Document-Level Language Models for Machine Translation

Frithjof Petrick[1,2] **Christian Herold**[1,2] **Pavel Petrushkov**[1]
Shahram Khadivi[1] **Hermann Ney**[2]
[1]eBay, Inc., Aachen, Germany
{cherold, ppetrushkov, skhadivi}@ebay.com
[2]Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{petrick, ney}@i6.informatik.rwth-aachen.de

Abstract

Despite the known limitations, most machine translation systems today still operate on the sentence-level. One reason for this is, that most parallel training data is only sentence-level aligned, without document-level meta information available. In this work, we set out to build context-aware translation systems utilizing document-level monolingual data instead. This can be achieved by combining any existing sentence-level translation model with a document-level language model. We improve existing approaches by leveraging recent advancements in model combination. Additionally, we propose novel weighting techniques that make the system combination more flexible and significantly reduce computational overhead. In a comprehensive evaluation on four diverse translation tasks, we show that our extensions improve document-targeted scores substantially and are also computationally more efficient. However, we also find that in most scenarios, back-translation gives even better results, at the cost of having to re-train the translation system. Finally, we explore language model fusion in the light of recent advancements in large language models. Our findings suggest that there might be strong potential in utilizing large language models via model combination.

1 Introduction

Machine translation (MT), the automatic translation of text from one language to another, has seen significant advancements in recent years, primarily driven by neural machine translation (NMT) models (Bahdanau et al., 2015; Vaswani et al., 2017). These models have demonstrated remarkable capabilities in capturing complex linguistic patterns and producing high-quality translations (Wu et al., 2016; Hassan et al., 2018). Nevertheless, most models to-date operate on sentence-level, i.e. translate sentences independently without the context of the surrounding document. Without access to such context, it is impossible for these MT systems to account for discourse-level phenomena such as resolution of ambiguous words and coherence. Unsurprisingly, automatic translations are perceived as much worse, when they are evaluated on entire documents rather than just at the sentence-level (Läubli et al., 2018, 2020; Maruf et al., 2022).

An obvious solution to this problem is to utilize context-aware MT models (Tiedemann and Scherrer, 2017). While document-level NMT models have been thoroughly studied in recent years, sentence-level MT remains the standard despite its inherent limitations. One of the main reasons for this is that most of the document-level approaches rely on parallel training data with document-level metadata. Most releases of large parallel training corpora lack this information and remain purely sentence-level (Bañón et al., 2020; Schwenk et al., 2021). In contrast, large amounts of document-level monolingual data are readily available for almost all domains and languages.

In this work, we strive to build a context-aware MT system that does not rely on any parallel document-level training data. Instead, we use monolingual documents to train a document-level language model (LM), which we fuse with an existing sentence-level MT model during translation. While existing work on LM fusion shows that the fused model is able to incorporate document-level context (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021), these approaches can be improved. Our work aims to do so in two main directions.

First, we acknowledge that NMT models implicitly learn the language modeling task during training. Recently, Herold et al. (2023) showed that estimating and neutralizing this internal LM can improve translation quality for sentence-level MT. We adapt their approach to document-level LM fusion and demonstrate that this also improves discourse modeling.

Second, the contribution of the fused MT model,

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 375–391
December 6–7, 2023.

the document-level LM and the internal LM must be balanced by a set of fusion scales. Existing work defines the fusion scales as static hyperparameters which are tuned on a validation set via an extensive grid search (Gülçehre et al., 2015; Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). In our work, we provide two simple alternatives to grid search which allow for automatically tuned context-dependent fusion scales. Our approaches eliminate the need for expensive tuning and further improve discourse-modelling.

The contributions of this work are as follows:

1. We propose multiple extensions to the existing approaches on document-level LM fusion for MT.

2. We compare our methods against two strong baselines: Back-translation, the to-date most popular way to utilize monolingual data for MT, and a task-specific LM re-ranking baseline for pronoun disambiguation. The comparison takes place over four diverse translation tasks in terms of general translation quality as well as specific context-dependant phenomena.

3. We present first results on fusing a large language model (LLM) with a sentence-level MT system.

2 Related Works

Most works on document-level NMT rely on parallel document-level data for system training.

Tiedemann and Scherrer (2017) propose to concatenate adjacent sentences on source and target side and input this into the NMT model which has the exact same architecture as the vanilla sentence-level transformer (Vaswani et al., 2017). Later, many works have proposed modifications to the architecture to better accommodate the additional context (Jean et al., 2017; Bawden et al., 2018; Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018; Miculicich et al., 2018; Maruf and Haffari, 2018). However, it has been shown that the simple concatenation approach performs as good, if not better than these more complicated variants (Lopes et al., 2020; Sun et al., 2022).

