Interspeech 2018
2-6 September 2018, Hyderabad

Low-Latency Neural Speech Translation

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber and Alex Waibel

Institute for Anthropomatics and Robotics
KIT - Karlsruhe Institute of Technology, Germany
firstname.lastname@kit.edu

Abstract

Through the development of neural machine translation, the quality of machine translation systems has been improved significantly. By exploiting advancements in deep learning, systems are now able to better approximate the complex mapping from source sentences to target sentences. But with this ability, new challenges also arise. An example is the translation of partial sentences in low-latency speech translation. Since the model has only seen complete sentences in training, it will always try to generate a complete sentence, though the input may only be a partial sentence. We show that NMT systems can be adapted to scenarios where no task-specific training data is available. Furthermore, this is possible without losing performance on the original training data. We achieve this by creating artificial data and by using multi-task learning. After adaptation, we are able to reduce the number of corrections displayed during incremental output construction by 45%, without a decrease in translation quality.

Index Terms: speech translation, low-latency

1. Introduction

Neural machine translation (NMT) is currently the state-of-the-art in machine translation, significantly improving translation quality in text translation [1] as well as in speech translation [2], where the translation input is the output from a speech recognizer. The main strength of neural machine translation is improved output fluency compared to traditional approaches, such as rule-based or statistical machine translation.

However, while the model is able to capture more complex dependencies between the source and target languages, it relies heavily on training data examples to do so. As a consequence, the model lacks the robustness at test time to handle data that is fundamentally different from what was seen in training. There are several scenarios where this can be observed. For example, if the input is incorrectly cased, or if a different dialect is presented at test time which has different spelling or phrasings.

In this work, we will concentrate on a speech translation use case in which the translation system is required to provide an initial translation in real time, before the complete sentence has been spoken. To this end, [3] presented an approach where partial sentences are translated and later replaced if necessary with the translations of the complete sentences. While we focus on this use case, the results of this work can be easily adapted to other use cases where there are differences between the training and testing scenarios.

When applying partial sentence translation to neural machine translation systems, we encounter the problem that the MT system has only been trained on complete sentences, and thus the decoder is biased to generate complete target sentences. When receiving inputs which are partial sentences, the translation outputs are not guaranteed to exactly match with the input content, which can be seen in Example 1.1. We observe that the translation is often "fantasized" by the model to be a full sentence, as would have occurred in the training data. In the example below, although the English input ends with *'all of'*, the system generates the translation *'todo el mundo'* (Engl. *'all of the world'*). In other cases, the decoder can fall into an over-generation state and repeat the last word several times (eg. 'debería, debería, debería').

Example 1.1. Examples of challenges in using NMT to translate spoken utterances.

English:	I encourage all of
Spanish:	yo animo a todo el mundo .
English:	now , I should
Spanish:	ahora debería , debería , debería .

In this work, we aim to remedy the problem of partial sentence translation in NMT. Ideally, we want a model that is able to generate appropriate translations for incomplete sentences, without any compromise during other translation use cases. Our approach involves using multi-task learning and the automatic generation of parallel corpora in which both the source and the target sentences are incomplete sentences.

2. Related Work

The main topic of our work is adapting to different types of inputs for neural machine translation. Previous works have focused on domain mismatch between the training data and test data [4]. In the case of speech translation, the model may only be exposed to specific issues arising from speech recognition outputs during test time. Since speech input can carry over errors from the ASR system to translation, it is necessary to adapt the model to noisier circumstances. To handle this scenario, previous work has proposed introducing artificially corrupted inputs at training time [5] or direct training on lattices produced by the speech recognition system [6].

Multi-task learning has commonly been used in various NLP problems to jointly train a single model for several well-established NLP tasks, reducing overhead and improving performance. Such implementations can be seen using the encoder-decoder model with attention mechanism [7], in which a single model is trained for part-of-speech tagging, named-entity tagging and machine translation simultaneously [8].

Regarding low-latency speech translation, various approaches to translating small text segments exist using statistical phrase-based models [9, 10, 11] or neural networks [12]. Due to the fact that the whole input sentence is not available, it is necessary to find a compromise between translation quality and latency. The decoding process of the neural models also needs to be changed to deal with a stream of inputs, which is non-negligible. It is also possible to use the revision strategy to

10.21437/Interspeech.2018-1055

update the partial translations, which has been implemented in practical systems [3].

