Unsupervised Machine Translation Demystified

Abstract

Recently, machine translation approaches which do not use any parallel corpus to train the machine translation models have been proposed in the literature. These approaches use only monolingual corpus available in the involved languages. The central theme of all these approaches is to learn a) language-agnostic representations and b) use Back-Translation (BT) mechanism to train the models in an unsupervised way. Several approaches have been proposed in the literature surrounding these two key steps. The focus of this tutorial is to cover the breadth of the literature on recent advances in Unsupervised Machine Translation. The tutorial will help the audience in getting started with unsupervised machine translation. The tutorial will span over three sections. In the first section, we will cover the fundamental concepts like cross-lingual embeddings, denoising auto-encoders, language model pre-training, Back Translation (BT), etc which are key to the success of Unsupervised Machine Translation. In the second section, the tutorial will provide a brief summary of recent works on unsupervised machine translation. The tutorial will cover both Phrase-Based Statistical Machine Translation systems as well as Neural Machine Translation systems. In the last section, we will talk about the limitations of the existing approaches for Unsupervised machine translation approaches and provide general guidelines for successful training of these systems. We also discuss case-studies from Indian languages and provide results obtained with U-MT over Indian language pairs. Finally, we talk about possible research directions.

1 Outline

We propose for a half-day tutorial on Unsupervised Machine Translation. The tutorial will cover the concept from the foundations of Unsupervised Machine Translation (U-MT) to recent advancements in the field. We will divide the tutorial into four parts. We shall begin the tutorial by explaining Machine Translation terminology, and a brief history of previous paradigms in the area. The main content of this tutorial starts with the establishment of the fundamentals of Unsupervised Machine Translation (U-MT). We, then, move towards discussing current research interests and trends in the area of U-MT.

In the first part, we will start with an overview of the Encode-Attend-Decode architecture of NMT, which is employed by most of the U-MT approaches, followed by explaining some basic concepts of MT which are the key components of U-MT. This includes a detailed discussion on Cross-lingual word embeddings, Denoising Autoencoder, Back-Translation, and Language Model pre-training techniques. We shall spend a significant amount of time to establish this groundwork required to understand the nitty-gritties of U-MT.

Once the audience is comfortable with the building blocks of U-MT, in the second part, we will introduce them with various U-MT approaches and the effect of each component on these methods. Here, we will describe U-NMT, U-SMT, hybrid U-MT approaches. We will, then, explain the latest state-of-the-art methods using Language Model pre-training.

Later, we will go in-depth and discuss some aspects which affect output quality, i.e., language relatedness, domain mismatch, along with case-studies from Indic languages. We will spend time on how language relatedness between source and target languages plays an important role in translation quality. We will also show that even if we get to achieve promising results without any parallel data, a hidden condition of source and target monolingual data to be of the similar domain is still present which is an important question for NLP researchers especially working with Indic languages.

We will conclude the session with summarisation and possible directions that we believe U-MT research should take.

2 Proposers

Rudra Murthy V is a research scientist at IBM since May 2020. He completed his PhD at IIT Bombay under the guidance of Prof. Pushpak Bhattacharyya. Rudra Murthy has published in key AI/NLP conferences such as NAACL, ACL, and TALLIP journal. He has worked in the areas of Machine Translation and Named Entity Recognition with focus on Indic languages.

Tamali Banerjee is a research scholar at IIT Bombay since January 2016. Her main research area is on Machine Translation for Indic languages. One of her works won the best paper award in the Second Workshop on Subword/Character Level Models (SCLeM 2018).

Jyotsana Khatri is a research scholar at IIT Bombay since July 2017. She is working on cross-lingual representation learning and machine translation with focus on Indian languages. Jyotsana has published in key AI/NLP conferences such as COLING, WMT etc.

Diptesh Kanojia is a PhD candidate at the IITB-Monash Research Academy, a joint PhD degree offered by the IIT Bombay, India, and Monash University, Australia. He is pursuing his PhD under the guidance of Prof. Pushpak Bhattacharyya, Prof. Gholamreza Haffari, and Prof. Malhar Kulkarni. Diptesh has published in key AI/NLP conferences such as AAAI, IJCAI, ACL, CoNLL, COLING, NAACL, LREC etc. These publications belong to various sub-fora of NLP, including but not limited to, cognate detection, machine translation, cognitive psycholinguistics, word sense disambiguation and computational phylogenetics; currently foraying into semantics and its applications to NLP. He has also reviewed papers for various conferences and journals like COLING, *ACL, EMNLP, AAAI, CoNLL, GWC, LREC, and TALLIP.

Dr. Pushpak Bhattacharyya has made seminal contributions in Natural Language Processing (NLP) and Machine Learning (ML), working in these fields for the last 30 years. In addition to being a professor at Dept. of CSE in IIT Bombay, he has been a visiting Professor at Stanford University (2004), University Joseph Fourier (2005, 2009, 2011, 2014), Distinguished lecturer at the University of Houston, USA (2012), and Visiting Scholar- MIT (1990). Dr. Bhattacharyya has carried out sponsored research projects on various NLP technologies, funded by the Ministry of Communication and Information Technology, Ministry of Human Resource Development and by IBM, Microsoft Research, Yahoo, HP Labs, and NEC. He has written a textbook called “Machine Translation”, published by CRC Press, USA (Taylor and Francis group) which brings to light foundational points pertaining to the translation of Indian languages. Google Scholar Citation which is the benchmark for Computer Science and Engineering shows approx.. 8000 career citations to Dr. Bhattacharyya’s papers with an h-index of 42.

References

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, EMNLP 2018*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised Neural Machine Translation with SMT as Posterior Regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *36th International Conference on Machine Learning, ICML 2019*.