Maybe the biggest challenge for document-level NMT is that most of the parallel MT training data is not document-level (Esplà-Gomis et al., 2019; Schwenk et al., 2021). Recently there has been some effort to restore document-level meta information from existing sentence-level corpora but this is a very time consuming and error-prone process (Ghussin et al., 2023). Therefore, approaches to document-level NMT have been proposed that utilize document-level monolingual data, of which typically large amounts are readily available.

One direction is to back-translate the document-level monolingual data to create synthetic parallel document-level data. The reverse system used for back-translation can be either sentence-level (Junczys-Dowmunt, 2019; Saleh et al., 2019; Post and Junczys-Dowmunt, 2023) or document-level (Sugiyama and Yoshinaga, 2019; Huo et al., 2020). A downside of this approach is that the final MT system has to be re-trained to incorporate the new synthetic data.

Another line of work uses document-level language models in combination with sentence-level translation models. Gülçehre et al. (2015) were the first to propose a log-linear combination of sentence-level language and NMT models, coining the term 'shallow fusion'. Recently, it was shown that the shallow fusion approach for sentence-level NMT can be improved by compensating for the implicitly learned internal language model of the NMT system (Herold et al., 2023). Regarding the integration of a document-level LM, earlier approaches simply use the LM for re-ranking the hypothesis of the sentence-level NMT model (Stahlberg et al., 2019; Yu et al., 2020). Several works have proposed to employ a log-linear combination between sentence-level NMT system and document-level LM (Garcia et al., 2019; Jean and Cho, 2020; Sugiyama and Yoshinaga, 2020). Both Jean and Cho (2020) and Sugiyama and Yoshinaga (2020) propose to also include the probabilities of the LM without context information in order to mitigate the influence of the current sentence on the LM probabilities. While our approach also uses the output of a sentence-level LM, it is conceptually different from the previous works in that we want to mitigate the influence of the internal LM from the NMT model, resulting in a different final formulation. To further improve LM incorporation, Jean and Cho (2020) propose to use subword-dependent fusion scales instead of a single scale per model.

Apart from back-translation and LM integration there exist some other ways to utilize additional monolingual document-level data for MT. Voita et al. (2019) train a document-level automatic post editing system on the monolingual data and use it to improve the hypotheses from a sentence-level NMT system in a two-pass approach. Several works utilize the additional data in a multi-task learning approach (Junczys-Dowmunt, 2019) or for pre-training (Zhu et al., 2020; Chen et al., 2021b; Liu et al., 2020; Chen et al., 2021a).

Very recently, LLMs have shown their potential for the task of document-level NMT (Wang et al., 2023). However, it is unclear how much parallel training samples were seen during the large scale pre-training on trillions of tokens.

3 Document-level Language Model Fusion

The sentence-level MT model translates a source sentence F into a target sentence $E := e_0^I$ of subwords e_i. In the document-level LM fusion approach, we additionally provide the k previous target-side sentences E_{-k}^{-1} as context[1].

3.1 Internal Language Model Neutralization

As the translation model already implicitly learns probabilities that are source-independent, directly fusing the MT model and the document-level LM overvalues the source-agnostic probabilities. Therefore, we estimate the internal LM of the MT model and in total combine three models during generation:

- the existing sentence-level MT model $p_{\text{TM}}(e_i) := p_{\text{TM}}(e_i \,|\, e_0^{i-1}, F)$,
- the LM $p_{\text{LM}}(e_i) := p_{\text{LM}}(e_i \mid e_0^{i-1}, E_{-k}^{-1})$ trained on monolingual documents with access to the previous target sentences E_{-k}^{-1},
- and a second LM $p_{\text{ILM}}(e_i) := p_{\text{ILM}}(e_i \,|\, e_0^{i-1})$ which estimates the internal LM probabilities implicitly learned by the MT model. We train this LM separately on the target-side of the MT training data, as we found that this approach works best for document-level MT when compared to other approaches presented by Herold et al. (2023). This comparison can be found in Appendix A.3.

