Relevance-guided Neural Machine Translation

Anonymous EMNLP submission

Abstract

With the advent of the Transformer architecture, Neural Machine Translation (NMT) results have shown great improvement lately. However, results in low-resource conditions still lag behind in both bilingual and multilingual setups, due to the limited amount of available monolingual and/or parallel data; hence, the need for methods addressing data scarcity in an efficient, and explainable way, is eminent. We propose an explainability-based training approach for NMT, applied in Unsupervised and Supervised model training, for translation of three languages of varying resources, French, Gujarati, Kazakh, to and from English. Our results show our method can be promising, particularly when training in low-resource conditions, outperforming simple training baselines; though the improvement is marginal, it sets the ground for further exploration of the approach and the parameters, and its extension to other languages.

1 Introduction

Unsupervised Neural Machine Translation (UNMT) has seen remarkable progress in recent years, with a very large number of methods proposed aiming to NMT when parallel data are few or non-existent for certain Language Pairs (Artetxe et al., 2017; Lample et al., 2018; Conneau et al., 2017; Wang and Zhao, 2021; Lample and Conneau, 2019; Song et al., 2019; Liu et al., 2020; Marchisio et al., 2020; Kim et al., 2020; Lample et al., 2017; Artetxe et al., 2019; Garcia et al., 2020; Su et al., 2019; Nguyen et al., 2022). Training techniques such as Back-Translation (Sennrich et al., 2015) and Auto-Encoding have been widely studied, in order to efficiently train NMT models under those data scarcity conditions to obtain high quality translation results. However, there is little work in enhancing Neural Machine Translation models with utilizing explainability of the model in order to improve quality of the output. We propose a method, based on Layer-wise Relevance Propagation (LRP), which leverages the contribution of the input tokens to the output, to boost NMT performance. Our results show LRP may be beneficial during model training for NMT output improvement, particularly in low-resource conditions and for specific well defined model setups.

2 Related Work

Layer-wise Relevance Propagation (LRP)

LRP was introduced by Bach et al. (2015), measuring the contribution of the input components, or the neurons of a network, to the next layer's output. Due to its nature, it is directly applicable to layer-wise architectures, and we extend its usage to the Transformer architecture, measuring the relevance of source and target sentences' tokens to the NMT output during training.

Explanations & Explanation-guided training

Several previous works outline and summarize the findings of explainability and interpetability-related research in NLP (Belinkov et al., 2020; Sun et al., 2021b; Tenney et al., 2020; Madsen et al., 2021; Danilevsky et al., 2020; Qian et al., 2021). Weber et al. (2022) provide a systematic review of explainable AI methods employed in improving various properties of Machine Learning models, such as performance, convergence, robustness, reasoning, efficiency and equality. Of these, of particular interest, and the focus of our work, are those that along with measuring feature importance and distinguishing relevant from irrelevant features, are utilized to augment the intermediate learned features, and improve model performance or reasoning (Anders et al., 2022; Sun et al., 2021a; Zunino et al., 2021a; Fukui et al., 2019; Zhou et al., 2016; Mitsuhara et al., 2019; Schiller et al., 2019; Zunino et al., 2021b).

In our paper, we modify the approach of Sun et al. (2021a), that proposes a model-agnostic LRP-guided training method for few shot classification, to improve model generalization to new classes, extending it to a transformer-based masked language model for NMT. Every intermediate feature representation $\mathrm{f_p}$ is weighted (multiplied) by its relevance $\mathrm{R(f_p)}$, with respect to the feature processing output, normalized in [-1,1]. The model is then trained on a loss function taking into account both predictions and given by the following formula

$$L = \xi * L_{\mathrm{ce}}(y, p) + \lambda * L_{\mathrm{ce}}(y, p_{\mathrm{lrp}}) \tag{1}$$

where ξ, λ are positive scalars. In this way, the features more relevant to the prediction are emphasized, while the less relevant ones are downscaled. Other recent works utilize LRP for improving model performance in the medical domain (Sefcik and Benesova, 2021), mitigating the influence of language bias for image captioning models (Sun et al., 2022).

