and "observe" tended to be used in the passive voice. However, "describe," "explain," and "introduce" tended not to be used in any particular voice. For example, the voice of the verb "introduce" could not be determined uniquely, because it was sometimes used in phrases like "This paper introduces ..." and, sometimes, "... are introduced." Therefore, it is possible that the translation model failed to learn the correspondence between Japanese and English.

Recently, recurrent neural networks (RNNs) such as encoder-decoder models have gained considerable attention in machine translation because of their ability to generate fluent sentences. However, compared to traditional statistical machine translation, it is not straightforward to interpret and control the output of the encoder-decoder models. Several attempts have been made to control the output of the encoder-decoder models. First, Kikuchi et al. (2016) proposed a new Long Short-Term Memory (LSTM) network to control the length of the sentence generated by an encoder-decoder model in a text summarization task. In their experiment, they controlled the sentence length while maintaining the performance compared to the results of previous works. Second, Sennrich et al. (2016) attempted to control the honorific in English-German neural machine translation (NMT). They trained an attentional encoder-decoder model using English (source) data to which the honorific information of a German (target) sentence was added. They restricted the honorific on the German side at the test phase.

Proceedings of the 3rd Workshop on Asian Translation,
pages 203–210, Osaka, Japan, December 11-17 2016.

Verb	# Active	# Passive		# Total
show	21,703	10,441	(32.5%)	32,144
describe	12,300	17,474	(58.7%)	29,774
explain	7,210	13,073	(64.5%)	20,283
introduce	6,030	9,167	(60.3%)	15,197
examine	3,795	11,100	(74.5%)	14,895
find	2,367	12,507	(84.1%)	14,874
observe	1,000	12,626	(92.7%)	13,626
All verbs	383,052	444,451	(53.7%)	827,503

Table 1: Number of occurrences of each voice in high-frequency verbs.

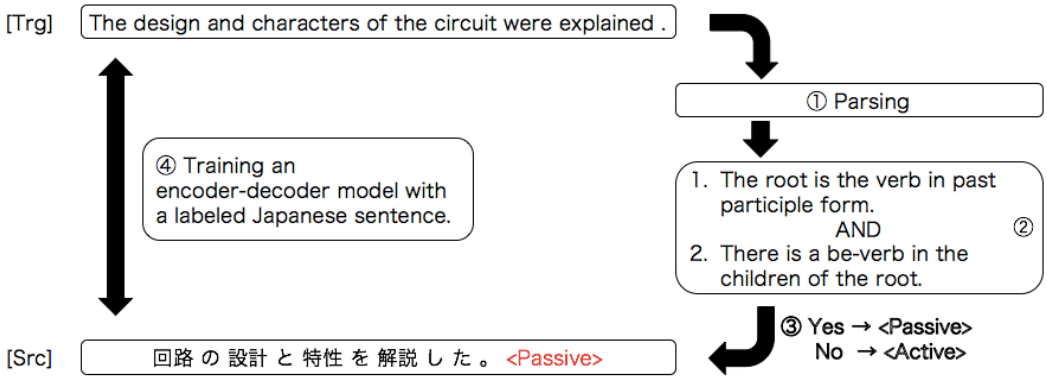

Figure 1: Flow of the automatic annotation for training an NMT.

Similar to Sennrich et al. (2016), this paper reports on our attempt to control the voice of a sentence generated by an encoder-decoder model. At the preprocessing phase, we determined the voice of the root phrase in the target side by parsing and added it to the end of the source sentence as a voice label. At the training phase, we trained an attentional encoder-decoder model by using the preprocessed source data. Lastly, we controlled the voice of the generated sentence by adding a voice label to the source sentence at the test phase. We tested several configurations: (1) controlling all sentences to active/passive voices, (2) controlling each sentence to the same voice as the reference sentence, and (3) predicting the voice using only the source sentence. The result showed that we were able to control the voice of the generated sentence with 85.0% accuracy on average. In the evaluation of the Japanese-English translation, we obtained a 0.73-point improvement in BLEU score compared to the NMT baseline, in the case of using the voice information of the references.

