Neural Machine Translation (NMT): Literature Review

June 18, 2024

Abstract

This literature review provides a comprehensive overview of Neural Machine Translation (NMT), covering advancements, challenges, and solutions in various areas including NMT models, unsupervised multilingual NMT, NMT frameworks, low-resource NMT models, open-source NMT toolkits, graph-based NMT, and training NMT models specifically for English. Key findings highlight improvements in translation accuracy, robustness, and adaptability, with significant progress in handling low-resource languages and unsupervised translation tasks. The review concludes with discussions on future research directions aimed at further enhancing the capabilities and reliability of NMT systems.

1 Introduction

Neural Machine Translation (NMT) has revolutionized the field of machine translation with significant advancements in translation accuracy and fluency. This literature review aims to summarize the current state of research in NMT, identify key challenges, and explore different methodologies and frameworks that have been proposed to address these issues. The review covers various aspects of NMT, including specific models, multilingual and unsupervised translation, frameworks, low-resource contexts, open-source toolkits, graph-based improvements, and training strategies for English.

2 Background

NMT relies on deep learning techniques to translate text from one language to another by modeling translation as a sequence-to-sequence problem. This approach has largely replaced traditional statistical methods, providing improved performance in terms of fluency and accuracy. However, NMT faces challenges such as handling low-resource languages, translation inaccuracies for specialized domains, and maintaining robustness against input perturbations and noisy data.

3 Neural Machine Translation Models

NMT models have seen notable advancements but continue to struggle with issues like translation inaccuracies, particularly with named entities, due to limited training data. The "Extract and Attend" approach improves accuracy by integrating dictionary lookups [Zeng et al., 2023]. The contrastive marking objective refines training by weighting correct and incorrect tokens differently, addressing exploration issues in translation space [Berger et al., 2023].

NMT's sensitivity to input perturbations has led to methods like Pseudo Label Training (PLT) for greater model stability [Hsu et al., 2023]. Interpretability issues have been addressed by tracking input token attributions [Ferrando et al., 2022]. Quality-aware decoding approaches, such as minimum Bayes risk decoding, enhance inference accuracy [Fernandes et al., 2022].

Despite progress, issues like correct token alignment and managing noisy data persist. Techniques like Token Dropout help in preventing overfitting and improving generalization [Zhang et al., 2020]. Volatility in NMT models highlights the need for robust training methods [Fadaee and Monz, 2020, Cheng et al., 2018]. Various strategies are also being explored to enhance translation quality under specific lexical and structural constraints [Zhang et al., 2019, Wang et al., 2019].

4 Unsupervised Multilingual NMT

Unsupervised Neural Machine Translation (UNMT) aims to improve translation quality without large quantities of human-translated data. Methods like XConST enhance zero-shot performance in multilingual NMT [?]. Evaluations of UNMT across diverse languages using Layer-wise Relevance Propagation have shown promising results in semantic similarity [Tourni and Wijaya, 2023]. Challenges in translating low-resource languages, like Yorùbá to English, have been highlighted [Akinade et al., 2023], and methods like unsupervised pivot translation aim to maintain language-specific characteristics while sharing high-level representations [Yang et al., 2018].

Iterative back translation has shown to be effective in synthetic bilingual data generation [Marie et al., 2018], and adding an artificial token for target language indication simplifies and improves multilingual NMT [Johnson et al., 2017]. These studies collectively showcase advancements in making NMT models more effective in unsupervised and low-resource multilingual contexts.

5 Neural Machine Translation Frameworks

NMT frameworks typically use an encoder-decoder architecture. Recent developments include Self-Knowledge Distillation with bidirectional decoders for better regularization [?]. Knowledge distillation approaches have been employed to compress deep models without performance loss [Li et al., 2020]. Multi-pass decoding with the "Rewriter-Evaluator" architecture helps iteratively improve translation quality [Li et al., 2021].

Efforts to simplify NMT architectures, such as encoder-free models and Multi-Dimensional LSTM, demonstrate competitive performance [Tang et al., 2019, Bahar et al., 2018]. Transformer models continue to benefit from new optimizations that allow deeper architectures and improved BLEU scores [Bapna et al., 2018, Zhang et al., 2018]. Innovative training and decoding strategies further enhance the practical application of NMT systems [Devlin, 2017, Wang et al., 2017, Eriguchi et al., 2016].

6 Low-Resource NMT Models

Large language models (LLMs) pretrained on extensive datasets have shown promise in NLP tasks, including NMT. Simul-LLM, a framework for finetuning LLMs for simultaneous translation, represents an important advancement [Agostinelli et al., 2023]. In low-resource settings, strategies such as joint dropout and memory-augmented adapters enhance generalization and translation quality [Araabi et al., 2023, ?].

The development of the first Luganda-English NMT model demonstrated significant progress in low-resource NMT [Kimera et al., 2023]. Transfer learning and curriculum-based training have proven effective in improving performance for low-resource language pairs [Arivazhagan et al., 2019, Zoph et al., 2016]. Adapting NMT systems to tackle linguistic variations among dialects and non-native speakers remains a pressing challenge [?Raunak et al., 2020].

