Improving Language Model Integration for Neural Machine Translation

Christian Herold **Yingbo Gao** **Mohammad Zeineldeen** **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{herold|ygao|zeineldeen|ney}@cs.rwth-aachen.de

Abstract

The integration of language models for neural machine translation has been extensively studied in the past. It has been shown that an external language model, trained on additional target-side monolingual data, can help improve translation quality. However, there has always been the assumption that the translation model also learns an implicit target-side language model during training, which interferes with the external language model at decoding time. Recently, some works on automatic speech recognition have demonstrated that, if the implicit language model is neutralized in decoding, further improvements can be gained when integrating an external language model. In this work, we transfer this concept to the task of machine translation and compare with the most prominent way of including additional monolingual data - namely back-translation. We find that accounting for the implicit language model significantly boosts the performance of language model fusion, although this approach is still outperformed by back-translation.

1 Introduction

Machine translation (MT) is the task of automatically translating text from one language to another. Nowadays, the dominant approach is neural machine translation (NMT), where a neural network is used to predict the probability of a sentence in the target language, given a sentence in the source language (Bahdanau et al., 2014; Vaswani et al., 2017). For this approach to be effective, a large number of bilingual training samples - consisting of sentences and their corresponding translations - is needed. This poses a challenge, especially when we want to build a system for a specific domain, where zero or only limited amounts of in-domain bilingual data are available.

In these situations, people turn towards monolingual text data, which is simply text in the source or target language and of which plenty exists for most languages and domains. Before NMT became feasible, the preferred way of incorporating additional monolingual data in the MT system was the usage of an external target-side language model (LM), which is trained on monolingual data to predict the probability of a sentence (Brown et al., 1990; Della Pietra, 1994; Zens et al., 2002).

However, with the rise of NMT, it was found that a technique called back-translation outperforms the LM incorporation by a large margin (Sennrich et al., 2016a). Back-translation is a two step process, where we first create synthetic parallel data by automatically translating target side monolingual data into the source language. Then, the final NMT system is trained on the combination of the real and synthetic parallel data. It was argued that the back-translation approach better suits the NMT framework because the NMT system implicitly learns an internal language model (ILM) as part of the training, which might interfere with an additional external LM (Sennrich et al., 2016a).

More recently, for automatic speech recognition (ASR), there have been works focusing on neutralizing this ILM before combination with an external LM and significant improvements were reported (McDermott et al., 2019; Variani et al., 2020; Meng et al., 2021; Zeyer et al., 2021; Zeineldeen et al., 2021). In this work, we adapt the methods for ILM compensation, developed for ASR, and test them for NMT. We compare against back-translation in different settings and find that ILM compensation significantly boosts the performance of LM fusion, although back-translation is still outperforming this approach for NMT. Also, applying ILM compensation on top of back-translation does not result in significant performance improvements.

2 Related Work

Several approaches to combine an LM and NMT model have been proposed in the past. Shallow

Findings of the Association for Computational Linguistics: ACL 2023, pages 7114–7123

fusion (SF) is the most straight forward way, using a weighted log-linear combination of the model output probabilities (Gulcehre et al., 2015, 2017). Deep fusion denotes the concatenation of the hidden states of NMT model and LM and requires joint fine-tuning of both models (Gulcehre et al., 2015, 2017). Simple fusion is similar to shallow fusion, but the NMT model is trained using information from a pre-trained LM (Stahlberg et al., 2018).

For the task of ASR, people recently have started to remove the ILM that is implicitly learned. The biggest question there is, how to best approximate the ILM. Approaches include: (1) training an additional LM on the target side of the parallel data (McDermott et al., 2019), (2) removing/averaging encoder information (Variani et al., 2020; Meng et al., 2021; Zeyer et al., 2021) and (3) training a small sub-network while freezing all other parameters (Zeineldeen et al., 2021).