2 Controlling Voice in Neural Machine Translation

2.1 The Control Framework

In Japanese-English translation, the voices of the source and target sentences sometimes differ because the use of the verbs between the source and the target languages is different. In particular, English uses the passive voice to change the word order of a sentence for object topicalization to encode the information structure. Thus, it is beneficial to control the syntactic structure of the English sentences for discourse-aware machine translation. Moreover, if the voice of the generated sentence fluctuates at each sentence, it is difficult to train a translation model consistently.

In this paper, we attempt to add a ability of voice control to an encoder-decoder model, based on Sennrich et al. (2016), which controls the honorifics in English-German neural machine translation. They restricted the honorifics of the generated sentence by adding the honorific information to the source side. Instead of the honorific information, we extracted the voice information of the target sentence as a gold standard label to annotate the source sentence. At the test phase, we specified the voice of the generated

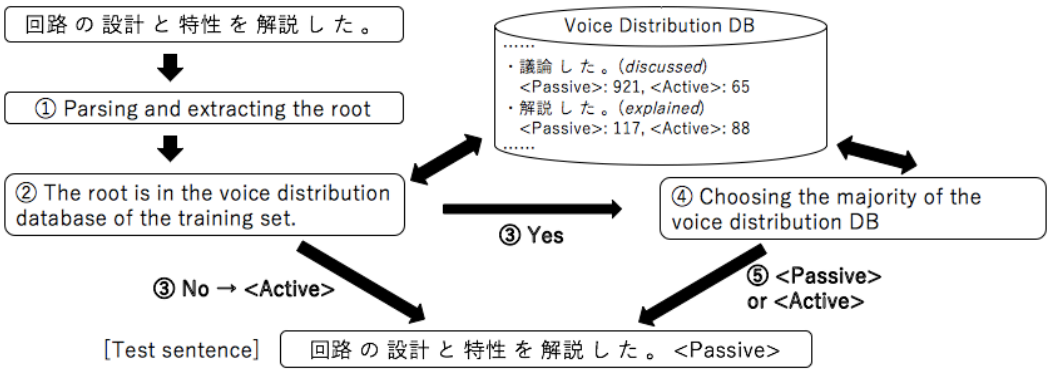

Figure 2: Flow of the voice prediction for testing an NMT.

sentence, and instructed the model to translate along with it.

In the following experiment, we used the attentional encoder-decoder model by Bahdanau et al. (2015). It is the same model that Sennrich et al. (2016) used. This model uses a bi-directional RNN as an encoder with attention structure. The proposed method can be adapted to any sequence-to-sequence model because it does not depend on the network structure.

2.2 Automatic Labeling of the Voice for Training

The performance of this method depends on the annotation performance of the voice at the training phase. Figure 1 shows the flow of the automatic annotation for training the attentional encoder-decoder model. We recognized the voice of the target (English) sentence by parsing. Then, the result of the parsing was checked to determine whether the root was a verb in the past participle form or not and whether it had a be-verb in the children or not. If both conditions were satisfied, the target sentence was recognized as being in the passive voice; otherwise, it was in the active voice[1]. For the voice controlling, we added a special token, <Active> or <Passive>, as a word to the end of the sentence, which became the input to the encoder. The special token, <Active> or <Passive>, encoded the voice of the root of the target sentence. The decoder considered only these tokens to determine the voice of the target sentence. For simplicity, we annotated only one voice for each sentence. In other words, if the sentence was a complex sentence, we selected the root verb for annotation. How the non-root verb must be treated in order to obtain the consistency of the document expression will be studied in a future work.

2.3 Voice Prediction for Testing

This study assumed that the voice label was determined in advance, but it was sometimes difficult to determine which label was suitable just from the source sentence alone. Even in this case, we had to add a voice label to the end of the source sentence to generate a target sentence because the proposed method necessarily uses a voice label.

Thus, we attempted to predict the voice for each sentence. Figure 2 shows the flow of the voice prediction. We investigated the voice distribution of the English verb in each root phrase of the Japanese side in the training data to predict the voice of the generated sentence.

At the test phase, we also obtained the root phrase of the Japanese sentence. If the root phrase was included in the training data, we added the majority label of the voice distribution in the training data as a predicted label. If the root phrase was not in the training data, the voice label was <Active>.