We multiply the model output probabilities and normalize them. The resulting probability distribution is now conditioned on both the source sentence F and the target-side context E_{-k}^{-1}:

$$p(e_i) := p(e_i \,|\, e_0^{i-1}, F, E_{-k}^{-1}) := \frac{p_{\text{TM}}^{\lambda_0}(e_i) \cdot p_{\text{LM}}^{\lambda_1}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0}(e') \cdot p_{\text{LM}}^{\lambda_1}(e') \cdot p_{\text{ILM}}^{-\lambda_2}(e')}. \quad (1)$$

Each model is weighted with a scalar $\lambda_0, \lambda_1, \lambda_2 \geq 0$, the internal LM is included with a negative exponent. We tune these fusion scales on the validation set for BLEU via a grid search over $\lambda_0, \lambda_1, \lambda_2 \in \{0, 0.1, \ldots, 1\}$.

Existing work on document-level LM fusion uses a similar formulation as our approach, but instead of neutralizing the internal LM of the MT model, it accounts for the sentence-level probabilities $p_{\text{LM}}(e_i \,|\, e_0^{i-1})$ of the document-level LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). In the particular case where there are no previous sentences available, this approach simply falls back to using only the sentence-level MT model probabilities. Our approach on the contrary can also leverage the gains obtained from sentence-level LM fusion and is theoretically more expressive.

3.2 Context-dependent Fusion Scales

Choosing appropriate fusion scales $\lambda_0, \lambda_1, \lambda_2$ in Equation 1 is crucial. Conventionally, the scales are tuned via grid search. This is problematic in three aspects:

1. Grid search is expensive. Testing e.g. ten possible values for each of the three model scales already requires translating the validation set 1000 times.
2. The tuning process depends on the tuning data, its domain and the tuning objective. E.g., the scales that optimize document-targeted metrics differ from the ones that maximize sentence-level translation quality (Sugiyama and Yoshinaga, 2021).
3. Fusion scales obtained by a hyperparameter grid search must be constant. Document-level context however is not uniformly useful for all predicted subwords.

In the following, we propose two simple alternatives to obtaining fusion scales with grid search that overcome the aforementioned issues.

[1] At the beginning of the document we only provide as many sentences as available.

3.2.1 On-the-fly Fusion Scales

During decoding, the next subword e_i is chosen to maximize the fused probability (Equation 1). We propose to also choose the fusion scales in a similar fashion and define them to maximize the fused model scores:

$$(\lambda_0, \lambda_1, \lambda_2) := \underset{(\lambda_0,\lambda_1,\lambda_2)}{\operatorname{argmax}} \frac{p_{\text{TM}}^{\lambda_0}(e_i) \cdot p_{\text{LM}}^{\lambda_1}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0}(e') \cdot p_{\text{LM}}^{\lambda_1}(e') \cdot p_{\text{ILM}}^{-\lambda_2}(e')}. \quad (2)$$

Our model maximizes over the discrete set $\lambda_0, \lambda_1, \lambda_2 \in \{0, 0.1, \ldots, 1\}$. This approach obviates the need for separate scale tuning entirely and only has a small overhead during generation.

3.2.2 Automatically Learned Fusion Scales

Alternatively, we propose to learn the fusion scales automatically using a small amount of training examples (F, E, E_{-k}^{-1}) with document-level context, similarly to Jean and Cho (2020). We obtain the training data by back-translating the monolingual data (see Section 5). Automatic learning allows us to implement subword-dependent fusion scales: We introduce a set of learnable parameters $\lambda_0(e), \lambda_1(e), \lambda_2(e)$ for each subword e from the target vocabulary and learn them automatically by optimizing the cross-entropy loss

$$(\lambda_0, \lambda_1, \lambda_2) := \underset{\lambda:\, V \to \mathbb{R}^3}{\operatorname{argmax}} \sum_{(F,E,E_{-k}^{-1})} \sum_i \log \frac{p_{\text{TM}}^{\lambda_0(e_i)}(e_i) \cdot p_{\text{LM}}^{\lambda_1(e_i)}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2(e_i)}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0(e')}(e') \cdot p_{\text{LM}}^{\lambda_1(e')}(e') \cdot p_{\text{ILM}}^{-\lambda_2(e')}(e')}. \quad (3)$$

Scale learning uses the same optimization parameters as the MT model was originally trained with. The scale parameters are initialized with a small variance around zero while all other parameters are frozen.

4 Document-level Language Model Pronoun Re-ranking

Besides consistency, the main problem of discourse-modelling are ambiguities. E.g. translating the English pronoun 'it' to German requires access to the noun that it refers to, which might only be found in a preceding sentence (Müller et al., 2019).