3 Method & Experiments

3.1 Model Translation Quality Evaluation

In our experiments we use a 6-layers 8-heads encoder-decoder transformer-based model, XLM (Lample and Conneau, 2019), following the training configurations and hyperparameters suggested by the authors. We use Byte Pair Encoding (Sennrich et al., 2016) to extract a 60k vocabulary, and have an embedding layer size of 1024, a dropout value and an attention layer dropout value of 0.1, and a sequence length of 256. We measure the quality of the Language Model (LM) with perplexity, and quality of the NMT output with BLEU (Papineni et al., 2002), both used as training stopping criteria, when there is no improvement over 10 epochs. All further parameter values are provided in the Appendix. We first pre-train a Language Model in each language with the MLM objective, which is then used to initialize the encoder and decoder of the NMT model. We further train a NMT model, using backtranslation (BT) and denoising auto-encoding (AE) with the monolingual data used for LM pretraining for UNMT, the Machine Translation (MT) objective for the Supervised NMT model, and BT+MT for the joint Unsupervised and Supervised approach.

3.2 Datasets

The languages we work with are English, French, Gujarati, Kazakh, and we're translating in all English-centric directions, English–French (En–Fr), French–English (Fr–En), English–Gujarati (En–Gu), Gujarati–English (Gu–En), English–Kazakh (En–Kk), Kazakh–English (Kk–En). For English and French, we use 5 million News Crawl 2007-2008 monolingual sentences for each language, and 23 million WMT14 parallel sentences. For Gujarati, we have 1.4 million sentences and for Kazakh we have 9.5M monolingual sentences, collected for both languages from Wikipedia, WMT 2018, 2019 and Leipzig Corpora (2016)[1]. As parallel data, we have 22k and 132k from the WMT 2019 News Translation Task[2] for Gu–En and Kk–En.As development and test sets, we use newstest2013 and newstest2014, respectively, for En–Fr and Fr–En, WMT19 for En–Gu and Gu–En and En–Kk and Kk–En.

3.3 Layer-wise Relevance Propagation (LRP)

We follow the method proposed by Voita et al. (2020) in calculating LRP in an encoder-decoder Transformer architecture. The Relevance Score is first propagated inversely through the decoder and then the encoder, up to the input layer of the architecture. The conservation principle only holds across all processed tokens, and the score is defined as relevance of the input neurons to the top-1 logit predicted by the Transformer model, and the sum of the input neurons' relevance is regarded as the token contribution. The total source and target sentence contributions to the result are defined as the summation of the Relevance of tokens in the source sentence, x and that of those in the target sentence, y, at generation step t.

$$R_t(source) = \sum_i x_i \tag{2}$$

$$R_t(target) = \sum_{j=1}^{t-1} y_j \tag{3}$$

At every step t, Relevance of the source and target sentences follow the conservation principle, summing up to 1. Moreover, for every target token past the currently generated one, Relevance Score is 0.

[1] https://wortschatz.unileipzig.de/en/download/
[2] http://data.statmt.org/news-crawl/

	En–Fr				Fr–En			
Parameters' Set Values	v1	v2	v3	Regular	v1	v2	v3	Regular
22k								
MT	29.9	24.67	30.24	31.12	30.12	26.79	**30.52**	30.63
BT+AE+MT	30.01	28.62	**32.15**	34.54	29.94	29.87	**33.03**	34.02
1m								
MT	**39.12**	38.12	35.44	41.25	**39.2**	38.43	36.04	41.33
BT+AE+MT	**37.58**	37.22	36.89	40.37	37.66	**37.95**	37.38	40.4
2.5m								
MT	**39.06**	36.02	35.57	40.46	**39.11**	36.28	36.09	40.71
BT+AE+MT	37.68	36.21	**37.74**	39.88	40.71	37.48	36.66	**38.15**
5m								
MT	**39.21**	37.55	39	41.52	**39.33**	38.01	39.2	41.18
BT+AE+MT	30.71	36.7	**37.48**	40.89	32.02	37.03	**37.67**	40.8
Other methods				45.9				-