3. Low-Latency Speech Translation

In the practice of simultaneous speech translation, translation quality is not the only criterion; it is also important to produce a translation for a spoken utterance in real-time and at low latency. Since a speaker's utterance can be arbitrarily long, it is necessary for the translation system to start operating before the speaker stops, in which case the system input will consist of incomplete spoken segments instead of full sentences.

We explore the translation revision method of [3], which has been successfully applied to statistical translation systems. The key idea of this method is that the system iteratively revises translations by re-translating new messages sent by the speech recognition component. These newly sent messages are either replacements of or concatenations to previous ones. As a result, the user sees the translation continually updated in the interface.

For example, for the sentence *'I encourage all of you'*, the system first receives only the beginning of the sentence *'I'*, with the intermediate translation being *'yo'*. Afterwards, it receives an update which is the continuation of the previous one: *'I encourage all of'*. The resulting translation from a typical neural model would be *'yo animo a todo el mundo'*, hypothesizes a final word. Finally, the whole source sentence is available, and the MT system will update the translation of the sentence to *'yo animo a todos ustedes'*.

As can be seen from the above example, in the last translation step the interface has to update the words *'todo el mundo'* for *'todos ustedes'*, which was generated only when the full sentence was available. As a result, we experience a delay which comes from the second to last translation step, which is longer than necessary. The interface also suffers from the update, since nearly half of the sentence needs to be replaced. Despite the fact that the final translation quality in the end does not depend on the processing of each segment, the intermediate translation outputs may change drastically due to source sequence updates. The problem is exacerbated by the fact that a neural machine translation model trained with normal parallel corpora is not able to flexibly generate translation for partial input segments, which were not available during training. The aim of our work is to build an online machine translation models which minimizes the number of words which need to be corrected until the full sentence has been seen. We aim to minimizing such criteria while maintaining translation quality for the complete utterance.

4. Partial Translation

As motivated in the introduction, an out-of-the-box NMT system struggles with partial input sentences. In order to improve the flexibility of the model, we investigate generating parallel corpora in which the input and output are also partial sentences. Subsequently, we adjust the training process to make use of the data in order to build a single system that is proficient at translating partial as well as full sentences.

4.1. Generating Partial Parallel Corpora

In order to build a system that is good at translating partial sentences, we need to build partial sentence training data. Since such data is not available, we investigated methods to build an artificial training set from standard parallel data. This has the advantage that the methods can be easily applied to any language pair and domain and no new data has to be collected.

Creating the source data is straightforward. Given a source sentence $S = s_1 \dots s_I$, we can generate I input samples $S^{(i)} = s_1 \dots s_i$ by selecting the first i words. The challenge arises from defining the correct translation for this source string. Since we are using this data in a low-latency speech translation system, the translation of the partial sentence should meet several conditions. First, it should be as long as possible in order to minimize the latency of the system. If we always used only the first word, we would not improve latency over a system that waits until the sentence is finished. Furthermore, to minimize the number of corrections, the translation of $S^{(i)}$ should be a substring of $S^{(i')}$ for all $i' > i$. Thus the translation of $S^{(i)}$ should be a substring of the reference translations.

One possible solution is to take the reference translation of the whole sentence. But, it is unrealistic to be able to generate the whole target sentence from only a single word in the source string. Therefore, we investigated two methods to select a reasonable substrings from the reference translation.

The first method is motivated by the idea that the translation should constantly generate longer target sequences when receiving longer source segments. Furthermore, word reordering may exist between two languages, for many languages sentence structure is similar. Consequently, a first approximation is to use the same proportion of words from the reference translation as we have from the source sentence.

One problem in this case is that we introduce additional noise. If the word order is different, we force the system to guess the words coming next in the source sentence. To avoid this problem, we first generate a word alignment using Giza++ [13] between the source and target sentences. Then, we select the longest prefix of the reference so that no target words in the prefix are aligned to source words that are not in the partial sentences:

$$T^{(i)} = \arg\max_{j \in J} \{t_1 \dots t_j | \vee j' \leq j : a(j') \leq i\} \quad (1)$$

4.2. Training Process

Multi-task training Given the artificially produced training data, a first step is to train a model on the newly created partial sentence data and use it for speech translation only. Since both tasks are very similar, we first pre-train a standard NMT system and then fine-tune the system to translate partial sentences.

