A Study of Punctuation Handling for Speech-to-speech Translation

Xiaolin Wang Andrew Finch Masao Utiyama Eiichiro Sumita
Advanced Speech Translation Research and Development Promotion Center
National Institute of Information and Communications Technology, Japan
{xiaolin.wang, andrew.finch, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

This paper is devoted to finding out proper methods of handling punctuation when translating un-punctuated text output from automatic speech recognition systems. Three methods of handling punctuation and two machine translation (MT) systems were studied on a Japanese-to-English parallel corpus of about 5 million sentence pairs. BLEU calculated without punctuation was employed as measurement of translation quality in order to reflect the quality of synthesized speech output. Experimental results show that, for both phrase-based MT systems and pre-ordering phrase-based MT systems, methods of predicting punctuation with either a hidden n-gram model or a monolingual machine translation systems yield translation systems which are close in quality to the one using oracle punctuation, and better than ignoring punctuation.

1 Introduction

The task of Spoken Language Translation (SLT) is to automatically translate text output by an Automatic Speech Recognition (ASR) system, into another language. SLT is an important application of machine translation (MT) because it takes one of the most natural forms of human communication – speech – as input and makes it accessible to speakers of another language [Peitz et al., 2011].

One of the many challenges of SLT (compared to normal text-to-text MT), is that the output of ASR does not contain punctuation. This paper is devoted to finding out what is the most effective method of processing and translating un-punctuated Japanese text.

In this paper, BLEU without punctuation was employed as the measure of translation quality. BLEU is one of the most widely used quality measurements for MT. Punctuation markers are normally taken as tokens during the calculation of BLEU. However, this paper is focused on interpretation systems that output speech instead of text, and therefore punctuation of the target language is less important. Ignoring the punctation on the target side may change the findings concerning different punctuation-handling methods, that have been reported in the related work.

The main contributions of this paper, compared to previous work on punctuation handling, are:

- Employing BLEU without punctuation to evaluate translation quality in order to reflect the quality of synthesized speech output;
- Testing on Japanese-to-English pre-ordering phrase-based MT systems as well as normal phrase-based MT systems;
- Testing on a large Japanese-to-English corpus.

2 Related Work

The work of Evgeny et al. [Matusov et al., 2006] and Peitz [Peitz et al., 2011] are most related to this paper. Three methods of handling punctuation were tested, as follows:

- training MT systems with un-punctuated both source and target text, and predicting punctuation on the output of MT which is therefore unpuctuated;
- training MT systems with un-punctuated source and punctuated target text;
- training MT systems with punctuation on both source and target, and predicting punctuation on the source side before applying MT.

Punctuation from two different methods was evaluated: a hidden n-gram model [Stolcke et al., 1998], and a monolingual MT system (see Section 3 for details). BLEU with punctuation was used to measure translation quality. Experimental results show that the first method performed best on the large vocabulary task of TC-STAR English-to-Spanish, and the

second method performed best on a small vocabulary 2006 Workshop on Spoken Language Translation (IWSLT) Chinese-to-English task.

the annual IWSLT open evaluation campaigns are the most related shared tasks to this paper. Participants are asked to translate ASR transcripts of TED talks. The tasks in IWSLT 2015 included language pairs of German to English, and English to Chinese, Czech, French, German, Tai and Vietnamese [Cettolo et al., 2015]. Peter et al. [Peter et al., 2015] used a monolingual hierarchical MT system which was augmented with a word class language model to predict punctuation. Ha et al. [Ha et al., 2015] used monolingual MT to predict punctuation, case information and sentence segmentation jointly. Kazi et al. [Kzai et al., 2015] trained a classifier based on a recurrent neural network to predict punctuation.

One difference between above work and ours is in the measurement of translation performance. Another difference is that our work focuses on the more challenging task of Japanese-to-English translation. The word order is very different for this language pair and pre-ordering has been shown to be especially effective to alleviate the problems arising from this. We chose to experiment on this pair because we hypothesized that punctuation will have an impact on the re-ordering process, and therefore its presence or absence may affect the quality of translation.

