Improving Japanese-to-English Neural Machine Translation by Voice Prediction

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

This study reports an attempt to predict the voice of reference using the information from the input sentences or previous input/output sentences. Our previous study presented a voice controlling method to generate sentences for neural machine translation, wherein it was demonstrated that the BLEU score improved when the voice of generated sentence was controlled relative to that of the reference. However, it is impractical to use the reference information because we cannot discern the voice of the correct translation in advance. Thus, this study presents a voice prediction method for generated sentences for neural machine translation. While evaluating on Japanese-to-English translation, we obtain a 0.70-improvement in the BLEU using the predicted voice.

1 Introduction

Recently, recurrent neural networks such as encoder-decoder models have gained increasing attention in machine translation owing their ability to generate fluent sentences. Controlling the output of the encoder-decoder model is difficult; however, several control mechanisms have been developed. For example, Sennrich et al. (2016) attempted to control honorifics in English-German neural machine translation (NMT). They trained an attentional encoder-decoder model (Bahdanau et al., 2015) using source data wherein the honorific information of a target sentence was represented by an additional word. They obtained a 3.2-point improvement in the BLEU score when the sentence was controlled to the same honorifics as the reference.

Similar to the research of Sennrich et al. (2016), Yamagishi et al. (2016) reported an attempt to control the voice of a generated sentence using an attentional encoder-decoder model. They added a label to the end of the source sentence using the voice information of the target sentence during training. Subsequently, they translated the source sentences with a specified voice by appending the voice information. As a result, 0.73-point improvement in BLEU was achieved if the reference information was used.

Although Yamagishi et al. (2016) showed the upper bound for the improvement, it is impractical to use the reference information in the test phase. Therefore, in this study, we develop a voice classifier using a logistic regression model with simple context features. Note that our previous experiments did not exclude intransitive verb from the training and testing process, which may result in over-estimation of the active voice. Thus, for fair comparison, our test data constructed in this paper only contain transitive verbs. Our results demonstrate 67.7% and 66.0% voice prediction accuracies for the target sentence translated from Japanese to English on Asian Scientific Paper Excerpt Corpus (ASPEC, Nakazawa et al. (2016)) and NTCIR PatentMT Parallel Corpus (NTCIR, Goto et al. (2013)), respectively. An evaluation of the translation shows the statistically significant improvements in the BLEU score when using the predicted voice. In addition, a manual inspection shows that the voice-controlled translation clearly produces more fluent translation than the baseline.

2 Voice Prediction

Our previous study (Yamagishi et al., 2016) did not build a voice classifier for voice control; we used the majority of voice for each verb in the training corpus. We reported that the majority vote did not consistently improve the BLEU score. In

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 277–282,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

contrast, this study develops a voice classifier using the following seven features. In this way, we can consider the context information of the source and target languages when predicting the voice of generated sentences. Note that we expect not only the quality of the voice prediction but also the quality of the translation to improve. These features are concatenated as a vector used to train a logistic regression model.

SrcSubj: Phrase embedding of the subject in a source sentence[1].

SrcPred: Phrase embedding of the predicate in a source sentence.

SrcPrevPred: Phrase embedding of the predicate in the previous source sentence.

SrcVoice: Voice of the source sentence.

TrgPrevObj: Word embedding of the objects in the previous target sentence.

TrgPrevVoice: Voice of the output sentences from previous three sentences.

TrgVoicePrior: The majority of target voice of each predicate phrase in a source sentence.

The phrase embeddings are the average of the all the word embeddings, except for alphabets, numerals, and punctuation marks in the phrase. All features are calculated from the information obtained from the main clauses. "Previous sentence" is flagged when an input sentence is not the first sentence in a document. We use SrcPrevPred, TrgPrevObj, and TrgPrevVoice to consider the information structure of a document.

TrgPrevVoice and TrgVoicePrior accept only three values, i.e., "Active," "Passive," and "No information." TrgVoicePrior represents the relation between the predicate of a source sentence and the voice of a target sentence. If the predicate of a test sentence is included in the training data, we obtain the majority of the voice distribution each predicate phrase in the training data. It can be noted that only the value of TrgVoicePrior was directly used as a label by Yamagishi et al. (2016); TrgVoicePrior was not used as a feature in the logistic regression model.

SrcVoice represents the voice of the source side. However, unlike English, it is difficult to formulate simple rules to obtain the voice of Japanese sentences[2]. Thus, this feature shows whether the sentence has an auxiliary verb of representing passive voice.