As an alternative to LM fusion, back-translation (Schwenk, 2008; Bertoldi and Federico, 2009; Sennrich et al., 2016a) has become the standard method for incorporating additional monolingual data for NMT. Some work has been done to improve this approach, including sampling (Edunov et al., 2018; Graça et al., 2019), tagging (Caswell et al., 2019) and block-BT (Popel et al., 2020). For sake of simplicity, we focus on the standard back-translation approach using beam search in this work.

Apart from using an external LM and back-translation, additional monolingual data can also be utilized by pre-training (Ramachandran et al., 2017; Zhu et al., 2019), multi-task-learning (Zhang and Zong, 2016; Domhan and Hieber, 2017) or post-editing (Junczys-Dowmunt and Grundkiewicz, 2016; Freitag et al., 2019). In principle, all these approaches can also be combined with LM fusion, potentially further improving the performance of the resulting system.

3 Internal LM Estimation

During decoding, given a source sentence f_1^J and a model $P(e_1^I|f_1^J)$, we want to find the translation $\hat{e}_1^{\hat{I}}$ that maximizes

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}} \left\{ P(e_1^I|f_1^J) \right\}.$$

In our framework, P is the combination of three models:

$$P(e_1^I|f_1^J) \propto P_{\text{MT}}(e_1^I|f_1^J) \cdot P_{\text{LM}}^{\lambda_1}(e_1^I) \cdot P_{\text{ILM}}^{-\lambda_2}(e_1^I)$$

where P_{MT}, P_{LM} and P_{ILM} are the probabilities of the NMT model, external LM (trained on additional monolingual data) and ILM respectively, and $\lambda_1, \lambda_2 \geq 0$. Note that the ILM gets a negative weight, because we want to neutralize its impact in this model combination. If $\lambda_2 = 0$, we fall back to standard shallow fusion.

In principle, the ILM can be exactly calculated from the NMT model by marginalizing over all source sentences f_1^J. However, this summation would be intractable. Instead, different ILM approximations have been proposed in the recent past for ASR, which we will briefly recall here. For a more in-depth discussion of the different approximation methods we refer the reader to Zeineldeen et al. (2021).

separate LM : The ILM is approximated by training a separate LM on the target side of the parallel training data.

$\boldsymbol{h = 0}$: The ILM is approximated by taking the fully trained NMT model $P_{\text{MT}}(e_1^I|f_1^J)$ and setting the encoder outputs h_1^J to 0.

$\boldsymbol{h = h_{\text{avg}}}$: Instead of setting all encoder outputs h_1^J to 0, we replace the vector h_j for each position j with the average $h_{\text{avg}\,j}$, extracted over the whole parallel training data.

$\boldsymbol{c = c_{\text{avg}}}$: Instead of h, we replace all context vectors c (the output of the encoder-decoder attention module) with the position-wise average over the whole parallel training data.

mini-self-attn : We replace the encoder-decoder attention of the fully trained NMT model with an additional self-attention module (with causal masking), which is then trained on the target side of the parallel training data while the rest of the NMT network is frozen. This is different from the *separate LM* approach because most of the parameters are still shared between NMT model and ILM, which might result in a better overall ILM approximation.[1]

4 Experiments

We perform experiments on four machine translation tasks, representing different data conditions.

[1] In their work, Zeineldeen et al. (2021) used a mini-LSTM network with the same dependencies as our mini-self-attention.

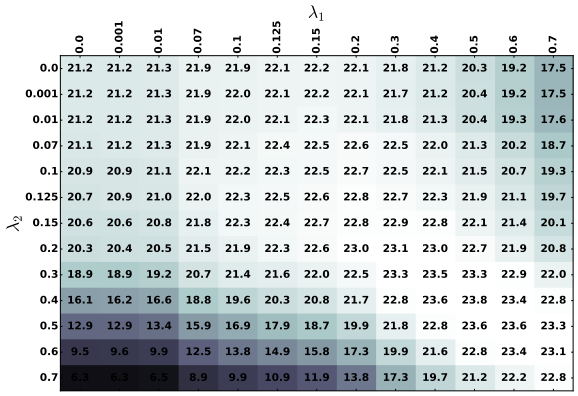

Figure 1: BLEU scores (percentage) on the validation set for the IWSLT En→De task for different weights of LM and ILM (*mini-self-attention*). λ_1 on the x-axis is the weight of the external LM while λ_2 on the y-axis is the (negative) weight of the ILM.

