Advances and Challenges in Unsupervised Neural Machine Translation

Rui Wang and Hai Zhao
Department of Computer Science and Engineering, Shanghai Jiao Tong University
Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
MoE Key Lab of Artificial Intelligence, AI Institute,
Shanghai Jiao Tong University, Shanghai, China
wangrui.nlp@gmail.com and zhaohai@cs.sjtu.edu.cn

Abstract

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 several challenges for UNMT. This tutorial first introduces the background and the latest progress of UNMT. We then examine a number of challenges to UNMT and give empirical results on how well the technology currently holds up.

1 Tutorial Content

1.1 Introduction

Machine translation (MT) is a classic topic in the NLP community. Since 2010s, deep learning methods have been adopted in neural MT (NMT) and NMT has achieved promising performances (Bahdanau et al., 2015). Recently, NMT has been adapted to the unsupervised scenario. Unsupervised NMT (UNMT) (Artetxe et al., 2018b; Lample et al., 2018a) only requires monolingual corpora, using a combination of diverse mechanisms such as an initialization with bilingual word embeddings, denoising auto-encoder, back-translation, and shared latent representation.

1.2 Methods

Cross-lingual language representation initialization. In supervised NMT, language representation initialization is not so necessary, because the bilingual corpus can help NMT learn the cross-lingual representation. In comparison, there is only monolingual corpus for UNMT. Therefore, the pre-trained unsupervised bilingual word embedding (Artetxe et al., 2017; Lample et al., 2018b) or unsupervised cross-lingual language model (Lample and Conneau, 2019) provide a naive translation knowledge to enable the back-translation to generate pseudo-parallel corpora at the beginning of the UNMT training.

Denoising auto-encoder: Noise obtained by randomly performing local substitutions and word reorderings (Vincent et al., 2010), is added to the input sentences to improve model learning ability and regularization. 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 (Sennrich et al., 2016). 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.

1.3 Recent Advances

USMT and UNMT. Since 2016, statistical MT (SMT) has been significantly over-passed by NMT. Lample et al. (2018c) and Artetxe et al. (2018a) proposed an alternative method, that is, unsupervised statistical machine translation (USMT) method. However, in the supervised scenario, the performance of USMT method is comparable with that of UNMT. In addition, several works (Marie and Fujita, 2018; Ren et al., 2019; Artetxe et al., 2019) combined UNMT and USMT to improve unsupervised machine translation performance. In WMT-2019, the unsupervised MT task (German-Czech) first-time became the officially task of WMT, and the system from NICT (Marie et al., 2019) won

the first place and achieved state-of-the-art performances by combining the USMT and UNMT. However, after the advanced pre-training technologies was developed, USMT became less important.

Advanced Pre-Training Technologies. Similar as other NLP tasks, the quality of language representation pre-training significantly affects the performance of UNMT. Several works focus on improving the language representation pre-training. Sun et al. (2019b) proposed to train UNMT jointly with bilingual word embedding agreement. More recently, it has been shown that the pre-trained cross-lingual language model (Lample and Conneau, 2019; Song et al., 2019) achieve better UNMT performance than the bilingual word embedding. In high-resource scenario, UNMT has achieved remarkable performance. However, the performance of low-resource UNMT is still far below expectations

Multilingualism. To improve the low-resource UNMT, multi-lingual UNMT (MUNMT) is proposed (Sun et al., 2020; Liu et al., 2020). The translation of low-resouce and zero-shot language pairs can be enhanced by the similar languages in the shared latent representation. In addition, the pivot-based methods are proposed. Leng et al. (2019) introduced unsupervised pivot translation for distant language pairs. The SJTU-NICT team used monolingual corpus together with parallel third-party languages to enhance the low-resource UNMT performance (Li et al., 2020b) and their system achieved the best performance in WMT-2020 unsupervised task (Li et al., 2020a).

1.4 Challenges

Most existing works focus on modeling UNMT systems and few works investigate the reason why UNMT works and the scenario where UNMT works. UNMT still has limit performance in the distant language pair and domain-specific scenarios.

Distant Language Pairs. we will first empirically show that the performances of UNMT in distant language pairs (Chinese/Japanese-English) are much worse than the similar language pairs (German/French-English). Then, we will show the hypotheses: 1) syntactic structures of distant language pairs are quit different. Without parallel supervision, it is very difficult for UNMT to learn the syntactic correspondence. 2) There are too few shared words/subwords in the distant language pair to learn the shared latent representation for UNMT. Finally, we will show some potential solutions, such as 1) syntactic methods (Eriguchi et al., 2016; Chen et al., 2017, 2018) and 2) artificial shared words/code-switching methods (Yang et al., 2020) and show the initial results.

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.

In this tutorial, we will empirically show different scenarios for unsupervised domain-specific neural machine translation. Based on these scenarios, we will show and analyze several potential solutions including batch weighting, data selection, and fine tuning methods, to improve the performances of domain-specific UNMT systems (Sun et al., 2019a).

