Agreement on Target-bidirectional Neural Machine Translation

Lemao Liu, Masao Utiyama, Andrew Finch, Eiichiro Sumita
National Institute of Information and Communications Technology (NICT)
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, Japan
{lmliu,first.last}@nict.go.jp

Abstract

Neural machine translation (NMT) with recurrent neural networks, has proven to be an effective technique for end-to-end machine translation. However, in spite of its promising advances over traditional translation methods, it typically suffers from an issue of unbalanced outputs, that arise from both the nature of recurrent neural networks themselves, and the challenges inherent in machine translation. To overcome this issue, we propose an agreement model for neural machine translation and show its effectiveness on large-scale Japanese-to-English and Chinese-to-English translation tasks. Our results show the model can achieve improvements of up to 1.4 BLEU over the strongest baseline NMT system. With the help of an ensemble technique, this new end-to-end NMT approach finally outperformed phrase-based and hierarchical phrase-based Moses baselines by up to 5.6 BLEU points.

1 Introduction

Recurrent neural network (RNN) has achieved great successes on several structured prediction tasks (Graves, 2013; Watanabe and Sumita, 2015; Dyer et al., 2015), in which RNNs are required to make a sequence of dependent predictions. One of its advantages is that an unbounded *history* is available to enrich the context for the prediction at the current time-step.

Despite its successes, recently, (Liu et al., 2016) pointed out that the RNN suffers from a fundamental issue of generating unbalanced outputs: that is to say the suffixes of its outputs are typically worse than the prefixes. This is due to the fact that later predictions directly depend on the accuracy of previous predictions. They empirically demonstrated this issue on two simple sequence-to-sequence learning tasks: machine transliteration and grapheme-to-phoneme conversion.

On the more general sequence-to-sequence learning task of machine translation (MT), neural machine translation (NMT) based on RNNs has recently become an active research topic (Sutskever et al., 2014; Bahdanau et al., 2014). Compared to those two simple tasks, MT involves in much larger vocabulary and frequent reordering between input and output sequences. This makes the prediction at each time-step far more challenging. In addition, sequences in MT are much longer, with averaged length of 36.7 being about 5 times longer than that in grapheme-to-phoneme conversion. Therefore, we believe that the history is more likely to contain incorrect predictions and the issue of unbalanced outputs may be more serious. This hypothesis is supported later (see Table 1 in §4.1), by an analysis that shows the quality of the prefixes of translation hypotheses is much higher than that of the suffixes.

To address this issue for NMT, in this paper we extend the agreement model proposed in (Liu et al., 2016) to the task of machine translation. Its key idea is to encourage the agreement between a pair of target-directional (left-to-right and right-to-left) NMT models in order to produce more balanced translations and thus improve the overall translation quality. Our contribution is two-fold:

- We introduce a simple and general method to address the issue of unbalanced outputs for

Proceedings of NAACL-HLT 2016, pages 411–416,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

NMT (§3). This method is robust without any extra hyperparameters to tune and is easy to implement. In addition, it is general enough to be applied on top of any of the existing RNN translation models, although it was implemented on top of the model in (Bahdanau et al., 2014) in this paper.

- We provide an empirical evaluation of the technique on large scale Japanese-to-English and Chinese-to-English translation tasks. The results show our model can generate more balanced translation results, and achieves substantial improvements (of up to 1.4 BLEU points) over the strongest NMT baseline (§4). With the help of an ensemble technique, our new end-to-end NMT gains up to 5.6 BLEU points over phrase-based and hierarchical phrase-based Moses (Koehn et al., 2007) systems. [1]

2 Overview of Neural Machine Translation

Suppose $\mathbf{x} = \langle x_1, x_2, \cdots, x_m \rangle$ denotes a source sentence, $\mathbf{y} = \langle y_1, y_2, \cdots, y_n \rangle$ denotes a target sentence. In addition, let $x_{<t} = \langle x_1, x_2, \cdots, x_{t-1} \rangle$ denote a prefix of $\mathbf{x}$. Neural Machine Translation (NMT) directly maps a source sentence into a target within a probabilistic framework. Formally, it defines a conditional probability over a pair of sequences $\mathbf{x}$ and $\mathbf{y}$ via a recurrent neural network as follows:

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}; \theta) &= \prod_{t=1}^{n} p(y_t \mid y_{<t}, \mathbf{x}; \theta) \\ &= \prod_{t=1}^{n} \mathbf{softmax}\big(g(h_t)\big)[y_t] \end{aligned} \tag{1}$$

where θ is the set of model parameters; h_t denotes a hidden state (i.e. a vector) of $\mathbf{y}$ at timestep t; g is a transformation function from a hidden state to a vector with dimension of the target-side vocabulary size; $\mathbf{softmax}$ is the softmax function, and $[i]$ denotes the i_{th} component in a vector.[2] Furthermore, $h_t = f(h_{t-1}, c(\mathbf{x}, y_{<t}))$ is defined by a recurrent function over both the previous hidden state h_{t-1} and the context $c(\mathbf{x}, y_{<t})$. [3] Note that both h_t and $c(\mathbf{x}, y_{<t})$ have dimension d for all t.

In this paper, we develop our model on top of the neural machine translation approach of (Bahdanau et al., 2014), and we refer the reader this paper for a complete description of the model, for example, the definitons of f and c. The proposed method could just as easily been implemented on top of any other RNN models such as that in (Sutskever et al., 2014).

3 Agreement on Target-bidirectional NMT

In this section, we extend the method in (Liu et al., 2016) to address this issue of unbalanced outputs for NMT. The key idea is to: 1) train two kinds of NMT, i.e. one generating targets from *left-to-right* while the other from *right-to-left*; 2) encourage the agreement between them by joint search.

3.1 Training

The training objective function for our **agreement** (or **joint**) model is formalized as follows:

$$\ell = \sum_{\langle \mathbf{x}, \mathbf{y} \rangle} \log p(\mathbf{y} \mid \mathbf{x}; \theta_1) + \log p(\mathbf{y}^r \mid \mathbf{x}; \theta_2) \tag{2}$$

where $\mathbf{y}^r = \langle y_n, y_{n-1} \cdots, y_1 \rangle$ is the reverse of sequence $\mathbf{y}$; $p(\mathbf{y} \mid \mathbf{x}; \theta_1)$ denotes the **left-to-right** model with parameters θ_1, while $p(\mathbf{y}^r \mid \mathbf{x}; \theta_2)$ denotes the **right-to-left** model with parameters θ_2, as defined in Eq.(1); and $\langle \mathbf{x}, \mathbf{y} \rangle$ ranges over a given training dataset. Following (Bahdanau et al., 2014), we employ AdaDelta (Zeiler, 2012) to minimize the loss ℓ.

Note that, in parallel to our efforts, Cheng et al. (2016) has explored the agreement idea for NMT close to ours. However, unlike their work on the agreement between source and target sides in the spirit of the general idea in (Liang et al., 2006), we focus on the agreement between left and right directions on the target side oriented to the natural issue of NMT itself. Although our model is orthogonal to theirs, one of our advantage is that our model does not rely on any additional hyperparameters to

[1] The absolute gains of our model can be expected to be further increased by applying the well-known techniques in (Jean et al., 2015; Luong et al., 2015) that address the problems presented by unknown words, but these techniques are beyond the scope of this paper.

[2] In that sense, y_t in Eq.(1) also denotes the index of this word in its vocabulary.

[3] Both hidden states and context vectors are dependent on the model parameter θ, but we remove it from the expressions here for simplicity.

encourage agreement, given that tuning such hyperparameters for NMT is too costly.