7 Open-Source NMT Toolkits

Open-source NMT toolkits such as OpenNMT, YANMTT, and NMT-Keras have greatly facilitated research by providing accessible frameworks for model development and enhancement [Klein et al., 2017, Dabre and Sumita, 2021, Álvaro Peris and Casacuberta, 2018]. VNMT's efficient use of the JIT format supports various translation tasks robustly [Quan et al., 2022]. These toolkits maintain competitive performance and are widely adopted in both academic and production environments.

8 Graph-Based NMT: Contextual Improvements

Graph-based NMT models facilitate more contextually consistent translations at the document level. Strategies like selective memory-augmented translation and data-adaptive context retrieval significantly improve translation quality across various benchmarks [Zhang et al., 2022, Zhang, 2021]. Representing documents as graphs and integrating them with Transformer architectures has shown substantial gains in translation performance [Xu et al., 2021].

Using context for resolving ambiguities and pronoun resolution in diverse domains has demonstrated task-specific advantages, although no universal architecture excels across all tasks [Huo et al., 2020, Fu et al., 2019, Wang et al., 2019]. Lightweight memory networks offer an effective way to adapt translations dynamically with minimal computational overhead [Tu et al., 2017].

9 Training NMT Models for English

General-domain NMT models often fail in specialized domains like e-commerce and legal documents due to unique terminologies. Methods such as the G2ST paradigm, incorporating self-contrastive semantic enhancement, improve domain-specific performance [?]. Handling linguistic variations across dialects and among non-native speakers requires tailored benchmarks and expert oversight [?Raunak et al., 2023]. Unified approaches for simultaneous translation of multiple tasks and solutions targeting robustness to input noise and adversarial attacks are critical for improving model trustworthiness [Liang et al., 2023, Weng et al., 2023].

Domain adaptation strategies, including back translation and curriculum-based training, continue to play a pivotal role in enhancing NMT systems' resilience and efficacy across various domains [Poncelas et al., 2019, Mohiuddin et al., 2022].

10 Discussion

The reviewed studies highlight significant advancements in NMT model robustness, accuracy, and adaptability across diverse languages and domains. Addressing challenges in low-resource and unsupervised contexts remains crucial for broadening NMT's applicability. Innovations in model architectures, training methods, and open-source toolkits collectively contribute to ongoing progress in the field.

While considerable progress has been made, issues such as translation inaccuracies, robustness against input perturbations, and domain-specific performance continue to present challenges. There is a growing need for methods that can effectively generalize across multiple languages and domains, incorporating both lexical and structural constraints into translation processes.

11 Conclusion

This literature review underscores the dynamic nature of NMT research, highlighting both achievements and persistent challenges. Future research should focus on refining learning algorithms, enhancing model robustness, and developing innovative strategies for low-resource languages and specialized domains. Continued improvement in NMT frameworks and open-source toolkits will facilitate further advancements, making NMT more reliable and widely applicable.

References

Victor Agostinelli, Max Wild, Matthew Raffel, Kazi Ahmed Asif Fuad, and Lizhong Chen. Simul-llm: A framework for exploring high-quality simultaneous translation with large language models, 2023. URL `https://arxiv.org/abs/2312.04691`.

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. ε kú ¡mask¿: Integrating yorùbá cultural greetings into machine translation, 2023. URL `https://arxiv.org/abs/2303.17972`.

Ali Araabi, Vlad Niculae, and Christof Monz. Joint dropout: Improving generalizability in low-resource neural machine translation through phrase pair variables, 2023. URL `https://arxiv.org/abs/2307.12835`.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. Massively multilingual neural machine translation in the wild: Findings and challenges, 2019. URL `https://arxiv.org/abs/1907.05019`.

Parnia Bahar, Christopher Brix, and Hermann Ney. Towards two-dimensional sequence to sequence model in neural machine translation, 2018. URL `https://arxiv.org/abs/1810.03975`.

Ankur Bapna, Mia Xu Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. Training deeper neural machine translation models with transparent attention, 2018. URL `https://arxiv.org/abs/1808.07561`.

Nathaniel Berger, Miriam Exel, Matthias Huck, and Stefan Riezler. Enhancing supervised learning with contrastive markings in neural machine translation training, 2023. URL `https://arxiv.org/abs/2307.08416`.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. Towards robust neural machine translation, 2018. URL `https://arxiv.org/abs/1805.06130`.

Raj Dabre and Eiichiro Sumita. Yanmtt: Yet another neural machine translation toolkit, 2021. URL https://arxiv.org/abs/2108.11126.

Jacob Devlin. Sharp models on dull hardware: Fast and accurate neural machine translation decoding on the cpu, 2017. URL https://arxiv.org/abs/1705.01991.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Tree-to-sequence attentional neural machine translation, 2016. URL https://arxiv.org/abs/1603.06075.