We propose an approach specific to the En→De language pair that directly targets the pronoun translation problem by re-ranking sentence-level hypotheses using a document-level LM. We first translate each sentence independently using the sentence-level MT model. Each sentence-level translation is expanded to a set of candidates by replacing the pronouns with all alternatives ('er', 'sie', 'es'). All candidate translations are then scored in context of the preceding sentences using a document-level LM, and we select the pronoun for which the LM score is highest.

This approach is very much tailored to the specific pronoun translation problem for this specific language pair. While it is theoretically possible to extend this approach to cover more cases, this will require extensive human effort and is probably not feasible in most scenarios. However, we include it here, because it serves as a reasonable baseline for this popular pronoun translation benchmark.

5 Document-level Back-translation

The to-date most popular way of utilizing monolingual data for MT is to create synthetic parallel training data via back-translation (Sennrich et al., 2016). We train a sentence-level backwards MT system on the parallel data and use it to translate the document-level monolingual data back into the source language. The sentence-level translations are concatenated to obtain synthetic parallel documents (Junczys-Dowmunt, 2019; Saleh et al., 2019; Sugiyama and Yoshinaga, 2019; Huo et al., 2020; Post and Junczys-Dowmunt, 2023).

To train the final systems we combine the authentic sentence-level parallel and the synthetic document-level data. Combining both data sources is not straightforward, because of their varying size and the difference between sentence/document-level context. Therefore, we first oversample the data accordingly to have roughly the same number of sentences in both parts. Secondly, we turn the authentic sentence-level parallel data into 'pseudo-documents' by concatenating them in a random order (Junczys-Dowmunt, 2019; Jean et al., 2019). This ensures that all training data has the same context size. We found this procedure to perform best when incorporating synthetic document-level data. For a detailed comparison, see Appendix A.5.

6 Experiments

6.1 Tasks

We evaluate our approaches on four different tasks of varying data conditions and domains. Three tasks are on publicly available data and a fourth task is based on a large scale internal dataset in the e-Commerce domain. All tasks include (sentence-level) parallel training data and document-level monolingual data from the same domain. The exact data conditions are provided in Appendix A.1.

The *News En→De* data consists of news articles while the *TED En→It* task consists of scientific talks. Both are low resource with less than 1M training samples in total. The *Subtitles En→De* data consists of subtitles from various TV shows and is medium size. Finally, the *e-Commerce En→De* task is about translating item descriptions from e-Commerce listings and the training data is large scale with more than 100M examples.

While the parallel training data for the three academic tasks does provide document-level metadata, our approaches do not make use of this information and we assume that the parallel training data is sentence-level for most experiments. We only make use of this information to provide a direct comparison against the setting where document-level parallel data is assumed to be available. As ParaCrawl, like most other large-scale web-crawled parallel datasets, is not a document-level corpus, we can not conduct these experiments for the e-Commerce task.

We preprocess each corpus with byte-pair encodings (Sennrich et al., 2016) using the SentencePiece toolkit (Kudo, 2018) learned on the parallel dataset with a shared vocabulary of 32k subwords (13.6k for TED). For the e-Commerce task we additionally use inline casing (Berard et al., 2019; Etchegoyhen and Gete, 2020).

6.2 Settings

We train transformer MT models in the 'base' configuration (Vaswani et al., 2017), implemented in Fairseq (Ott et al., 2019). For the LMs we use a similar architecture but without the encoder. Our document-level models use the same architecture as the sentence-level models, we simply include context sentences by concatenating the previous two source and target sentences to all training examples, separated by a reserved symbol (Tiedemann and Scherrer, 2017).

Details on the optimization algorithm are given in Appendix A.1. The final model is selected based on the validation set perplexity. We then perform beam search with beam size 12 and length normalization. Document-level decoding uses the 'last sentence' search strategy as described in Herold and Ney (2023b).

The document-level LMs are trained on a combination of target-side of the sentence-level parallel and document-level monolingual data. Regardless of the task, we train the LMs for 300k update steps with batch size 90k, 10 % dropout, and 10 % label smoothing.

For the LM fusion experiments with non-static fusion scales, we restrict the search space to only consider scale combinations where $\lambda_0 = 1$ and $\lambda_1 = \lambda_2$. A direction comparison is given in Appendix A.4. For back-translation, we use beam search with beam size 4 and increase the training time proportionally to the new data size.

6.3 Evaluation

Document-level evaluation is challenging, as intersentential context usually is only relevant for a small fraction of words. Further, conventional metrics like BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020) do not appropriately measure how well document-level context is considered for those words where context does matter (Läubli et al., 2018, 2020; Maruf et al., 2022). However, we still report BLEU using Sacrebleu (Post, 2018) and COMET[2] on the task-specific in-domain test sets to evaluate the general MT quality.