3 Experiments

3.1 The Control Framework

Here we explain the voice controlling method proposed by Yamagishi et al. (2016) for Japanese-to-English NMT. We parse the target sentence and then evaluate the result to determine whether the ROOT is a past participle and whether it has a be-verb in the children. If both the conditions are satisfied, the target sentence is considered "passive." Otherwise, it is considered "active." The voice information is added to the end of the source sentence as a word. Finally, we create a new training corpus using labeled sentences. In the test phase, an `<Active>` or `<Passive>` label is added to the end of the source sentences to generate sentences in the desired voice.

3.2 Settings

We experimented with four labeling patterns.

ALL_ACTIVE: All sentences to active voice.

ALL_PASSIVE: All sentences to passive voice.

REFERENCE: Each sentence to the same voice as that of the reference sentence.

PREDICT: Each sentence to the predicted voice.

We mainly use the ASPEC (Nakazawa et al., 2016) in this experiment. The ASPEC comprises abstracts from scientific papers. We reconstructed the ASPEC as a document-level bilingual corpus. Sentences with more than 50 words from the training data are deleted, and the parallel documents that comprise continuous sentences are collected. As a result, the number of sentences in the training data is 1,103,336 (329,025 documents; the average number of sentences per document is 3.35). The original test data comprises 453 documents (four sentences in each document). Thus, it has 1,812 sentences in total. To evaluate voice control accuracy, we select 100 active sentences and 100 passive sentences from the top of the original test data. As stated in Section 1, sentence pairs whose ROOT of the reference is an intransitive verb are

[1] We extract the NP with the nominative case particle.

[2] In Japanese, the auxiliary verbs "れる (*reru*)" or "られる (*rareru*)" are typically used in the passive voice. However, they are also used to represent possibility or honorifics. It is difficult to apply simple rule to distinguish their usage.

omitted from the test data because it may not be possible to generate the passive sentences.

We also use the NTCIR Corpus (Goto et al., 2013) to investigate the corpus-specific tendencies. As a result of the preprocessing used with the NTCIR, the number of sentences in this training data is 1,169,201. The NTCIR10 development data and test data are used, which include 2,741 and 2,300 sentences, respectively. Note that we could not reconstruct this corpus at a document-level one. Therefore, our voice classifier only uses the sentence-level features in the experiments of voice prediction experiments.

The experimental results are based on accuracy, BLEU scores (Papineni et al., 2002), and human evaluation. Two types of accuracy were considered, i.e., voice accuracy and control accuracy. Voice accuracy is calculated as the agreement between the voice of the reference and that of the generated sentence. Control accuracy is calculated as the agreement between the label and the voice of the generated sentence. Note that only one evaluator performs annotation. We do not consider subject and object alternation because this evaluation only focuses on the voice of the sentence. We show two BLEU scores, i.e., BLEUall and BLEU200. BLEUall represents the score evaluated using all official test data, and BLEU200 represent the score evaluated using arranged test data described earlier. We statistically evaluate the BLEU scores using the bootstrap resampling implemented in Travatar[3]. The human evaluation involves pairwise comparison between the baseline results and the REFERENCE results (Base:REF) or between the baseline results and the PREDICT results (Base:PRED). The evaluator of this comparison is only one. Note that the evaluator differs from the voice label annotator.

We use CaboCha[4] (Ver. 0.68) to parse the Japanese sentences, and the Stanford Parser[5] (Ver. 3.5.2) to parse the English sentences. Scikit-learn (Ver. 0.18) is used to implement logistic regression. The word embeddings[6] (Mikolov et al., 2013) that we use as the features for voice prediction are trained using the source side of training corpus of ASPEC with 100 dimensions. The voice-labeling performance is 95%.

We obtain two NMT models, one trained using the original corpus and the other trained using the labeled corpus. The former model is the baseline. These models are optimized by Adagrad (learning rate: 0.01). The vocabulary size is 30,000, the dimensions of the embeddings and hidden units are 512, and the batch size during training is 64. We train both the models for 15 epochs. We use Chainer (Ver. 1.18; Tokui et al. (2015)) to implement NMT models proposed by Bahdanau et al. (2015). We train Word2Vec with all 3,008,500 sentences in the ASPEC original training data for initializing the word vectors. Likewise, we use the source side of the training corpus of NTCIR to train Word2Vec for the experiments of NTCIR.

4 Results and Discussion

4.1 Voice Classifier Result

Table 1 summarizes the results of label prediction and an ablation test for feature selection. Yamagishi et al. (2016) reported that the accuracy of the majority voice was 63.7% on the ASPEC. Therefore, we obtained slight improvements in those scores by using the regression model with several features.