Table 1: BLEU scores for Supervised, and Unsupervised + Supervised NMT Layerwise Relevance Propagation-guided experiments, for En–Fr, Fr–En. *AE*, *BT* and *MT* stand for Auto-Encoding loss, Back Translation loss and Machine Translation loss, respectively. Test and validation sets are from newstest2013-14 for French. State of the art results (*Other methods*) for En–Fr come from `http://www.deepl.com/press.html`, `http://nlpprogress.com/english/machine_translation.html`.

	en–gu				gu–en			
Parameters' Set Values	v1	v2	v3	Regular	v1	v2	v3	Regular
22k								
MT	2	2.18	**2.19**	1.04	0.69	0.7	**0.73**	2.65
BT+AE+MT	0.76	**1.36**	0.89	1.16	0.71	**1.06**	0.67	2.19
Other methods				0.1				0.3

Table 2: BLEU scores for Supervised, and Unsupervised + Supervised NMT Layerwise Relevance Propagation-guided experiments, for En–Gu, Gu–En. *AE*, *BT* and *MT* stand for Auto-Encoding loss, Back Translation loss and Machine Translation loss, respectively. Test and validation sets are from WMT19 for Gujarati. State of the art results (*Other methods*) can be found in `https://github.com/google-research/bert/blob/master/multilingual.md`.

	en–kk				kk–en			
Parameters' Set Values	v1	v2	v3	Regular	v1	v2	v3	Regular
22k								
MT	2.1	3	**3.2**	2.4	2.2	2.1	**2.7**	2.6
BT+AE+MT	1.6	**3**	2.3	2.8	2.1	**2.6**	2.	2.9
132k								
MT	4.8	5.6	5.3	5.2	6.8	**8.5**	8.4	8
BT+AE+MT	5.2	**6.8**	6.4	6.6	8.7	**9.4**	9.2	8.9
Other methods				2.5[3]				7.4

Table 3: BLEU scores for Supervised, and Unsupervised + Supervised NMT Layerwise Relevance Propagation-guided experiments, for En–Kk, Kk–En. *AE*, *BT* and *MT* stand for Auto-Encoding loss, Back Translation loss and Machine Translation loss, respectively. Test and validation sets are from WMT19. State of the art results (*Other methods*) can be found in `https://github.com/google-research/bert/blob/master/multilingual.md`.

3.4 LRP-weighted training

Following Sun et al. (2021a), we attempt to utilize LRP contributions during training, and examine performance. In our case, the representation of every intermediate source or target token x_i, with Relevance Score $R_t(x_i)$, is reweighted by its score at each layer, and included in a new loss term, $L_{\text{ce}}(y, x_i)$. The loss then is the weighted sum of the previous and the new terms, each weighted by parameters ξ, λ respectively, for which we experiment with three sets of values:

$$\xi, \lambda = \{v_1 = \{1, 0.5\}, v_2 = \{0, 1\}, v_3 = \{1, 1\}\}. \tag{4}$$

In the first layer we only weigh the word embedding of the token. We hypothesize that in this way, the tokens with a higher contribution to the NMT result are enhanced, while the effect of the ones contributing less is reduced.

4 Results & Discussion

In Tables 1, 2, 3, we present our results for LRP-guided training, in certain low- and high-resource Semi-Supervised and Supervised experiments, for all languages and directions, providing the regular NMT model results as our baselines.

We see that for En–Fr and Fr–En NMT, in Table 1, the method fails to outperform baseline NMT results in all cases. The model translation quality is usually on par with baselines, and state-of-the art results on high scale experiments, and small differences in BLEU scores in the range of 0.1-0.5 can be considered negligible. Among the three hypermarameter settings, choosing v1 for training seems to outperform the other two in the majority of experiments under a MT-only setting. Results could indicate unsuitability of the method in low-resource settings; the original method was after all proposed in a few-shot classification context, hence we also seek more promising results in low-resource NMT experiments, examined below.