3 Methods

Two methods of predicting punctuation were tested in this paper. The first method was so-called hidden n-gram model [Stolcke et al., 1998]. Hidden n-gram models treat punctuation as hidden events, and tag a stream of word tokens with these hidden events occurring between words. The algorithm underlying hidden n-gram models is a hidden Markov model in which word-event pairs correspond to states and the words to observations.

The second method was to employ a monolingual MT system to translate un-punctuated text into punctuated text[Matusov et al., 2006, Ha et al., 2015]. In this monolingual MT system, the "source language" is un-punctuated text, and "target language" is punctuated text. In addition, the translation process is constrained to be monotonic. The motivation for this method is that, the additional models and tuning techniques of a MT system may help to predict punctuation more correctly. In contrast, a hidden n-gram is based on a language model only [Peitz et al., 2011].

4 Experiments

4.1 Experimental Settings

Experiments were performed on a Japanese-to-English translation task. The experimental corpora were a union of corpora from multiple sources, including shared tasks such as the Basic Travel Expression Corpus [Takezawa et al., 2002], NTCIR Patent Machine Translation [Goto et al., 2013], crawled web data and several in-house parallel resources. Table 1 presents the statistics of sentences and words in the training, development and test sets.

The corpora were processed using a standard procedure for MT. The Japanese text was segmented into words using Mecab [Kudo, 2005]. The English text was tokenized with the tokenization script released with Europarl corpus [Koehn, 2005] and converted to lowercase.

For the MT systems, MOSES was employed as the base phrase-based MT system [Koehn et al., 2007]. GIZA++ with the default Moses settings [Och and Ney, 2000] was employed to align the training corpora. 5-gram interpolated modified Kneser-Ney smoothed language models were learned from the target side of the training corpora using the SRILM [Stolcke et al., 2002] tools. The pre-ordering toolkit was an in-house implementation which was based on binary syntactic constituent tree, and a neural network classifier which was trained in a supervised manner to predict whether or not to swap a constituent node.

For predicting punctuation, the "hidden-ngram" tool from the SRILM toolkit was employed as the hidden n-gram model in our experiments. The language models used by "hidden-ngram" were 5-gram interpolated modified Kneser-Ney smoothed language models were learned from the source side of the training corpora. The monolingual MT systems used for predicting punctuation followed the same settings of bilingual MT systems. The systems were trained on the un-punctuated and punctuated version of the source side of the training corpus.

4.2 Experimental Results

Experimental results on translation quality are presented in Table 2. BLEU without punctuation was the main performance measurement. RIBES's with and without punctuation were also presented since RIBES was effective for distant lanugage pairs including Japanese-English [Isozaki et al., 2010]. In addition, the direct performance of predicting punctuation is shown in Table 3. We make the following observations from these results:

- Predicting punctuation yielded better performance than not predicting punctuation. The gain in BLEU points ranged from 0.67 to 0.88;

Corpus	# Sent. Pairs	Japanese		English	
		# Tokens†	# Words	# Tokens†	# Words
Training	5,134,941	106,044,671	93,672,553	84,371,311	74,733,865
Develop	6,000	180,058	160,688	130,923	114,042
Test	6,000	121,383	104,410	76,188	63,517

Table 1: Experimental Corpora. † Including punctuation.

SMT System	Punctutation Handling		Evaluation (Punc.)		Evaluation (No Punc.)	
	Training Corpus	Test Corpus	RIBES	BLEU	RIBES	BLEU
Phrase-based	Oracle Punc.	Oracle Punc.†	.7184†	.1223†	.6310†	.1125†
	Oracle Punc.	Hidden ngram Punc.	**.7188**	**.1239**	**.6314**	**.1133**
	Oracle Punc.	Monolingual MT Punc.	.7150	.1222	.6308	.1124
	No Punc.	No Punc.	–	–	.6016	.1045
Pre-ordering Phrase-based	Oracle Punc.	Oracle Punc.†	.7405†	.1403†	.6712†	.1311†
	Oracle Punc.	Hidden-ngram Punc.	.7338	**.1398**	.6593	**.1292**
	Oracle Punc.	Monolingual MT Punc.	**.7346**	.1388	**.6642**	.1289
	No Punc.	No Punc.	–	–	.6294	.1222

Table 2: Quality of Machine Translation. † For comparison only, not available in practical applications.