First, we discuss the result using ASPEC. The model using SrcPred, TrgPrevVoice, and TrgVoicePrior obtains the highest accuracy. The most important feature is SrcPred. The words included in the predicate phrase have some tendencies in each voice. TrgVoicePrior comprises the majority of the information in the training data. It is possible that this feature is inaccurate for verbs having no voice skewness tendency. TrgPrevVoice is also a useful feature to predict the voice, except for the first sentence in each document. SrcVoice is not a useful feature because the voice of the source sentence is not always the same as that of the target sentence. The voice concordance rate between languages is 53.5% on ASPEC.

Accuracy decreases using the other features. SrcSubj and TrgPrevObj are useless because many source sentences do not contain any subject and many target sentences do not contain any object. SrcPrevPred is ineffective because the voice seems to be determined by the discourse structure of the target sentences. We consider that the voice of output sentence is influenced by the words included in the previous outputs. However, our classifier only requires the voice information for the previous outputs.

Second, we discuss the result obtained using the

[3] http://www.phontron.com/travatar/index.html

[4] https://taku910.github.io/cabocha/

[5] http://nlp.stanford.edu/software/lex-parser.shtml

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

Feature \ Corpus	ASPEC										NTCIR				
SrcSubj	✓		✓	✓	✓	✓	✓	✓			✓		✓	✓	✓
SrcPred	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
SrcPrevPred	✓	✓	✓		✓	✓	✓	✓			—	—	—	—	—
SrcVoice	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓		✓
TrgPrevObj	✓	✓	✓	✓	✓		✓	✓			—	—	—	—	—
TrgPrevVoice	✓	✓	✓	✓	✓	✓		✓	✓	✓	—	—	—	—	—
TrgVoicePrior	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	
Accuracy (%)	67.2	67.3	65.2	67.2	66.9	67.3	65.9	65.4	67.5	**67.7**	65.7	65.9	65.9	65.8	**66.0**

Table 1: Results of label prediction and ablation test for feature selection. "✓" represents "used", and "—" represents "cannot be used" in each column.

NTCIR corpus. Herein, the highest accuracy is 66.0%, which is obtained by the model using three features, i.e., SrcSubj, SrcPred, and SrcVoice. SrcVoice is the best predictive feature because the voice concordance rate is 63.3% on NTCIR. When we examine the top 50 most frequent predicates in both corpora, auxiliary verbs that represent passiveness are found in five predicates in the ASPEC, while auxiliary verbs are found in 17 predicates in the NTCIR. If a source sentence includes those 17 predicates, the generated sentence tends to be a passive sentence. It is not clear why the effective features for voice prediction differ in these corpora because the percentages of sentences that do not have the subject are quite similar.

4.2 Translation Result

Table 2 shows the voice controlling results. "Other" indicates that the generated sentence is unreadable and that it does not include a verb.

Table 2(a) shows the result using ASPEC. The baseline model tends to generate a passive sentence, although the number of active sentences is greater than that of the passive sentences in the training data. This occurs because the generated sentence using a transitive verb tends to be a passive sentence under the general condition due to the fact that active sentences in the training data may contain an intransitive verb. We perform Base:REF comparison in which a human evaluator assessed that REFERENCE is better than the baseline model. We can obtain further improvement if we can appropriately change the voice of the generated sentence to that of the reference. With PREDICT, we obtain a 0.70-point improvement in the BLEUall score. The score of Base:PRED is close to that of Base:REF. Although we do not use the reference information in PREDICT, we obtain a promising result in human evaluations using the proposed method.

We observe the same tendencies when using the NTCIR, as shown in Table 2(b). The improvement to the BLEUall score between Baseline and REFERENCE is less than that in the ASPEC experiment. If an NMT model tends to generate sentences in a particular voice, the voice control method fixes this tendency. We can observe this tendency in the voice accuracy of Baseline; however, this is not observable in the training data. Thus, the voice control method becomes more effective when voice accuracy is low.

4.3 Discussion of Translation

Table 2 shows that it was difficult to generate active sentences. It is difficult for the model to generate an appropriate subject when a sentence is forced to become an active sentence despite it should be a passive sentence. The model tends to generate appropriate subjects only if a high-frequency verb is included in the generated sentence. In the NTCIR experiment, the model tends to generate the passive sentences, even though it is forced to produce active sentences when the source sentence has an auxiliary verb which represents passiveness. This tendency is not observed with the ASPEC because this corpus includes fewer sentences with such auxiliary verbs than the NTCIR. The reasons why ALL_PASSIVE obtains high accuracy is that these problems do not occur when generating the passive sentences.