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.

2 Relevance to the Computational Linguistics Community

This tutorial makes an attempt to review the latest progress on UNMT by introducing advances and challenges for UNMT. MT is a classic topic in the NLP community. Recently, UNMT has attracted great interest in the researchers in both the MT/NLP community and industry.

This tutorial is primarily towards researchers who have a basic understanding of deep learning based NLP. We believe that this tutorial would help the audience more deeply understand UNMT.

3 Type of the Tutorial: Cutting-edge

We introduce the cutting-edge technologies. This tutorial is primarily towards researchers who have a basic understanding of deep learning based NLP,

Presenter: Rui Wang		Presenter: Hai Zhao	
1. Introduction of MT (30 min)	2. Methods for UNMT (70 min)	3. Challenges in UNMT (60 min)	4. Summary (20 min)
1.1 Statistical MT (SMT)	2.1 USMT and UNMT	3.1 Distant Language Pairs	4.1 Conclusion
1.2 Neural MT (NMT)	2.2 Advanced Pre-Training Technologies	3.2 Domain Adaptation	4.2 Future Trends
	2.3 Multilingualism	3.3 Training Efficiency	
– Coffee Break – (30 min)			

Table 1: Tutorial outlines

and it is supposed to widen and deepen the understanding of cutting-edge NLP for the audience.

4 Tutorial Outlines

We will present our tutorial in three hours. The detailed tutorial outlines are shown in Table 1.

5 Specification of Any Prerequisites for the Attendees

This tutorial is primarily aimed at researchers who have a basic understanding of NLP and deep learning.

6 Small reading list

- Neural Machine Translation: the basic method "*Neural machine translation by jointly learning to align and translate*" (Bahdanau et al., 2015) and the related deep learning backgrounds "*Deep learning*" (LeCun et al., 2015).
- UNMT: the basic methods "*Unsupervised neural machine translation*" (Artetxe et al., 2018b) and "*Unsupervised machine translation using monolingual corpora only*" (Lample et al., 2018a). State-of-the-art UNMT systems (Marie et al., 2019; Li et al., 2020a).

7 Presenters

1. Dr. Rui Wang, Tenured Researcher, Advanced Translation Technology Laboratory, National Institute of Information and Communications Technology (NICT), Japan

wangrui.nlp@gmail.com

https://wangruinlp.github.io

His research focuses on machine translation (MT), a classic task in NLP. His recent interests are traditional linguistic based and cutting-edge machine learning based approaches for MT. He (as the first or the corresponding authors) has published more than 30 MT papers in top-tier NLP/ML/AI conferences and journals, such as ACL, EMNLP, ICLR, AAAI, IJCAI, IEEE/ACM transactions, etc. He has also won several first places in top-tier MT shared tasks, such as WMT-2018, WMT-2019, WMT-2020, etc.

He has given several tutorial and invited talks in conferences, such as CWMT, CCL, etc. He served as the area chairs of ICLR-2021 and NAACL-2021.

2. Dr. Hai Zhao, Professor, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China.

zhaohai@cs.sjtu.edu.cn

http://bcmi.sjtu.edu.cn/~zhaohai

His research interest is natural language processing. He has published more than 120 papers in ACL, EMNLP, COLING, ICLR, AAAI, IJCAI, and IEEE TKDE/TASLP. He won the first places in several NLP shared tasks, such as CoNLL and SIGHAN Bakeoff and top ranking in remarkable machine reading comprehension task leaderboards such as SQuAD2.0 and RACE.

He has taught the course "natural language processing" in SJTU for more than 10 years. He is ACL-2017 area chair on parsing, and ACL-2018/2019 (senior) area chairs on morphology and word segmentation.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Unsupervised statistical machine translation.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, San Diego, CA.

Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, Akihiro Tamura, Eiichiro Sumita, and Tiejun Zhao. 2017. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2852, Copenhagen, Denmark.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2018. Syntax-directed attention for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4792–4799, New Olreans, LA.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-sequence attentional neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 823–833, Berlin, Germany.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, Vancouver, Canada.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*, Vancouver, Canada.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183, Florence, Italy.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020a. SJTU-NICT's supervised and unsupervised neural machine translation systems for the WMT20 news translation task. *arXiv preprint arXiv:2010.05122*.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020b. Reference language based unsupervised neural machine translation. In *The 2020 Conference on Empirical Methods in Natural Language Processing: ACL Findings*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Benjamin Marie and Atsushi Fujita. 2018. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy. Association for Computational Linguistics.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised neural machine translation with SMT as posterior regularization. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 241–248, Honolulu, Hawaii, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936, Long Beach, California, USA.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019a. An empirical study of domain adaptation for unsupervised neural machine translation. *CoRR*, abs/1908.09605.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019b. Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *ACL*, pages 1235–1245, Florence, Italy.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Code-switching pre-training for neural machine translation. *arXiv: 2009.08088*.