To better evaluate the improvements from the document-level approaches, we focus on selected sentences for which document-level context is known to be important. Here, we report on two test sets focusing on ambiguities. The En→De *pronouns* test set released by Müller et al. (2018) was curated from OpenSubtitles shows and contains 12k examples. Most examples require previous sentences as context to properly translate the English pronoun 'it' with German 'er', 'sie' or 'es'. Further, the *gender-referring professions* test sets released as contextual part of MT-GenEval (Currey et al., 2022) are available for various target languages and focus on a wider range of ambiguous words, e.g. whether 'the teacher' should be translated with 'die Lehrerin' or 'der Lehrer' in German. Again, context from the previous sentences is required to determine the correct translation. We use these test sets for En→De and En→It which both comprise approx. 1.1k examples that were created by translating Wikipedia articles.

Computing BLEU and COMET on these chal-

[2] Using the `wmt22-comet-da` model (Rei et al., 2020)

lenge test sets better reflects how well a MT system handles document-level context. An even more specific metric can be obtained by focusing only on the ambiguous words. Previous work commonly reports an accuracy metric that is based on contrastive scoring, which is computed by comparing the model probabilities of the reference against a set of contrastive examples (Müller et al., 2018). This metric however can be misleading, as it not based on the generated translation but rather just on scoring. MT systems with high contrastive scores often perform poorly when their generated hypothesis is evaluated (Post and Junczys-Dowmunt, 2023). Instead, we focus on translation-based document-targeted metrics.

On the pronouns test set, we compute a pronoun F1-score as proposed by Herold and Ney (2023a). This metric directly compares the pronouns of the hypothesis and the reference and is based on the BLONDE metric (Jiang et al., 2022). On the professions test set, we report the translation-based accuracy metric suggested by their curators (Currey et al., 2022). Further, for the Subtitles system we also report a formality F1-score on its test set as proposed by Herold and Ney (2023a).

6.4 Results

We evaluate our approaches to utilize monolingual document-level data on the four MT tasks. We apply them in two settings where a) we assume that all parallel data is purely sentence-level, and b) also the parallel data is document-level.

In an effort to compare to previous work, we re-implement LM fusion with static scales without subtracting the internal LM which was independently proposed by Jean and Cho (2020) and Sugiyama and Yoshinaga (2021). These works subtract the intersentential probabilities of the external LM instead. Further, we also re-implement the non-static scales predicted with a 'merging module' learned on parallel document-level data as proposed by Jean and Cho (2020).

We first evaluate our approaches on conventional metrics to measure their general MT performance. Then, we focus on the document-targeted challenge sets to quantify how well they utilize document-level context.

6.4.1 Conventional Metrics

We start by evaluating on the in-domain test sets of the four MT tasks using the conventional MT metrics. Here, we do not expect to see much improvements coming from the document-level context. The results are presented in Table 1.

Adding monolingual data gives the largest improvements on News and small improvements on the e-Commerce task. On these two tasks, the monolingual data is in-domain and the improvements are likely because of the domain. On Subtitles and TED we do not see any improvements as Subtitles already has a large amount of in-domain parallel data and the TED monolingual data is slightly out-of-domain. We verified the domain effect by training sentence-level LMs on equal amounts of data from the target-side of the parallel and monolingual corpora and comparing their perplexities on the test sets. Details are provided in Appendix A.2.

None of the presented approaches significantly decreases translation performance in terms of conventional metrics. The only exception is the back-translation which when added to the Subtitles and TED document-level baseline performs worse in BLEU. In COMET however, this decrease is less prevalent.

6.4.2 Document-targeted Metrics

The results on the document-targeted test sets are shown in Table 2. First we discuss the scenario without access to document-level parallel training data.

LM fusion. Adding monolingual documents to the sentence-level baseline with the existing approaches from Jean and Cho (2020) and Sugiyama and Yoshinaga (2021) improves scores only marginally by on average +0.5 % absolute F1 score on the pronouns test set and no improvements on the professions set. In comparison, our approach on LM fusion with the neutralization of the internal LM performs better: E.g., the variant with on-the-fly scales on average improves the pronoun F1 score by +2.4 % and the professions ac-

[3] External baseline by Herold and Ney (2023b)

[4] External baseline by Huo et al. (2020)

[5] Re-implementation of LM fusion with neutralization of the intersentential LM probabilities instead of the internal LM, as introduced by Jean and Cho (2020) and Sugiyama and Yoshinaga (2021)

[6] Re-implementation of the 'merging module' approach by Jean and Cho (2020). This approach uses parallel document-level data for scale learning.