The disadvantage of this approach is that the performance on complete sentences might drop, since NMT models tend to rapidly forget what they have learned before. In order to have a system that is able to generate high quality translations of both complete sentences and partial sentences, we opt to use multi-task learning, treating these as two separate tasks. In our approach, we randomly subsample the partial sentence training data to make it the same size as the original training data, so that the model can put equal emphasis on both tasks. The mixed training data then has twice as many sentences as the baseline system, but significantly less than the system using all partial sentences. Then, we fine-tune the NMT system on both tasks: translating complete sentences as well as the partial counterparts.

Sequence level optimization Beside multi-task learning, we can also guide the search operation of the model so that the generated output is better matched to the source input. We use reinforcement learning with policy gradient methods [14, 15] to

train the model to maximize the GLEU score [16], which is the combination of n-gram precision and recall. This reward function restricts the model from generating sentences that are too long. Since this method is known to have high variance gradients, we follow the method in [17], which estimates a baseline using greedy search to reduce the variance.

5. Experimental Results

We evaluate the method on three different languages pairs: English-Spanish, English-French and German-English.

5.1. System description

For all experiments, we trained systems on the Europarl [18] and the WIT-TED corpora [19] and tested on test sets from the IWSLT evaluation campaign. All systems were adapted to the TED domain by fine-tuning on the in-domain TED data. For the English→Spanish and English→French directions, we also optimize the models towards GLEU scores [16] using reinforcement learning (RL). For partial sentence translation (both data generation and training), we utilize only the TED corpus. We used the OpenNMT-py toolkit [20] to train the systems. For each language pair, we jointly trained BPE [21] for the source and target languages.

5.2. Evaluation metrics

We evaluated the translation quality using the BLEU score [22]. Since the ASR output uses automatic sentence segmentation, we need to re-segment the translation to fit the reference translations. Therefore, we used the method described in [23], where the automatic translation is re-segmented in a way that minimizes the word error rate to the reference.

In addition, we also need to measure the extent to which we are able to reduce the number of corrections in the spoken language translation (SLT) system. To do so, we roll out all updates from the ASR system and translate each. For each updated translation, we measure the overlap between pairs of consecutive updates s_t and s_{t+1} and calculate the amount of re-writing necessary to produce s_{t+1} after s_t. Specifically, the number of corrected words is calculated by the length of the translation of s_t, minus the length of the common prefix both translations. As illustrated in the example in Section 3, the final update would lead to 3 corrected words (Word Up). Since an intermediate word change will force the user to reread all following words, our metric also counts all words following the first corrected word as corrected. We also report the number of messages where at least one word is corrected (Messg. Up.).

5.3. Experiments

Initial results Our initial results on the English→Spanish translation task are shown in Table 1. We report results in BLEU on the test and validation set with full sentence translation. Next, we report results on the test data with all possible prefixes, and finally, results on the ASR output. In the initial experiments, we use length ratio to determine the length of the reference for the partial translations as described in Section 4.1. The baseline system is only trained on complete sentences. All other systems use the baseline system, and continue training using different strategies. The system "Partial" is fine-tuned on all partial sentences. As shown in the first two lines of Table 1, the final translation quality drops significantly by $\sim$1 BLEU point. On the other hand, the BLEU score calculated on only partial sentences improves by $\sim$3 BLEU points. As shown by the number of tokens, the length of translations is reduced by 25%. So, a major problem of the baseline system is that it generates translations that are too long for the partial sentences. When testing on the ASR output in the last two columns, we see that the translation quality of the final hypothesis also drops, but the number of words which are updated is reduced by 45%. Also, the number of messages where at least one word changed is reduced by 20%. The system in the third line uses multi-task learning. In this case, the system is trained to perform both tasks: translation of partial and full sentences. Using this technique, we can combine the advantages of both models and maximize the translation quality of the final hypothesis, while minimizing the number of updates. The system has the same translation quality as the baseline system, with the same reduction in updates as the partial system.