3.2 Approximate Joint Search

Given a source sentence $\mathbf{x}$ and model parameters $\langle \theta_1, \theta_2 \rangle$, decoding can be formalized as follows:

$$\hat{\mathbf{y}} = \mathop{\mathbf{argmax}}_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}; \theta_1) \times p(\mathbf{y}^r \mid \mathbf{x}; \theta_2)$$

As pointed out by (Liu et al., 2016), it is NP-hard to perform an exact search, and so we adapt one of their approximate search methods for the machine translation scenario. The basic idea consists of two steps: 1) run beam search for forward and reverse models independently to obtain two k-best lists; 2) re-score the union of two k-best lists using the joint model to find the best candidate. We refer to the reader to (Liu et al., 2016) for further details.

4 Experiments

We conducted experiments on two challenging translation tasks: Japanese-to-English (JP-EN) and Chinese-to-English (CH-EN), using case-insensitive BLEU for evaluation.

For the JP-EN task, we use the data from NTCIR-9 (Goto et al., 2011): the training data consisted of 2.0M sentence pairs, The development and test sets contained 2K sentences with a single referece, respectively. For the CH-EN task, we used the data from the NIST2008 Open Machine Translation Campaign: the training data consisted of 1.8M sentence pairs, the development set was nist02 (878 sentences), and the test sets are were nist05 (1082 sentences), nist06 (1664 sentences) and nist08 (1357 sentences).

Four baselines were used. The first two were the conventional state-of-the-art translation systems, phrase-based and hierarchical phrase-based systems, which are from the latest version of well-known Moses (Koehn et al., 2007) and are respectively denoted as Moses and Moses-hier. The other two were neural machine translation systems implemented using the open source NMT toolkit (Bahdanau et al., 2014):[4] left-to-right NMT (**NMT-l2r**) and right-to-left NMT (**NMT-r2l**). The proposed joint model (**NMT-J**) was also implemented using NMT (Bahdanau et al., 2014).

Systems	Prefix	Suffix
NMT-l2r	29.4	25.4
NMT-r2l	26.2	26.7
NMT-J	29.5	28.6

Table 1: Quality of 5-word prefixes and suffices of translations in the JP-EN test set, evaluated using partial BLEU.

We followed the standard pipeline to train and run Moses. GIZA++ (Och and Ney, 2000) with grow-diag-final-and was used to build the translation model. We trained 5-gram target language models using the training set for JP-EN and the Gigaword corpus for CH-EN, and used a lexicalized distortion model. All experiments were run with the default settings except for a distortion-limit of 12 in the JP-EN experiment, as suggested by (Goto et al., 2013).[5] To alleviate the negative effects of randomness, the final reported results are averaged over five runs of MERT.

To ensure a fair comparison, we employed the same settings for all NMT systems. Specifically, except for the maximum sequence length (seqlen, which was to 80), and the stopping iteration which was selected using development data, we used the default settings set out in (Bahdanau et al., 2014) for all NMT-based systems: the dimension of word embedding was 620, the dimension of hidden units was 1000, the batch size was 80, the source and target side vocabulary sizes were 30000, and the beam size for decoding was 12. Training was conducted on a single Tesla K80 GPU, and it took about 6 days to train a single NMT system on our large-scale data.

4.1 Results and Analysis on the JP-EN Task

In §1, it was claimed that NMT generates unbalanced outputs. To demostrate this, we have to evaluate the partial translations, which is not trivial (Liu and Huang, 2014). Inspired by (Liu and Huang, 2014), we employ the idea of partial BLEU rather than potential BLEU, as there is no future string concept during NMT decoding. In addition, since the lower n-gram (for example, 1-gram) is easier to be aligned to the uncovered words in source side,

[4] See https://github.com/lisa-groundhog/GroundHog/tree/master/experiments/nmt.

[5] This configuration achieved the significant improvements over the default setting on JP-EN.