Table 3 summarizes the examples of the generated sentences on the experiment using ASPEC. Example 1 shows that the voice of the generated sentence was appropriately controlled in the case of a single sentence. The voice controlling method only annotates voice information for the main clause; however, some input sentences are complex sentences. Examples 2 and 3 show the results of the subordinate clause and coordinate clause, respectively. The voice of the main clause is different from that of the subordinate clause at "to be Passive" in Example 2, although the voice of the main clause is the same as that of the coordinate clause in Example 3.

Experiment	# Active	# Passive	# Other	Voice acc.	Control acc.	BLEU200	BLEUall	Base:REF	Base:PRED
Reference	100	100	—	—	—	—	—	—	—
Baseline	31	163	6	60.5%	—	20.60	17.16	80	76
ALL_ACTIVE	147	44	9	57.5%	73.5%	20.22	—	—	—
ALL_PASSIVE	6	189	5	51.0%	94.5%	20.18	—	—	—
REFERENCE	82	113	5	89.0%		****22.47**	****18.78**	120	—
PREDICT	74	118	8	64.0%	89.0%	21.05	*17.86	—	124

(a) Experiments using ASPEC.

Experiment	# Active	# Passive	# Other	Voice acc.	Control acc.	BLEU200	BLEUall
Reference	100	100	—	—	—	—	—
Baseline	69	127	3	66.0%	—	31.80	29.29
ALL_ACTIVE	127	69	4	71.0%	63.5%	31.91	—
ALL_PASSIVE	17	186	3	55.0%	93.0%	32.32	—
REFERENCE	80	116	4	89.5%		****33.90**	***29.80**
PREDICT	73	122	5	69.5%	83.0%	33.16	29.59

(b) Experiments using NTCIR.

Table 2: Performance of voice control, BLEU score, and the result of the human evaluations in each corpus. These scores are calculated by original test data except for BLEUall. * represents the p-value < 0.05, and ** represents the p-value < 0.01 over the baseline.

Example 1	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 .
	To be Active	this paper introduces recent topics on the recycling .
	To be Passive	recent topics on the recycling are presented .
Example 2	Source	また，ドットの形状及び結晶性は温度に依存することも分かった．
	Reference	it was also proven that the shape and crystallinity of the dots were dependent on temperatures .
	To be Active	the morphology and the crystallinity of the dots depended on the temperature .
	To be Passive	it was also found that the shape and the crystallinity of the dots depend on the temperatures .
Example 3	Source	超電導材料開発のためのデータベースを構築し、材料設計用演えきシステムの開発を行った。
	Reference	a database for development of superconducting material was constructed , and deduction system for material design was developed .
	To be Active	we constructed a database for the development of superconducting materials and developed a deduction system for material design .
	To be Passive	a database for the development of superconducting materials was constructed , and the <unk> system for material design was developed .

Table 3: Examples of the generated sentences on the experiment using ASPEC.

Clause type		ALL_ACTIVE	ALL_PASSIVE
Coordinate	# Active	22	8
	# Passive	19	40
	# Total	41	48
	Agreement	63.4%	77.1%
Subordinate	# Active	29	21
	# Passive	15	13
	# Total	44	34
	Agreement	55.5%	38.2%

Table 4: The number of coordinate or subordinate clause in each voice on ASPEC.

Table 4 summarizes the results of the coordinate and subordinate clauses on the experiment of ASPEC. "Agreement" in this table represents the concordance rate between the voice of the main clause and that of each clause. If this rate is high, the voice of all clauses in a sentence is controlled to the same voice as the added label. The voices of the dependent clauses are not controlled, although the voice of the main clause can be controlled at high accuracy. As mentioned previously, the translation model tends to generate a passive sentence when it is expected to generate a transitive verb. We recognize the same tendency in the generation of the voice of the coordinate clause. Conversely, the voice of the subordinate clause tends to be active because the "be-verb + adjective" structure or "be-verb + noun" structure tends to be used in the subordinate clause, as in the abovementioned example. Hence, the proposed method greatly influences the main clause to which the voice information is added, although it also affects the dependent clauses.

5 Conclusion

This paper reported an attempt to predict the voice of reference sentence to improve the translation quality using the voice controlling method. We used simple features to train the logistic regression model. As a result, we predicted the voice of the reference sentences at 67.7% accuracy on ASPEC and at 66.0% on NTCIR, respectively. We observed difference of important features between the corpora. Certain improvement in BLEU score was achieved using the voice classifier results to output generated sentences in Japanese-to-English NMT. We will attempt to improve the quality of machine translation using context information of a document, including the voice information.

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)*.

Isao Goto, Bin Lu, and Benjamin K Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NII Testbeds and Community for Information access Research Conference (NTCIR)*.

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, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. 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.

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)*.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.