Data		Method	News		Subtitles		TED		e-Commerce	
parallel	mono.		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
sent.	-	baseline (prev. work)	32.8[3]	-	37.3[4]	-	34.2[3]	-	-	-
		baseline (ours)	32.7	82.8	37.3	87.9	34.8	86.1	36.4	89.2
	doc.	(Jean, 2020; Sugiyama, 2021)[5]	33.1	83.2	37.2	87.8	34.6	86.2	37.1	89.6
		(Jean, 2020)[6]	32.9	83.0	37.3	87.9	34.5	86.2	36.6	89.2
		LM: static	34.8	84.2	37.2	87.8	34.9	86.2	**37.3**	89.6
		LM: on-the-fly	34.7	83.9	37.2	87.9	34.9	86.2	36.8	**89.7**
		LM: auto. learned	34.4	83.8	37.4	87.8	34.7	86.2	36.8	89.0
		LM: re-rank pronouns	32.6	82.7	36.9	87.8	n.a.		36.4	89.2
		back-translation	37.1	85.2	37.2	87.6	35.1	86.6	36.2	89.3
		+ LM: static	**37.4**	**85.6**	**37.6**	87.7	35.2	86.6	35.0	88.9
		+ LM: on-the-fly	37.2	85.4	37.1	87.6	34.8	86.6	35.9	89.4
		+ LM: auto. learned	37.2	85.3	37.3	87.6	34.9	86.6	36.2	89.5
doc.	-	baseline	32.5	82.9	39.5	**88.2**	**35.4**	86.5	n.a.	
	doc.	LM: static	35.1	84.3	38.9	**88.2**	35.2	**86.7**		
		LM: on-the-fly	34.5	84.1	39.0	88.0	35.1	**86.7**		
		LM: auto. learned	34.8	84.1	39.3	88.2	35.2	86.6		
		LM: re-rank pronouns	32.3	82.8	39.1	88.1	n.a.			
		back-translation	37.2	85.3	37.5	87.8	34.6	86.5		

Table 1: Utilizing document-level monolingual data using different methods, reporting on the in-domain test sets of each task. BLEU and COMET are given in percentage. Best results for each column are highlighted.

Data		Method	News		Subtitles			TED	e-Commerce	
parallel	mono.		pron.	proff.	pron.	proff.	form.	proff.	pron.	proff.
sent.	-	baseline (prev. work)	45.3[3]	-	41.1[3]	-	59.4[3]	-	-	-
		baseline (ours)	45.1	65.9	41.7	65.3	57.2	65.4	42.6	63.7
	doc.	(Jean, 2020; Sugiyama, 2021)[5]	46.0	65.0	42.3	65.8	58.1	65.1	42.7	64.0
		(Jean, 2020)[6]	45.1	64.7	41.9	65.8	57.7	65.4	42.5	63.5
		LM: static	45.5	65.5	42.5	66.3	58.4	65.4	42.8	64.4
		LM: on-the-fly	48.0	65.5	44.2	65.9	58.9	66.4	44.4	66.2
		LM: auto. learned	46.7	64.9	42.8	65.5	58.6	65.6	44.0	65.2
		LM: re-rank pronouns	48.0	66.1	57.5	65.5	57.2	n.a.	**54.5**	64.0
		back-translation	48.7	80.5	52.3	67.0	58.5	65.1	42.9	67.1
		+ LM: static	48.5	80.6	53.1	68.3	53.8	65.4	42.6	66.0
		+ LM: on-the-fly	48.9	**81.3**	52.8	67.3	60.4	65.4	46.3	**70.5**
		+ LM: auto. learned	48.9	80.5	52.0	67.6	59.9	65.4	46.2	65.7
doc.	-	baseline	**55.9**	71.2	67.2	70.8	61.9	67.2	n.a.	
	doc.	LM: static	55.3	70.8	67.5	71.1	61.5	66.8		
		LM: on-the-fly	55.8	72.3	67.8	**71.9**	61.4	**67.6**		
		LM: auto. learned	55.7	71.5	**67.4**	71.0	61.6	**67.6**		
		LM: re-rank pronouns	50.9	71.5	62.6	70.8	61.9	n.a.		
		back-translation	52.1	79.4	62.8	67.3	**62.0**	65.7		

Table 2: Document-targeted evaluation of the different approaches utilizing document-level monolingual data. We report the pronoun F1 score (Herold and Ney, 2023a), gender-referring professions accuracy (Currey et al., 2022) and the formality F1 score on the Subtitles test set (Herold and Ney, 2023a), all given in percentage. Best results for each column are highlighted.

curacy by +0.9 %. Compared to static scales, both on-the-fly and automatically learned scales yield small improvements and further do not involve the expensive grid search.