3 Experiments

We conducted two types of evaluations: evaluation of the controlling accuracy and evaluation of the machine translation quality. We tested the following four patterns of labeling the voice features to evaluate

[1]Strictly speaking, we checked whether the target sentence was in the passive voice or not, but we did not distinguish "not in passive voice" from "active voice."

	# Active	# Passive	# Error	Accuracy	BLEU
Reference	100	100	0	—	—
Baseline (No labels)	74	117	9	(72.0)	20.53
ALL_ACTIVE	151	36	13	75.5	19.93
ALL_PASSIVE	17	175	8	87.5	19.63
REFERENCE	97	94	9	89.5	**21.26**
PREDICT	72	121	7	87.5	20.42

Table 2: Accuracy of voice controlling and BLEU score of the translation.

the extent to which the voice of the generated sentence was controlled correctly.

ALL_ACTIVE. Controlling all target sentences to the active voice.

ALL_PASSIVE. Controlling all target sentences to the passive voice.

REFERENCE. Controlling each target sentence to the same voice as that of the reference sentence.

PREDICT. Controlling each target sentence to the predicted voice.

There were two reasons for testing ALL_ACTIVE and ALL_PASSIVE: to evaluate how correctly we counld control the voice, and to discuss the source of errors. In REFERENCE, the generated sentences tended to be natural. However, in ALL_ACTIVE and ALL_PASSIVE, the generated sentences were sometimes unnatural in terms of the voice. We identified these sentences to investigate the reasons why these errors occurred.

We checked the voice of the generated sentence and calculated the accuracy manually because the performance of voice labeling depends on the performance of the parser. We used the Stanford Parser (ver. 3.5.2) to parse the English sentence. The labelling performance was 95% in this experiment. We used CaboCha (ver. 0.68; Kudo and Matsumoto (2002)) to obtain the root phrase of the Japanese sentence in PREDICT. If the sentence was a complex sentence, we checked the voice of the root verb[2].

The test data of ASPEC consisted of 1,812 sentences in total. The evaluation data for the voice controlling consisted of 100 passive sentences and 100 active sentences chosen from the top of the test data. We did not consider subject and object alternation because this evaluation only focused on the voice of the sentence. Only one evaluator performed an annotation. In this experiment, the accuracy was calculated as the agreement between the label and the voice of the generated sentence. "Error sentence" means the root verb of the generated sentence could not be distinguished manually, or it did not include a verb, and so on. The baseline was an attentional encoder-decoder by Bahdanau et al. (2015), which does not control the voice. In the evaluation of the Japanese-English translation, we calculated the BLEU (Papineni et al., 2002) score with the test data of all 1,812 sentences.

At the training phase, we used 827,503 sentences, obtained by eliminating sentences with more than 40 words in the first 1 million sentences of the ASPEC. Word2Vec[3] (Mikolov et al., 2013) was trained with all 3 million sentences of ASPEC. The vocabulary size was 30,000[4]. The dimension of the embeddings and hidden units was 512. The batch size was 128. The optimizer was Adagrad, and the learning rate was 0.01. We used Chainer 1.12 (Tokui et al., 2015) to implementing the neural network.

4 Result and Discussion

4.1 Experiments with a Gold Voice Label

Table 2 shows the accuracy of the voice control and the BLEU score of the translation[5]. In the baseline, our system tended to generate a passive sentence compared to the voice distribution of the reference

[2]Even if the root phrase of the Japanese sentence was semantically different from the root of the English sentence, we still checked the voice of the root of the English sentence without considering the meanings.

[3]https://radimrehurek.com/gensim/models/word2vec.html

[4]We did not perform any processing of unknown words because we focused on the control of the voice.

[5]In this experiment, the BLEU score was calculated before the detokenization because we focused on the voice controlling. We submitted our system for the crowdsourcing evaluation after the detokenization.

because the number of passive sentences was greater than that of the active sentences in the training data. The accuracy of the baseline was calculated as the agreement between the voice of the generated sentence and that of the reference.

