A Survey of Advances and Challenges in Unsupervised Neural Machine Translation

Rui Wang, Haipeng Sun, Masao Utiyama, and Eiichiro Sumita
National Institute of Information and Communications Technology
{wangrui, sun.haipeng, mutiyama, eiichiro.sumita}@nict.go.jp

1 Introduction

Unsupervised cross-lingual language representation initialization methods, together with mechanisms such as denoising and back-translation, have advanced unsupervised neural machine translation (UNMT), which has achieved impressive results. Meanwhile, there are still challenges for UNMT. This paper first introduces the background and the latest progress of UNMT. We then examine a number of challenges to UNMT and analyze the future trend of UNMT.

2 Advances

Recently, neural machine translation (NMT) has been adapted to the unsupervised scenario where NMT is trained without any bilingual data. Unsupervised NMT (UNMT) [1, 2] only requires monolingual corpora, using a combination of diverse mechanisms such as an initialization with bilingual word embeddings, denoising auto-encoder [3], back-translation [4], and shared latent representation. A shared encoder was used to encode the source sentences and decode them from a shared latent space [1, 2]. The difference is that [2] used a single shared decoder and [1] leveraged two independent decoders for each language. [5] used two independent encoders for each language with a weight-sharing mechanism to overcome the weakness of retaining the uniqueness and internal characteristics of each language. [6] achieved remarkable results in several similar languages such as English-French by concatenating two bilingual corpora as one monolingual corpus and using monolingual embedding pre-training in the initialization step.

Recent adavances: [7] utilized auxiliary languages to boost UNMT model. [8] proposed an extract-edit approach, to extract and then edit real sentences from the target monolingual corpora instead of back-translation. [9] proposed to train UNMT with bilingual word embedding agreement. [10] introduced unsupervised pivot translation for distant language pairs. More recently, [11, 12] introduced the pre-trained cross-lingual language model to achieve better UNMT performance.

USMT and UNMT: [6, 13] proposed an alternative method, that is, unsupervised statistical machine translation (USMT) method. [14, 15, 16] combined UNMT and USMT to improve unsupervised machine translation performance. In 2019, the unsupervised MT task (German-Czech) first-time became the officially task of WMT-2019, and the system from NICT [17] won the first place and achieved state-of-the-art performances by combining the USMT and UNMT.

Supervised and unsupervised NMT: [17] also explored the relationship between supervised and unsupervised NMT. First, they found that the pseudo-parallel data generated by unsupervised MT can be directly used as training data to train a pseudo-supervised NMT system with significant improvement. In addition, they found that NICT's system which achieved comparable performance with online (supervised) systems [18], if it was fine-tuned by several thousands of parallel sentences.

We summarize the performances of recent UNMT works in Table 1.

表 1: Performance (BLEU score) of recent UNMT works.

Method	Fr-En	En-Fr	De-En	En-De	Ja-En	En-Ja
Artetxe *et al.* [1]	15.56	15.13	n/a	n/a	n/a	n/a
Lample *et al.* [2]	14.31	15.05	13.33	9.64	n/a	n/a
Yang *et al.* [5]	15.58	16.97	14.62	10.86	n/a	n/a
Lample *et al.* [6]	24.20	25.10	21.00	17.20	n/a	n/a
Sun *et al.* [9]	25.87	28.38	22.67	18.29	17.22	23.64
Lample *et al.* [11]	33.4	33.3	26.4	34.3	n/a	n/a

3 Methods

State-of-the-art UNMT [6] can be decomposed into four components: bilingual word embedding initialization, denoising auto-encoder, back-translation, and shared latent representations.

Bilingual word embedding initialization: Compared with supervised neural machine translation, there is no bilingual data for training. Instead, the pre-trained unsupervised BWE provides a naive translation knowledge to enable the back-translation to generate pseudo-parallel corpora at the beginning of the training.

Denoising auto-encoder: Noise obtained by randomly performing local substitutions and word reorderings [19, 20], is added to the input sentences to improve model learning ability and regularization. Consequently, the input data are continuously modified and different at each epoch. The denoising auto-encoder model objective function would be optimized by maximizing the probability of encoding a *noisy* sentence and reconstructing it.

Back-translation: The back-translation plays a key role in achieving unsupervised translation relying only on monolingual corpora in each language. The pseudo-parallel sentence pairs produced by the model at the previous iteration have been used to train the new translation model.

Sharing latent representations: Encoders and decoders are (partially) shared for two languages. Therefore, the two languages must use the same vocabulary. The entire training of UNMT needs to consider back-translation between the two languages and their respective denoising processing.

4 Challenges

Effect on specific language: Most existing works focus on modeling UNMT systems and few works have investigated the effect of UNMT in specific languages. The performances of UNMT in similar language pairs (French/German-English) are dramatically better than that in distant language pairs (Chinese/Japanese-English) translation tasks. The UNMT performance in the distant language pairs needs to be solved.

One of the syntactic structures of languages are different. Without parallel supervision, it is very difficult for UNMT to learn the syntactic correspondence. Syntax information has been shown that it can improve the performances of NMT [21, 22, 23, 24, 25, 26, 27, 28].

Domain adaptation methods for UNMT have not been well-studied although UNMT has recently achieved remarkable results in some specific domains for several language pairs. For UNMT, addition to inconsistent domains between training data and test data for supervised NMT, there also exist other inconsistent domains between monolingual training data in two languages. Actually, it is difficult for some language pairs to obtain enough source and target monolingual corpora from the same domain in the real-world scenario.