LM re-ranking pronouns. Our LM re-ranking approach was specifically tailored towards the pronouns test set. We see most improvements on this test set, while the document-targeted metrics on the other test sets remain mostly unchanged. For both the Subtitles and the e-Commerce task, LM re-ranking is the best approach of utilizing document-level monolingual data for this specific test set in the absence of document-level parallel data. On News however, the gains are less prevalent: Our analysis finds that even though the LM in this case can predict the pronouns correctly, the general translation quality of the baseline on this test set is low and therefore this model often fails to generate any pronouns at all. This again highlights the discrepancy between scoring- and generation-based metrics.

Back-translation. In a direct comparison to LM fusion, back-translation outperforms LM fusion despite our improvements over the existing work. Back-translation on average improves the pronouns F1 score by +4.8 % and the professions accuracy by +4.9 % over the sentence-level baseline. This may also highlight the importance of source-side document-level context as the LM based approaches do not have access to this. Still, both back-translation and LM fusion can be combined and this yields further improvements: The best performing approach not relying on document-level parallel data is to use both document-level back-translation and then LM fusion with on-the-fly scales, this method achieves on average +6.2 % F1 score on the pronouns and +6.0 % professions accuracy.

Parallel document-level data. The three baselines trained on parallel document-level data perform much better than the sentence-level baseline: The document-level baselines score on average +18.0 % better on the pronouns F1 score and +4.2 % better on the professions accuracy than their sentence-level counterparts. In addition, the systems trained on parallel documents also perform better than the sentence-level systems with additional monolingual documents in almost all cases. This concludes that on these three tasks, having access to parallel document-level data is much more effective than utilizing monolingual document-level data, even though our monolingual corpora are much larger than the parallel ones.

Method	contrastive pronoun acc.		
	News	Subtitles	e-Comm.
sentence-level baseline	49.0	46.4	46.1
(Jean, 2020; Sugiyama, 2021)[5]	53.4	48.8	47.4
(Jean, 2020)[6]	49.2	46.8	45.5
LM: static	55.2	49.5	48.5
LM: on-the-fly	55.9	53.4	51.3
LM: auto. learned	53.0	50.1	50.5
LM: re-rank pronouns	65.7	73.9	64.8
back-translation	56.5	57.9	47.3
+ LM: static	57.7	61.6	47.5
+ LM: on-the-fly	57.9	61.1	54.3
+ LM: auto. learned	56.7	59.0	54.1
document-level baseline	**67.9**	**84.0**	n.a.

Table 3: Scoring-based, contrastive accuracies on the pronouns test set (Müller et al., 2018) for the three En→De tasks, reported in percent.

Further including monolingual document-level data to the document-level baselines does not generally give additional improvements. In particular, LM pronoun re-ranking decreases performance in this setting as the MT model itself is already better at predicting the correct pronoun than the LM trained on the document-level monolingual data.

Contrastive scores. Previous work on document-level MT commonly evaluates document-level MT systems using contrastive scoring (e.g., Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). As a direct comparison, we report the contrastive accuracies on the pronouns test set in Table 3. The trend is often similar to the translation-based metrics in Table 2, however scoring-based improvements are much more pronounced. Our experiments also show that strong contrastive accuracies do not necessarily lead to improvements on the generated hypothesis. For example, on the News task, the contrastive scores of the LM pronoun re-ranking approach and the document-level baseline are similar but their translation-based scores differ strongly (c.f. Table 2).

6.4.3 Computational Cost

We have shown that both the on-the-fly scales and the automatically learned scales improve document-targeted scores over static scores obtained via grid search. Another downside of grid search is that the tuning process is quite expensive. In Table 4, we illustrate that a grid search with 11^3 parameters (as is used in this work) on a single GPU can easily take multiple days. The on-the-fly scales do not

Method	Time	
	Preparation	Search
LM: static	7187 min	5.4 min
LM: on-the-fly	0 min	6.5 min
LM: auto. learned	8.3 min	5.4 min

Table 4: Total time necessary to tune different fusion scale variants on a single GPU, as well as the time spent during translation. We measure the time used to translate the News validation set.