Marzieh Fadaee and Christof Monz. The unreasonable volatility of neural machine translation models, 2020. URL https://arxiv.org/abs/2005.12398.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. Quality-aware decoding for neural machine translation, 2022. URL https://arxiv.org/abs/2205.00978.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer, 2022. URL https://arxiv.org/abs/2205.11631.

Han Fu, Chenghao Liu, and Jianling Sun. Reference network for neural machine translation, 2019. URL https://arxiv.org/abs/1908.09920.

Benjamin Hsu, Anna Currey, Xing Niu, Maria Nădejde, and Georgiana Dinu. Pseudo-label training and model inertia in neural machine translation, 2023. URL https://arxiv.org/abs/2305.11808.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. Diving deep into context-aware neural machine translation, 2020. URL https://arxiv.org/abs/2010.09482.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation, 2017. URL https://arxiv.org/abs/1611.04558.

Richard Kimera, Daniela N. Rim, and Heeyoul Choi. Building a parallel corpus and training translation models between luganda and english, 2023. URL https://arxiv.org/abs/2301.02773.

Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Crego, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation, 2017. URL https://arxiv.org/abs/1709.03815.

Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. Learning light-weight translation models from deep transformer, 2020. URL https://arxiv.org/abs/2012.13866.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation, 2021. URL https://arxiv.org/abs/2105.14802.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. Unified model learning for various neural machine translation, 2023. URL https://arxiv.org/abs/2305.02777.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. Nict's neural and statistical machine translation systems for the wmt18 news translation task, 2018. URL https://arxiv.org/abs/1809.07037.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale, and Shafiq Joty. Data selection curriculum for neural machine translation, 2022. URL https://arxiv.org/abs/2203.13867.

Alberto Poncelas, Maja Popovic, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. Combining smt and nmt back-translated data for efficient nmt, 2019. URL https://arxiv.org/abs/1909.03750.

Nguyen Hoang Quan, Nguyen Thanh Dat, Nguyen Hoang Minh Cong, Nguyen Van Vinh, Ngo Thi Vinh, Nguyen Phuong Thai, and Tran Hong Viet. Vinmt: Neural machine translation toolkit, 2022. URL https://arxiv.org/abs/2112.15272.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. On long-tailed phenomena in neural machine translation, 2020. URL https://arxiv.org/abs/2010.04924.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. Leveraging gpt-4 for automatic translation post-editing, 2023. URL https://arxiv.org/abs/2305.14878.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. Understanding neural machine translation by simplification: The case of encoder-free models, 2019. URL https://arxiv.org/abs/1907.08158.

Isidora Chara Tourni and Derry Wijaya. An empirical study of unsupervised neural machine translation: analyzing nmt output, model's behavior and sentences' contribution, 2023. URL https://arxiv.org/abs/2312.12588.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to remember translation history with a continuous cache, 2017. URL https://arxiv.org/abs/1711.09367.

Mingxuan Wang, Zhengdong Lu, Jie Zhou, and Qun Liu. Deep neural machine translation with linear associative unit, 2017. URL https://arxiv.org/abs/1705.00861.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. Exploiting sentential context for neural machine translation, 2019. URL https://arxiv.org/abs/1906.01268.

Rongxiang Weng, Qiang Wang, Wensen Cheng, Changfeng Zhu, and Min Zhang. Towards reliable neural machine translation with consistency-aware meta-learning, 2023. URL https://arxiv.org/abs/2303.10966.

Mingzhou Xu, Liangyou Li, Derek. F. Wong, Qun Liu, and Lidia S. Chao. Document graph for neural machine translation, 2021. URL https://arxiv.org/abs/2012.03477.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Unsupervised neural machine translation with weight sharing, 2018. URL https://arxiv.org/abs/1804.09057.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Xu Tan, Tao Qin, and Tie yan Liu. Extract and attend: Improving entity translation in neural machine translation, 2023. URL https://arxiv.org/abs/2306.02242.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation, 2020. URL https://arxiv.org/abs/2004.11867.

Linlin Zhang. Context-adaptive document-level neural machine translation, 2021. URL https://arxiv.org/abs/2104.08259.

Wei Zhang, Youyuan Lin, Ruoran Ren, Xiaodong Wang, Zhenshuang Liang, and Zhen Huang. Language model-driven unsupervised neural machine translation, 2019. URL https://arxiv.org/abs/1911.03937.

Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. Asynchronous bidirectional decoding for neural machine translation, 2018. URL https://arxiv.org/abs/1801.05122.

Xu Zhang, Jian Yang, Haoyang Huang, Shuming Ma, Dongdong Zhang, Jinlong Li, and Furu Wei. Smdt: Selective memory-augmented neural document translation, 2022. URL https://arxiv.org/abs/2201.01631.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation, 2016. URL https://arxiv.org/abs/1604.02201.

Álvaro Peris and Francisco Casacuberta. Nmt-keras: a very flexible toolkit with a focus on interactive nmt and online learning, 2018. URL https://arxiv.org/abs/1807.03096.