[29] empirically show different scenarios for unsupervised domain-specific neural machine translation. Based on these scenarios, they show and analyze several potential solutions including batch weighting, data selection, and fine tuning methods, to improve the performances of domain-specific UNMT systems [29].

Efficiency: Compared with NMT, the training time of UNMT increased rapidly. In addition, learning sharing latent representations ties the performance of both translation directions, especially for distant language pairs, while denoising dramatically delays convergence by continuously modifying the training data. Efficient training of UNMT is also an issue that needs to be solved.

5 Conclusion

In this paper, we first survey the the background and the latest progress of UNMT. We then examine a number of challenges to UNMT and and analyze the future trend of UNMT.

参考文献

[1] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *ICLR*, 2018.

[2] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," in *ICLR*, 2018.

[3] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[4] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *ACL*, 2016.

[5] Z. Yang, W. Chen, F. Wang, and B. Xu, "Unsupervised neural machine translation with weight sharing," in *ACL*, 2018.

[6] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," in *EMNLP*, 2018.

[7] C. Xu, T. Qin, G. Wang, and T. Liu, "Polygon-net: A general framework for jointly boosting multiple unsupervised neural machine translation models," in *IJCAI*, 2019.

[8] J. Wu, X. Wang, and W. Y. Wang, "Extract and edit: An alternative to back-translation for unsupervised neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, Jun. 2019, pp. 1173–1183.

[9] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, and T. Zhao, "Unsupervised bilingual word embedding agreement for unsupervised neural machine translation," in *ACL*, Florence, Italy, Jul. 2019, pp. 1235–1245.

[10] Y. Leng, X. Tan, T. Qin, X.-Y. Li, and T.-Y. Liu, "Unsupervised pivot translation for distant languages," in *ACL*, 2019.

[11] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019.

[12] K. Song, X. Tan, T. Qin, J. Lu, and T. Liu, "MASS: masked sequence to sequence pre-training for language generation," in *ICML*, 2019.

[13] M. Artetxe, G. Labaka, and E. Agirre, "Unsupervised statistical machine translation," in *EMNLP*, 2018.

[14] B. Marie and A. Fujita, "Unsupervised neural machine translation initialized by unsupervised statistical machine translation," *CoRR*, vol. abs/1810.12703, 2018.

[15] S. Ren, Z. Zhang, S. Liu, M. Zhou, and S. Ma, "Unsupervised neural machine translation with SMT as posterior regularization," in *AAAI*, 2019.

[16] M. Artetxe, G. Labaka, and E. Agirre, “An effective approach to unsupervised machine translation,” in *ACL*, Jul. 2019.

[17] B. Marie, H. Sun, R. Wang, K. Chen, A. Fujita, M. Utiyama, and E. Sumita, “NICT’s unsupervised neural and statistical machine translation systems for the WMT19 news translation task,” in *WMT*, 2019.

[18] L. Barrault, O. Bojar, M. R. Costa-jussÃ, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. MÃller, S. Pal, M. Post, and M. Zampieri, “Findings of the 2019 conference on machine translation (wmt19),” in *ACL*, 2019.

[19] F. Hill, K. Cho, and A. Korhonen, “Learning distributed representations of sentences from unlabelled data,” in *NAACL*, 2016.

[20] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W. Ma, “Dual learning for machine translation,” in *NIPS*, 2016.

[21] A. Eriguchi, K. Hashimoto, and Y. Tsuruoka, “Tree-to-sequence attentional neural machine translation,” in *ACL*, 2016.

[22] A. Eriguchi, Y. Tsuruoka, and K. Cho, “Learning to parse and translate improves neural machine translation,” in *ACL*, 2017.

[23] H. Chen, S. Huang, D. Chiang, and J. Chen, “Improved neural machine translation with a syntax-aware encoder and decoder,” in *ACL*, 2017.

[24] J. Li, D. Xiong, Z. Tu, M. Zhu, M. Zhang, and G. Zhou, “Modeling source syntax for neural machine translation,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 688–697. [Online]. Available: http://aclweb.org/anthology/P17-1064

[25] S. Wu, D. Zhang, N. Yang, M. Li, and M. Zhou, “Sequence-to-dependency neural machine translation,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 698–707. [Online]. Available: http://aclweb.org/anthology/P17-1065

[26] K. Chen, R. Wang, M. Utiyama, L. Liu, A. Tamura, E. Sumita, and T. Zhao, “Neural machine translation with source dependency representation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017, pp. 2846–2852. [Online]. Available: https://www.aclweb.org/anthology/D17-1304

[27] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, “Syntax-directed attention for neural machine translation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, New Olreans, LA, 2018, pp. 4792–4799. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16060/16008

[28] C. Ma, A. Tamura, M. Utiyama, E. Sumita, and T. Zhao, “Improving neural machine translation with neural syntactic distance,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2032–2037. [Online]. Available: https://www.aclweb.org/anthology/N19-1205

[29] H. Sun, R. Wang, K. Chen, M. Utiyama, E. Sumita, and T. Zhao, “An empirical study of domain adaptation for unsupervised neural machine translation,” *CoRR*, vol. abs/1908.09605, 2019.