ALL_ACTIVE and ALL_PASSIVE demonstrated that the voice could be controlled with high performance. The BLEU score became lower than the baseline because some sentences were transformed into different voices regardless of the contexts and voice distribution. In other words, active sentences in the test data included sentences whose root verb of the reference was an intransitive verb. Even in that case, we forced the voice of the generated sentence to become passive in ALL_PASSIVE. As a result, the voice of some sentences did not become passive, compared to other sentences that were controlled to become passive sentences if not natural.

REFERENCE achieved the highest accuracy, and its voice distribution was close to that of the references. As mentioned earlier, the voice of REFERENCE was more natural than that of ALL_ACTIVE or ALL_PASSIVE. We obtained a 0.73-point improvement in the BLEU score compared to the baseline[6]. Therefore, we found that there is room for improvement if we can correctly predict the voice of the reference.

PREDICT used the labels predicted from the voice distribution. It tended to generate a passive sentence compared to the baseline. The controlling accuracy was 87.5% because the voice distributions were skewed in many verbs. However, the agreement rate between the predicted and the reference voices was 63.7%. Therefore, PREDICT failed to predict the voice of the reference, especially with high-frequency verbs, resulting in decrease in the BLEU score. We leave the prediction of the voice of references as a future work.

We show the output examples in Table 3. Examples 1, 2, and 3 are the success cases, whereas Examples 4 and 5 are the failure cases.

Examples 1 and 2 showed that the voice of the generated sentence was correctly controlled. When a passive sentence was changed into an active sentence, a subject was needed. Both examples generated adequate subjects depending on the context. In Example 3, although the voice was controlled, the subject and object were not exchanged. Besides this example, there were many sentences that persisted the "be-verb + verb in past participle form" structure when adding the <Passive> label was added. For example, the "... can be done ..." structure was changed into the "... is able to be done ..." structure. In this experiment, we did not evaluate whether the subject and object were exchanged, but it may be necessary to distinguish these patterns for the purpose of improving the coherence of the discourse structure.

In Example 4, it was impossible to make a passive sentence because the root verb in the target sentence should be an intransitive verb. Most of the active sentences in ALL_PASSIVE should stay active sentences that used intransitive verbs. Like Example 3, there were many sentences that were successfully controlled by using the "be found to be ..." structure when an intransitive verb was included as a root verb. Example 5 showed the case wherein the voice could not be controlled despite the attempt to control it to the active voice. The frequency of the voice of the verb "detect" in the training data consisted of 468 active-voice sentences and 2,858 passive sentences. When we forced the voice of the generated sentence to become active, the result of generation tended to fail sometimes if we input the verb that had few examples of active sentences in the training data. The subject should be generated if we forced the voice of the generated sentence to become active. However, the encoder-decoder model did not know what to generate as a subject if the training data had only a few examples of an active sentence for that verb. On the other hand, when we forced the voice of the generated sentence to become passive, we failed to find any tendencies of this type of the failure. We would like to do some additional investigation on the tendency of this result.

4.2 Experiments with Predicated Voice: TMU at WAT 2016

Table 4 shows the results of two methods submitted for the shared task at WAT 2016 (Nakazawa et al., 2016a). The BLEU, RIBES (Isozaki et al., 2010), and AMFM (Banchs et al., 2015) were calculated

[6]We were not able to submit REFERENCE for the human evaluation because we were not allowed to look at the references in WAT 2016.

Example 1	Source	熱戻り反応の機構を議論した。
	Reference	This paper discusses the mechanism of the heat return reaction.
	Controlling to Active	We discuss the mechanism of the thermal return reaction.
	Controlling to Passive	The mechanism of the thermal return reaction is discussed.
Example 2	Source	リサイクルに関する最近の話題を紹介した。
	Reference	Recent topics on recycling are introduced.
	Controlling to Active	This paper introduces recent topics on recycling.
	Controlling to Passive	Recent topics on recycling are introduced.
Example 3	Source	自己組織化構造に分子の形と分子間相互作用が大きく影響する。
	Reference	Molecular shape and intermolecular interaction influence self-assembled structures greatly.
	Controlling to Active	The molecular structure and molecular interaction greatly affect the self-organization structure.
	Controlling to Passive	The molecular structure and molecular interaction are greatly affected by the self-organization structure.
Example 4	Source	テロメラーゼ活性は生殖細胞と癌細胞で高い。
	Reference	Telomerase activity is high in reproductive cells and cancer cells.
	Controlling to Active	The telomerase activity is high in the reproductive cell and cancer cells.
	Controlling to Passive	The telomerase activity is high in the reproductive cell and cancer cells.
Example 5	Source	その結果, thz 波は stj でのトンネリング電流信号として検出できる。
	Reference	Consequently, the thz waves can be detected as tunneling current signals at stj.
	Controlling to Active	As a result, the thz wave can be detected as a current current signal in the <unk>.
	Controlling to Passive	As a result, the thz wave can be detected as a current current signal in the <unk>.