A different model behavior is observed for all cases but one in English to Gujarati NMT, in Table 2 when either the MT-only or BT-AE-MT objectives are used in training. This is an interesting finding - we can hypothesize that LRP-guided training might be more useful when translating into highly complex morphological languages such as Gujarati; Results certainly outperform previous state-of-the-art approaches, however more research including other languages and potentially other parameter values is required to verify that observation.

Very encouraging is also the case of LRP-guided training in En–Kk and Kk–En NMT. In a large number of settings, training with Relevance guidance improves NMT BLEU scores significantly compared to our regular models' results. More specifically, we see improvement in both low- (22k) and mid-resource (132k) experiments, with v2-parameterized models to outperform our baselines in the majority of cases. Also, all experiments we are able to perform significantly better than the state of the art current result, hence relevance guidance shows great potential again in NMT experiments where few parallel data are available.

5 Conclusions

We perform a series of Semi-Supervised and Supervised Neural Machine Translation experiments, using an explainability-based metric, namely Relevance-guided propagation, during training; we leverage the measure of influence of the input and intermediate layer outputs to the NMT result, in an attempt to improve NMT for three quite different languages, lying in both high- and low- resource data regimes. Our results, though showing marginal and very small improvements, indicate that Layerwise-relevance propagation shows potential in boosting NMT quality when training in small data scenarios. Further exploration of the method, different model hyperparameter setups, and expansion of our method to other languages is strongly recommended as a next step to identify the efficiency and robustness of the proposed method.

Limitations

Training a large Neural Machine Translation model from scratch is a hard task computationally, and employing LRP-guidance during training significantly raises training time, the amount and usage of required computational resources, and the complexity of the training process, calling for more efficient training solutions, in terms of memory distribution of the model and parallelization. These factors constitute the limitations of our approach, and allowed us to launch a small number of experiments, hence addressing those factors and expanding to more languages, in more efficient training and computational ways, is a strong requirement for further generalization of the method.

Ethics Statement

Several ethical concerns ought to be addressed when working with large language models regarding quality, toxicity and bias related to their training process and output (Bender et al., 2021; Chowdhery et al., 2022; Brown et al., 2020),of which the authors of the paper are aware in their work.

References

Christopher J Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. 2022. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10705–10714.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P Parikh. 2020. A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *arXiv preprint arXiv:2108.04840*.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2022. Refining low-resource unsupervised translation by language disentanglement of multilingual model. *arXiv preprint arXiv:2205.15544*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kun Qian, Marina Danilevsky, Yannis Katsis, Ban Kawas, Erick Oduor, Lucian Popa, and Yunyao Li. 2021. Xnlp: A living survey for xai research in natural language processing. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 78–80.

Dominik Schiller, Tobias Huber, Florian Lingenfelser, Michael Dietz, Andreas Seiderer, and Elisabeth André. 2019. Relevance-based feature masking: Improving neural network based whale classification through explainable artificial intelligence.

Frantisek Sefcik and Wanda Benesova. 2021. Improving a neural network model by explanation-guided training for glioma classification based on mri data. *arXiv preprint arXiv:2107.02008*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.

Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, and Alexander Binder. 2022. Explain and improve: Lrp-inference fine-tuning for image captioning models. *Information Fusion*, 77:233–246.

Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. 2021a. Explanation-guided training for cross-domain few-shot classification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7609–7616. IEEE.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Qiu Han, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021b. Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2020. Analyzing the source and target contributions to predictions in neural machine translation. *arXiv preprint arXiv:2010.10907*.

Rui Wang and Hai Zhao. 2021. Advances and challenges in unsupervised neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–21.

Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. 2022. Beyond explaining: Opportunities and challenges of xai-based model improvement. *arXiv preprint arXiv:2203.08008*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Andrea Zunino, Sarah Adel Bargal, Pietro Morerio, Jianming Zhang, Stan Sclaroff, and Vittorio Murino. 2021a. Excitation dropout: Encouraging plasticity in deep neural networks. *International Journal of Computer Vision*, 129(4):1139–1152.

Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. 2021b. Explainable deep classification models for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3233–3242.