The exact data conditions and statistics are provided in the Appendix A. For all tasks, the additional monolingual data, as well as the test sets, are in the news domain. The monolingual data comes from Newscrawl[2] where we sample ca. 10M sentences for LM training and back-translation. For **IWSLT En→De** and **IWSLT En→It**, the parallel training data consists of around 200k sentence pairs and is in the scientific-talks-domain, coming from the IWSLT17 Multilingual Task (Cettolo et al., 2017). For this setting, we expect the biggest improvements from the additional monolingual data, since the parallel data is out-of-domain. For **NEWS En→De**, the parallel training data (around 300k sentence pairs) is in the news domain, coming from the NewsCommentaryV14 corpus[3]. Finally, **WMT14 En→De** is a standard NMT benchmark used by Vaswani et al. (2017) where the parallel training data consists of around 3.9M sentence pairs and is of mixed domain.

We tokenize the data using byte-pair-encoding (Sennrich et al., 2016b; Kudo, 2018) with 15k joint merge operations (40k for WMT14). The models are implemented using the fairseq toolkit (Ott et al., 2019) following the transformer base architecture (Vaswani et al., 2017). The details of the training setups can be found in Appendix A. All systems are trained until the validation perplexity no longer improves and the best checkpoint is selected using validation perplexity as well. We use beam-search with beam-size 12 and utilize SacreBLEU (Post, 2018) to calculate BLEU (Papineni et al., 2002)

[2] https://data.statmt.org/news-crawl/

[3] https://data.statmt.org/news-commentary/v14/

Method	valid-PPL
separate LM	109.9
$h = 0$	251.3
$h = h_{\text{avg}}$	240.9
$c = c_{\text{avg}}$	244.2
mini-self-attention	108.4

Table 1: Perplexities of the validation set for the IWSLT En-De task using different ILM model approximations.

ILM	λ_1	λ_2	BLEU	TER
-	0	0	28.9	52.8
-	0.15	0.0	30.0	52.3
separate LM	0.5	0.3	31.2	50.9
$h = 0$	0.5	0.3	30.8	51.3
$h = h_{\text{avg}}$	0.5	0.3	31.1	51.1
$c = c_{\text{avg}}$	0.5	0.3	30.6	51.5
mini-self-attn	0.5	0.4	31.7	50.0

Table 2: Translation performance of the different ILM variants on the test set of the IWSLT En-De task. BLEU and TER are reported in percentage.

and TER (Snover et al., 2006). We report BLEU and TER since we are most familiar with these metrics and to be comparable with previous works. However, we acknowledge that these metrics might have some biases and in future work it might be worth utilizing additional metrics like COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020). Additionally, in future work we should separate our test sets for original source and target text to better understand the effect of translationese in both training and test data, as this might very much influence the improvements we see, especially in the case of back-translation (Freitag et al., 2020).

4.1 Comparison of ILM Approximations

We start by analyzing the ILM neutralization approaches on the IWSLT En→De task and then verify the results on the other tasks.

We implement and re-train (if applicable) all the different ILM approximation methods discussed in Section 3. The resulting perplexities on the validation set are listed in Table 1. The variants *separate LM* and *mini-self-attention* have been trained directly using the language model objective, so it is no surprise that they exhibit a much lower perplexity than the other approaches. However, it can be argued that a lower perplexity of the ILM does not necessarily correspond to a better approximation

Method	IWSLT En-De		IWSLT En-It		NEWS En-De		WMT14 En-De	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
baseline *external*	-	-	-	-	†32.3	-	‡27.3	-
baseline *ours*	28.9	52.8	24.1	58.9	32.8	49.0	27.7	56.5
+SF	30.0	52.3	24.8	58.8	33.2	49.8	28.1	56.6
+ILM (*separate LM*)	31.2	50.9	26.0	57.8	34.7	47.6	28.8	55.3
+ILM (*mini-self-attn*)	31.7	50.0	26.1	57.0	35.1	47.5	29.1	54.8
back-translation	34.1	47.4	27.2	56.9	35.7	45.8	29.5	54.7
+SF +ILM (*mini-self-attn*)	34.1	47.6	27.3	56.7	35.7	46.0	29.8	54.3

Table 3: Comparison of LM fusion and back-translation on the four MT tasks. BLEU and TER are reported in percentage. External baselines are from † Kim et al. (2019) and ‡ Vaswani et al. (2017).

of the implicit language model.