Performance w.r.t the artificial data In the second set of experiments, we analyzed the use of the artificial data. We used the alignment-based method to generate the references for the partial sentences. In this case, we again fine-tuned used multi-task learning. As shown in the results in Table 1, there is no clear performance difference between the two approaches. When translating text input, the system using length-ratio references is better, while the system using alignment-based methods is better on partial sentence and speech translation. Since the length-ratio based method is simpler, we used this approach for the remainder of this work.

Sequence-level optimization Finally, we also used reinforcement learning (RL) to optimize the performance of the system directly towards BLEU. These systems are first trained using cross-entropy and continue training using reinforcement learning. Here again, we have a baseline system trained only on the full sentences, and a multi-task system trained on both the full and partial sentences (final two lines of Table 1). As above, we observe that with multi-task learning, we do not lose performance on full sentences, while we can significantly reduce the number of updates. In this case, the number of words updated is further reduced, reaching more than 50% less than the baseline.

English→French We also performed experiments using two English to French systems, which are summarized in Table 2. We again have a baseline system trained with cross-entropy, and a baseline system where training is continued with RL. The models were evaluated on the full sentences and the mixed set of complete and partial sentences, as well as on the ASR output. Similar to the other translation directions, we improved translation performance on partial sentences and reduced the number of rewritten tokens for the SLT output by using multi-task learning. Interestingly, using reinforcement learning also helped us improve the performance on partial sentences. The RL criteria evaluates the n-gram precision as well as the recall of the translation, which is punished when the generated output is too long. Both methods can be combined to achieve the overall best performance, which reduced the rewritten tokens by up to 50% without compromising translation performance.

German to English Finally, we also performed experiments on English to German as shown in Table 3. Again, we can improve the number of rewrites needed to produce the final output.

System	Valid (tst2011)	Test (tst2012)	TEDTest Partial		SLT (tst2010)		
	BLEU	BLEU	BLEU	length (tokens)	BLEU	Word Up	Mssg. Up.
Baseline	36.86	31.33	26.66	509K	25.97	182K	15.0K
Partial	35.45	30.29	29.48	375K	25.54	98K	11.8K
Multi-task	37.05	31.27	30.09	376K	26.00	101K	12.0K
Align. ref.	37.13	31.06	30.29	371K	26.30	98K	11.5K
RL	37.21	31.25	30.08	540K	26.61	179K	15.1K
RL + Multi	37.50	31.21	30.31	377K	26.77	82K	11.5K

Table 1: *Results for English to Spanish*

System	tst2010		SLT(tst2010)		
	Final	Mix	BLEU	Word Up	Mssg. Up.
Baseline	34.11	31.18	23.84	216K	16.3K
Multi	34.40	34.71	23.83	128K	13.5K
RL	35.08	34.09	24.31	140K	15.0K
RL +Multi	34.84	42.51	24.23	99K	12.1K

Table 2: *Results for English to French*

For this language pair, however, the number of updates messages is only slightly reduced. The reason for this might be the larger reordering needed between the two language pairs.

System	SLT(tedX2015)		
	BLEU	Words Up.	Messg. Up
Baseline	15.52	246K	23.6K
Multi-task	15.64	172K	23.1K

Table 3: *Results for German to English*

5.4. Examples

In addition to the evaluation in the last section, improvements using our approach can be seen through examples for English-Spanish and German-English, shown in Table 4.

In both cases, we see that the baseline system is not able to generate translations for very short sequences. In this case, the last word is repeated several times. In addition, since the NMT system is tested on input it is not accustomed to, we see that the NMT decoder relies more heavily on language modeling information and completes the sentence in a way that is typical in the target language, regardless of the source input. For example, we see the added *and so on* in the first message for the German to English system. The multi-task system, however, has been trained to handle partial sentences and is therefore able to generate a correct translation.

Finally, another interesting point is how the systems handle punctuation. The baseline model for German to English is only able to generate the correct translation if the sentences ends with a punctuation mark. This can be seen in the two last examples, which contain the same words, but only the second has punctuation. The multi-task system, in contrast, is able to generate the correct translation before the input is correctly punctuated. While most errors happen in very short (one or two words) partial sentences, longer partial sentences can also be problematic because of issues like punctuation, which suggests that ignoring short sentences is not a proper solution.