Method	F_1	Precision	Recall
Hidden-ngram Punc.	0.7019	**0.8706**	0.5880
Monolingual MT Punc.	**0.7487**	0.8460	**0.6714**

Table 3: Performance of Predicting Punctuation.

- Predicting punctuation through a monolingual MT system is more accurate than using a hidden ngram in terms of F_1 score. The motivation presented in Section 3 is valid;
- Even though the two methods of predicting punctuation noticeably different in terms of F_1 score, they yielded similar translation quality. The difference in BLEU points ranged from 0.03 to 0.09.
- In phrased-based MT, predicted punctuation yielded comparable quality compared to oracle punctuation (0.01 to 0.08 BLEU points); in pre-ordering phrase-based MT, predicted punctuation yielded clearly worse performance than oracle punctuation (0.19 to 0.22 BLEU points).
- Phrase-based MT is quite insensitive to the punctuation quality. Predicting punctuation using a hidden-ngram model or a monolingual MT system or even using oracle punctuation all lead. to similar translation quality.
- Pre-ordering is more sensitive to the punctuation quality. A substantial difference between using oracle punctuation and predicted punctation was observed. However, we found no difference in performance among the methods predicted punctuation.

5 Conclusion

This paper explored multiple methods of handling punctuation for spoken language translation. We draw the following conclusions from the experiments.

According the findings in this paper, to develop a speech-to-speech Japanese-to-English translation system, a good strategy would be to train an MT system on punctuated text, and then punctuate the output from an ASR system using either a hidden n-gram model or a monolingual MT system, before translation.

References

[Cettolo et al., 2015] Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2015). The IWSLT 2015 Evaluation Campaign. In *Proceedings of the twelveth International Workshop*

on Spoken Language Translation (IWSLT), Da Nang, Veitnam, pages 2–14.

[Goto et al., 2013] Goto, I., Chow, K. P., Lu, B., Sumita, E., and Tsou, B. K. (2013). Overview of the patent machine translation task at the ntcir-10 workshop. In *Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-10*.

[Ha et al., 2015] Ha, T.-L., Niehues, J., Cho, E., Mediani, M., and Waibel, A. (2015). The KIT Translation Systems for IWSLT 2015. In *Proceedings of the twelveth International Workshop on Spoken Language Translation (IWSLT), Da Nang, Veitnam*, pages 62–69.

[Isozaki et al., 2010] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (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. Association for Computational Linguistics.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

[Kudo, 2005] Kudo, T. (2005). Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

[Kzai et al., 2015] Kzai, M., Thompson, B., Salesky, E., Anderson, T., Erdmann, G., Hansen, E., Ore, B., Young, K., Gwinnup, J., Hutt, M., and May, C. (2015). The MITLL-AFRL IWSLT 2015 Systems. In *Proceedings of the twelveth International Workshop on Spoken Language Translation (IWSLT), Da Nang, Veitnam*, pages 23–30.

[Matusov et al., 2006] Matusov, E., Mauser, A., and Ney, H. (2006). Automatic sentence segmentation and punctuation prediction for spoken language translation. In *IWSLT*, pages 158–165.

[Och and Ney, 2000] Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1086–1090. Association for Computational Linguistics.

[Peitz et al., 2011] Peitz, S., Freitag, M., Mauser, A., and Ney, H. (2011). Modeling punctuation prediction as machine translation. In *IWSLT*, pages 238–245.

[Peter et al., 2015] Peter, J.-T., Toutounchi, F., Peitz, S., Bahar, P., Guta, A., and Ney, H. (2015). The RWTH Aachen German to English MT System for IWSLT 2015. In *Proceedings of the twelveth International Workshop on Spoken Language Translation (IWSLT), Da Nang, Veitnam*, pages 15–22.

[Stolcke et al., 2002] Stolcke, A. et al. (2002). Srilman extensible language modeling toolkit. In *INTERSPEECH*.

[Stolcke et al., 1998] Stolcke, A., Shriberg, E., Bates, R. A., Ostendorf, M., Hakkani, D., Plauche, M., Tür, G., and Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. In *ICSLP*, pages 2247–2250.

[Takezawa et al., 2002] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC*, pages 147–152.