Systems	dev	test
Moses	27.9	29.4
Moses-hier	28.6	30.2
NMT-l2r	31.5	32.4
NMT-r2l	31.5	32.6
NMT-J	33.0	34.1
NMT-l2r-5	32.6	33.7
NMT-r2l-5	33.0	34.3
NMT-J-5	**33.8**	**35.0**
NMT-l2r-10	32.5	33.6
NMT-r2l-10	33.0	34.2

Table 2: BLEU comparison of the proposed model NMT-Joint with three baselines on JP-EN task.

which might negatively affect the absolute statistics of evaluation,[6] we employ the partial 4-gram as the metric to evaluate the quality of partial translations (both prefixes and suffixes). In Table 1, we can see that the prefixes are of higher quality than the suffixes for a single left-to-right model (NMT-l2r). In contrast to this, it can be seen that our joint model (NMT-J) that includes one left-to-right and one right-to-left model, successfully addresses this issue, producing balanced outputs.

Table 2 shows the main results on the JP-EN task. From this table, we can see that, although a single NMT model (either left-to-right or right-to-left) comfortably outperforms the Moses and Moses-hier baselines, our simple NMT-J (with one l2r and one r2l NMT model) obtain gains of 1.5 BLEU points over a single NMT. In addition, the more powerful joint model NMT-J-5, which is an ensemble of five l2r and five r2l NMT models, gains 0.7 BLEU points over the strongest NMT ensemble NMT-r2l-5, i.e. an ensemble of five r2l NMT models. The ensemble of joint models achieved considerable gains of 5.6 and 4.8 BLEU points over the state-of-the-art Moses and Moses-hier, respectively. To the best of our knowlege, it is the first time that an end-to-end neural machine translation system has achieved such improvements on the very challenging task of JP-EN translation.

Systems	nist05	nist06	nist08
Moses	35.4	33.7	25.0
Moses-hier	35.6	33.8	25.3
NMT-l2r	34.2	34.9	27.7
NMT-r2l	34.0	34.1	26.9
NMT-J	36.8	36.9	28.5
NMT-l2r-5	37.0	37.5	28.2
NMT-r2l-5	36.9	37.1	27.3
NMT-J-5	**37.5**	**38.9**	**28.8**

Table 3: BLEU comparison of the proposed model NMT-Joint with baselines on CH-EN task.

One might argue that our NMT-J-5 contained ten NMT models in total, while the NMT-l2r-5 or NMT-r2l-5 only used five models, and thus such a comparison is unfair. Therefore, we integrated ten NMT models into the NMT-r2l-10 ensemble. In Table 2, we can see that NMT-r2l-10 is not necessarily better than NMT-r2l-5, which is consistent with the findings reported in (Zhou et al., 2002).

4.2 Results on the CH-EN Task

Table 3 shows the comparison between our method and the baselines on the CH-EN task.[7] The results were similar in character to the results for JP-EN. The proposed joint model (NMT-J-5) consistently outperformed the strongest neural baseline (NMT-l2r-5), an ensemble of five l2r NMT models, on all the test sets with gains up to 1.4 BLEU points. Furthermore, our model again achieved substantial gains over the Moses and Moses-hier systems, in the range 1.9∼5.2 BLEU points, depending on the test set.

5 Related Work

Target-bidirectional transduction techniques were pioneered in the field of machine translation (Watanabe and Sumita, 2002; Finch and Sumita, 2009; Zhang et al., 2013). They used the techniques for traditional SMT models, under the IBM framework (Watanabe and Sumita, 2002) or the feature-driven linear models (Finch and Sumita, 2009; Zhang et al., 2013). However, the target-bidirectional techniques

[6] In training SMT (Liu and Huang, 2014), we update weights towards higher BLEU translations and thus we care more about the relative statistics of BLEU; but in this paper, we care more about the absolute statistics, in order to show how severe the problem of unbalanced outputs is.

[7] We did not run NMT-l2r-10 and NMT-r2l-10, because it is too time-consuming to train 10 NMT models on both target directions and especially NMT-r2l-10 is not necessarily better than NMT-r2l-5 as shown in Table 2.

we have developed for the unified neural network framework, target a pressing need directly motivated by a fundamental issue suffered by recurrent neural networks.