	English to Spanish
Input:	now,
Baseline:	ahora ,
Multi-task:	ahora ,
Input :	now, I should
Baseline:	ahora debería , debería , debería .
Multi-task:	ahora debería
Input :	now, I should men
Baseline:	ahora debería hombres hombres .
Multi-task:	ahora debería
Input :	now, I should mention that this
Baseline:	ahora debería mencionar esto .
Multi-task:	ahora , debo mencionarlo .
	German to English
Input:	Und
Baseline:	And and and and and so on.
Multi-task:	And
Input :	Und ich habe
Baseline:	And I have
Multi-task:	And I have
Input :	Und ich empfehle Ihnen
Baseline:	And I recommend you to you
Multi-task:	And I recommend you
Input :	Und ich empfehle Ihnen .
Baseline:	And I recommend you .
Multi-task:	And I recommend you .

Table 4: *Examples for English-Spanish and German-English*

6. Conclusion

Low latency translation is important for real-time speech translation systems. To address this challenge, we improve upon a mechanism to translate partial speech input and make updates in real-time. Our main contribution is to propose a simple method to deal with scenarios where data at inference time is different from the training data, which can be resolved with adaptation. We first showed that using simple techniques to generate artificial data are effective to get more fluent output with less correction. We also illustrated that multi-task learning can help adapt the model to the new inference condition, without losing the original capability to translate full sentences. Combining these two ideas, we are able to maintain high quality translation at low latency, minimizing the number of corrected words by 45%, which significantly improves user experience for practical applications.

7. Acknowledgements

This work was supported by the Carl-Zeiss-Stiftung.

8. References

[1] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 131–198. [Online]. Available: http://www.aclweb.org/anthology/W/W16/W16-2301

[2] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann, "Overview of the iwslt 2017 evaluation campaign," in *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, Tokio, Japan, 2017.

[3] J. Niehues, T. S. Nguyen, E. Cho, T.-L. Ha, K. Kilgour, M. Müller, M. Sperber, S. Stüker, and A. Waibel, "Dynamic transcription for low-latency speech translation," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)*, San Francisco, California, USA, 2016, pp. 2513–2517.

[4] C. Kobus, J. M. Crego, and J. Senellart, "Domain control for neural machine translation," *Proceddings of Recent Advances in Natural Language Processing (RANLP 2017)*, 2016.

[5] M. Sperber, J. Niehues, and A. Waibel, "Toward Robust Neural Machine Translation for Noisy Input Sequences," in *International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2017.

[6] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Neural Lattice-to-Sequence Models for Uncertain Inputs," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[7] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," *CoRR*, vol. abs/1511.06114, 2015.

[8] J. Niehues and E. Cho, "Exploiting linguistic resources for neural machine translation using multi-task learning," in *Proceedings of the Second Conference on Statistical Machine Translation (WMT 2017)*, Copenhagen, Dennmark, Denmark, 2017.

[9] V. K. R. Sridhar, J. Chen, S. Bangalore, A. Ljolje, and R. Chengalvarayan, "Segmentation Strategies for Streaming Speech Translation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, Atlanta, Georgia, USA, 2013, pp. 230–238.

[10] Y. Oda, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Optimizing Segmentation Strategies for Simultaneous Speech Translation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, Maryland, USA, 2014.

[11] H. S. Shavarani, M. Siahbani, R. M. Seraj, and A. Sarkar, "Learning Segmentations that Balance Latency versus Quality in Spoken Language Translation," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam, 2015.

[12] J. Gu, G. Neubig, K. Cho, and V. O. K. Li, "Learning to translate in real-time with neural machine translation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, 2017, pp. 1053–1062.

[13] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[14] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[15] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[16] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[17] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," *arXiv preprint arXiv:1612.00563*, 2016.

[18] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings The Tenth Machine Translation Summit (MT Summit X)*, vol. 5, Phuket, Thailand, 2005.

[19] M. Cettolo, C. Girardi, and M. Federico, "Wit: Web inventory of transcribed and translated talks," in *Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, 2012, pp. 261–268.

[20] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," in *Proceedings of the 55st Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, Canda, 2017.

[21] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54st Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany, 2016.

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, Pennsylvania, 2002, pp. 311–318.

[23] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating machine translation output with automatic sentence segmentation," in *Proceedings of the Second International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburgh, USA, 2005.