In order to effectively use the external LM and the ILM during decoding, we need to optimize the weights λ_1 and λ_2 (see Section 3). We do this via a grid search over the validation set by optimizing for the highest BLEU score. The resulting grid for the *mini-self-attention* ILM variant on the IWSLT En→De task is shown in Figure 1.

The NMT system by itself has a BLEU[%] score of 21.2. By log-linear combination with just the external LM ($\lambda_2 = 0$, vanilla shallow fusion) we can gain around 1% absolute improvement on the validation set with the best choice of $\lambda_1 = 0.15$. By including the ILM with a negative weight, we can get further improvements, up to a final score of 23.8 BLEU[%]. [4] Interestingly, the best performance is reached when $\lambda_1 \approx \lambda_2$ and with the ILM neutralization, the external LM can be assigned a much bigger weight compared to the case $\lambda_2 = 0$. We find that for all ILM approximation variants, the optimal weights are similar, and that the TER scores on the validation set follow an almost identical pattern. The final performance of each variant on the test set is shown in Table 2.

We want to point out, that the improvements we see on the validation set transfer nicely to the test set with the same tuned weights λ_1 and λ_2. This is because, in our experiments, the validation and test sets are of the exact same domain. In some additional experiments we found that the optimal values for these weights are indeed domain specific and have to be re-tuned if the system were to be optimized for a different domain. All ILM approximation variants lead to a significant performance improvement over simple shallow fusion. Out of all ILM approximations, the *mini-self-attention* approach performs best, which is the same observation that Zeineldeen et al. (2021) made for ASR.

[4]For the *mini-self-attention* ILM variant, we also performed a more fine-grained search for $0.3 < \lambda_1, \lambda_2 < 0.6$ which did not result in further improvements.

4.2 Comparison to Back-Translation

For the back-translation experiments, we train NMT systems on the same parallel training data in the reverse direction and then translate a total of 10M sentences from the monolingual target data (the same data used for training the external LM). Afterwards, the final systems are trained on the combination of real and synthetic data. The final results for all four MT tasks are shown in Table 3.

We observe the same trend for all four MT tasks. In general, the improvements from the additional monolingual data are getting smaller, when the amount of parallel training data increases. In almost all cases, shallow fusion gives a small improvement over just using the NMT system. ILM neutralization again improves consistently over simple shallow fusion, with the *mini-self-attn* approximation variant always performing the best. Back-translation out-performs language model integration on all four tasks, although the gap is getting smaller the more parallel training data is available.

We also combine back-translation with the best ILM approximation approach (*mini-self-attn*). This does not further increase translation quality, with the exception of the WMT14 task, where we see a small improvement. In general, the ILM approach performs the closest to back-translation on the WMT14 task, so it might be worthwhile to apply this concept to an even bigger MT task.

5 Conclusion

We re-visit the method of language model integration for neural machine translation. We implement and experiment with a new approach of neutraliz-

ing the implicit language model, which has already shown promising result for the task of automatic speech recognition. We find that ILM neutralization significantly improves the translation quality compared to standard shallow fusion. However, back-translation as an alternative way to incorporate additional monolingual data, still outperforms the approaches using an external language model. Therefore, for future work we will focus on scenarios where back-translation can not be applied effectively, e.g. when the quality of the initial NMT system is too bad to create helpful synthetic data.

Acknowledgements

This work was partially supported by the project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A, and by NeuroSys which, as part of the initiative "Clusters4Future", is funded by the Federal Ministry of Education and Research BMBF (03ZU1106DA).