Target-directional neural network models have also been successfully employed in (Devlin et al., 2014). However, their approach was concerned with feedforward networks, which can not make full use of rich contextual information. As a result, their models could only be used as features (i.e. submodels) to augment traditional translation techniques in contrast to the end-to-end neural network framework for machine translation in our proposal.

Our approach is related to that in (Bengio et al., 2015) in some sense. Both approaches can alleviate the mismatch between the training and testing stages: the history predictions are always correct in training while may be incorrect in testing. Bengio et al. (2015) introduce noise into history predictions in training to balance the mistmatch, while we try to make the history predictions in testing as accurate as those in training by using of two directional models. Therefore, theirs focuses on this problem from the view of training instead of both modeling and training as ours, but it is possible and promising to apply their approach to optimize our joint model.

6 Conclusion

In this paper, we investigate the issue of unbalanced outputs suffered by recurrent neural networks, and empirically show its existence in the context of machine translation. To address this issue, we propose an easy to implement agreement model that extends the method of (Liu et al., 2016) from simple sequence-to-sequence learning tasks to machine translation.

On two challenging JP-EN and CH-EN translation tasks, our approach was empirically shown to be effective in addressing the issue; by generating balanced outputs, it was able to consistently outperform a respectable NMT baseline on all test sets, delivering gains of up to 1.4 BLEU points. To put these results in the broader context of machine translation research, our approach (even without special handling of unknown words) achieved gains of up to 5.6 BLEU points over strong phrase-based and hierarchical phrase-based Moses baselines, with the help of an ensemble technique.

Acknowledgments

We would like to thank the three anonymous reviewers for helpful comments and suggestions. In addition, we would like to thank Rico Sennrich for fruitful discussions.

References

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

[Bengio et al.2015] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

[Cheng et al.2016] Yong Cheng, Shiqi Shen, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Agreement-based joint training for bidirectional attention-based neural machine translation. *CoRR*, abs/1512.04650.

[Devlin et al.2014] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of ACL*.

[Dyer et al.2015] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of ACL-IJCNLP*.

[Finch and Sumita2009] Andrew Finch and Eiichiro Sumita. 2009. Bidirectional phrase-based statistical machine translation. In *Proceedings of EMNLP*.

[Goto et al.2011] Isao Goto, Bin Lu, Ka-Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR-9*.

[Goto et al.2013] Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of ACL*.

[Graves2013] Alex Graves. 2013. Generating sequences with recurrent neural networks. *CoRR*.

[Jean et al.2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of ACL-IJCNLP*.

[Koehn et al.2007] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi,

B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of ACL: Demonstrations*.

[Liang et al.2006] Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*.

[Liu and Huang2014] Lemao Liu and Liang Huang. 2014. Search-aware tuning for machine translation. In *Proceedings of EMNLP*.

[Liu et al.2016] Lemao Liu, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2016. Agreement on target-bidirectional lstms for sequence-to-sequence learning. In *Proceedings of AAAI*.

[Luong et al.2015] Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of ACL-IJCNLP*.

[Och and Ney2000] Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*, pages 440–447.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.

[Watanabe and Sumita2002] Taro Watanabe and Eiichiro Sumita. 2002. Bidirectional decoding for statistical machine translation. In *Proceeding of COLING*.

[Watanabe and Sumita2015] Taro Watanabe and Eiichiro Sumita. 2015. Transition-based neural constituent parsing. In *Proceedings of ACL-IJCNLP*.

[Zeiler2012] Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*.

[Zhang et al.2013] Hui Zhang, Kristina Toutanova, Chris Quirk, and Jianfeng Gao. 2013. Beyond left-to-right: Multiple decomposition structures for smt. In *HLT-NAACL*, pages 12–21.

[Zhou et al.2002] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. 2002. Ensembling neural networks: Many could be better than all. *Artif. Intell.*