Table 3: Examples of the generated sentences

System	BLEU	RIBES	AMFM	HUMAN
NMT Baseline	16.89	0.700849	0.546038	—
6 ensemble	**18.45**	**0.711452**	0.546880	**+25.000**
PREDICT	18.29	0.710613	**0.565270**	+16.000

Table 4: Evaluation scores of WAT 2016.

automatically, and HUMAN was evaluated by the pairwise crowdsourcing. Note that the NMT baseline is different from the baseline of the voice controlling experiment reported in the previous section.

6 ensemble: We performed an ensemble learning of the NMT baseline. Because of the lack of time, we trained the baseline NMT only twice. Thus, we chose three models that showed the three highest BLEU scores from all epochs of the development set for each NMT baseline, resulting in 6 ensemble. As a result, BLEU score achieves 18.45. It improves 1.56 point compared with the result of the single NMT Baseline.

PREDICT (2016 our proposed method to control output voice): We submitted our system in the configuration of PREDICT for pairwise crowdsourcing evaluation. It improved by 1.40 points in the BLEU score compared to the NMT baseline. Since we did not perform an ensemble learning for PREDICT, we expected a similar improvement in the BLEU score if we combined multiple models of PREDICT using an ensemble technique.

5 Related Work

An NMT framework consists of two recurrent neural networks (RNNs), called the RNN encoder-decoder, proposed by Cho et al. (2014) and Sutskever et al. (2014). The accuracy of NMT improves by using the attention structure (Bahdanau et al., 2015; Luong et al., 2015). However, the optimization of an RNN using log-likelihood does not always yield a satisfactory performance depending on the tasks at hand. For example, one may prefer a polite expression for generating conversation in a dialog system. Thus, several methods have been proposed several methods to control the output of encoder-decoder models.

First, Kikuchi et al. (2016) tried to control the length of the sentence generated by an encoder-decoder model in a text summarization task. They proposed four methods for restricting the length in the text summarization task and compared them. In their result, they obtained a learning-based decoder for

controlling the sentence length without compromising on the quality of the generated sentence.

Second, Sennrich et al. (2016) tried to control the honorifics in the task of English-German NMT. They trained an attentional encoder-decoder model by modifying the English data to include the honorific information of the German side. The result showed that the accuracy of enforcing the honorifics to the sentence was 86%, and that of constraining the sentence to not have the honorifics was 92%. They obtained an improvement of 3.2 points in the BLEU score when the sentence was limited to the gold honorifics as the reference sentence.

6 Conclusion

This paper reported on our attempt to control the voice of the sentence generated by in an encoder-decoder model. At the preprocess phase, we determined the voice of the root verb of the target language by parsing, and added a voice label to the end of the source sentence as a special token. At the training phase, we trained an attentional encoder-decoder model by using a preprocessed parallel corpus. At the test phase, we restricted the target sentence to have a particular voice by specifying a voice label in the encoder. The result showed that we were able to control the voice of the generated sentence with 85.0% accuracy on average. In the evaluation of the Japanese-English translation, we obtained a 0.73-point improvement in the BLEU score by using gold voice labels compared to the baseline.

Our future work includes making a supervised classifier for predicting the voice, controlling another stylistic expression, and implementing the control function into the network structure such as a gate in an LSTM.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rafael E. Banchs, Luis F. D'Haro, and Hizhou Li. 2015. Adequacy-fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1328–1338.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2016a. Overview of the 3rd workshop on Asian translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, Osaka, Japan, October.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016b. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208.

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 on Association for Computational Linguistics (ACL)*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 35–40.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in the 2015 Conference on Neural Information Processing Systems (NIPS)*.