Limitations

The approach of language model integration for neural machine translation is analyzed and compared to the de-facto standard method of back-translation. Due to constrained resources, this work has several limitations. We focus on translation of text in a single domain, namely news-articles. Different domains might exhibit different behaviour. For the back-translation experiments, we use beam search to create the synthetic data, other methods like sampling were not considered. When combining the synthetic and real parallel data, there are additional methods like tagging and block-wise batching, which we did not utilize in this work. Finally, we compare against the most commonly used LM fusion approach, i.e. shallow fusion. There exist other LM fusion techniques which might exhibit different behaviour when used in combination with ILM neutralization.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuitho Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Vincent J Della Pietra. 1994. The mathematics of statistical machine translation: Parameter estimation. *Using Large Corpora*, page 223.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine*

Translation (Volume 1: Research Papers), pages 45–52, Florence, Italy. Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Erik McDermott, Hasim Sak, and Ehsan Variani. 2019. A density ratio approach to language model fusion in end-to-end automatic speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 434–441. IEEE.

Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, and Yifan Gong. 2021. Internal language model estimation for domain-adaptive end-to-end speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 243–250. IEEE.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

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.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Prajit Ramachandran, Peter J Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation: Papers*, pages 182–189, Waikiki, Hawaii.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (hat). In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6139–6143. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mohammad Zeineldeen, Aleksandr Glushko, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2021. Investigating methods to improve language model integration for attention-based encoder-decoder asr models. In *Interspeech*, pages 2856–2860.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence*, pages 18–32. Springer.

Albert Zeyer, André Merboldt, Wilfried Michel, Ralf Schlüter, and Hermann Ney. 2021. Librispeech transducer model with internal language model prior correction. *arXiv e-prints*, pages arXiv–2104.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2019. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

A Appendix

All validation and test sets are from the WMT news translation tasks (Farhad et al., 2021). The validation/test sets are WMT newstest2015/newstest2018 for IWSLT En→De and NEWS En→De, newssyscomb2009/newstest2009 for IWSLT En→It and newstest2013/newstest2014 for WMT14 En→De. Data statistics can be found in Table 4.

task	dataset	domain	# sent.
IWSLT	train	scientific-talks	210k
En→De	valid	news	2.2k
	test	news	3k
	mono.	news	9.7M
IWSLT	train	scientific-talks	232k
En→It	valid	news	500
	test	news	2.5k
	mono.	news	10.0M
NEWS	train	news	330k
En→De	valid	news	2.2k
	test	news	3k
	mono.	news	9.7M
WMT14	train	mixed	3.9M
En→De	valid	news	3k
	test	news	3k
	mono.	news	10.0M

Table 4: Data statistics for all tasks.

We use dropout 0.3 and label-smoothing 0.2 for IWSLT En→De, IWSLT En→It and NEWS En→De and dropout 0.3 and label-smoothing 0.1 for WMT14 En→De. The resulting NMT models had ca. 51M parameters for IWSLT En→De, IWSLT En→It and NEWS En→De and ca. 67M parameters for WMT14 En→De. The NMT training took around 24h for IWSLT En→De, IWSLT En→It and NEWS En→De and around 150h for WMT14 En→De on a single NVIDIA GeForce RTX 2080 Ti graphics card. The language models had ca. 26M parameters for IWSLT En→De, IWSLT En→It and NEWS En→De and ca. 41M parameters for WMT14 En→De. All language model trainings took around 150h on a single NVIDIA GeForce RTX 2080 Ti graphics card. Due to computational limitations, we report results only for a single run.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☒ A2. Did you discuss any potential risks of your work?
The authors do not foresee potential risks of this work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 Experiments

☑ B1. Did you cite the creators of artifacts you used?
Section 4 Experiments

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All artifacts that were used allow such usage for research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All artifacts that were used allow such usage for research purposes.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We only use standard datasets which allow usage for research purposes.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 Experiments

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section Appendix

C ☑ Did you run computational experiments?

Section 4 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 Experiments

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section Appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 Experiments

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.