LM	perplexity		contrastive acc.	
	news	e-comm.	pron.	proff.
NewsCrawl	17.0	44.5	62.8	**63.4**
LLaMA	**9.2**	**11.8**	**80.0**	62.3

Table 5: Comparing the small in-domain LM trained on NewsCrawl against the LLM LLaMA.

require any preparation time as they are obtained entirely during search, in which the overhead is small. The automatically scales on the other hand can be learned in just a few minutes and do not have any overhead in decoding.

6.4.4 Large Language Model Integration

Recently, large language models (LLMs) which are trained on large corpora and long context sizes received a lot of attention (e.g., Brown et al., 2020; Touvron et al., 2023). In particular, they have also been able to perform document-level MT (Zhang et al., 2023; Hendy et al., 2023; Karpinska and Iyyer, 2023; Wang et al., 2023). This raises the natural question whether LLMs can improve document-level LM fusion.

We experiment on the News task and compare our own small LM with 35M parameters trained on 2.2B tokens from the in-domain German NewsCrawl corpus against the 13B parameter version of LLaMA (Touvron et al., 2023), which was trained on a total of 1000B tokens. LLaMA's training data includes various domains and languages. Only a small fraction of its data is German. The small LM provides two sentences context while we query the LLM with 200 tokens context. We re-train our MT model and the small LM using the LLaMA tokenizer. This leads to slightly worse performance compared to our previous experiments as the LLaMA tokenization was learned on general-domain English data. For decoding we use a beam size of 4.

Table 5 shows the perplexities of both LMs and their contrastive scores on the document-targeted test sets[7]. Both LMs use the same vocabulary and thus their perplexities are comparable. Because it is in general unclear whether test sets are or are not included in LLM training data, we also include the e-Commerce test set which was translated by ourselves for the purpose of cross-validation. On both test sets, the LLM perplexities are much better than the ones of the small in-domain LM. LLaMA's contrastive scores are also much better on the pronouns test set.

LM Fusion		news		e-Commerce	
LM	Scales	BLEU	COMET	BLEU	COMET
(none)	-	31.2	81.3	13.6	70.5
NewsCrawl	static	33.2	83.0	14.4	72.5
	on-the-fly	33.2	82.8	14.3	72.2
LLaMA	static	**34.6**	**84.2**	**16.5**	**75.0**
	on-the-fly	33.4	83.9	13.7	72.9

Table 6: Comparing fusion with a small LM and a LLM on general test sets.

Table 6 shows the performance of LM fusion with the two LMs in BLEU and COMET. Both LMs notably improve translation, but the LLM translation quality is best. Fusion with LLaMA yields +3.4 % absolute improvements on the in-domain test set. Improvements on the e-Commerce test set are similar, indicating that the gains are not an effect of data leakage of the test set into the training data. While the on-the-fly scales and the static scales perform similarly for the small LM, on-the-fly scales do not perform as well for the LLM.

The improvements measured on the in-domain test sets are likely not because of document-level context but rather due to the increased amount of data. Therefore, we continue our evaluation with the document-targeted scores. Table 8 depicts the results. On these metrics, the LLM outperforms the small LM by an even larger margin. In general, the improvements are correlated to their contrastive scores (c.f. Table 5).

6.5 Extended Analysis on Automatically Learned Fusion Scales

In our experiments we use the validation set of the News task to find the best working methods. We share some insights in the following.

How are the automatically learned scales distributed? Figure 1 shows the distribution of the au-

[7] The professions test set was released without target-side context, which we therefore created ourselves by translating the source-side context with a commercial MT system.

Fusion Scales Learning			λ	valid set		doc.-targeted	
Scales	Crit.	Train Set		BLEU	COMET	pron.	proff.
none	-	-	0.0	24.5	80.9	45.1	65.9
subword-agnostic	grid search	valid set	0.40	25.4	81.8	46.5	65.1
	CE	valid set	0.34	25.5	81.7	46.4	65.0
		synthetic	0.46	25.3	81.6	**47.1**	65.2
subword-dependent	CE	valid set	-	**26.6**	**81.8**	46.1	65.0
		synthetic	-	25.4	81.6	46.9	65.4

Table 7: Automatically learning subword-dependent and -agnostic fusion scales on the News task. We employ the restriction $\lambda_0 := 1$, $\lambda := \lambda_1 = \lambda_2$.

LM Fusion		document-targeted		
LM	Scales	pron.	proff.	form.
(none)	-	44.5	65.7	33.4
NewsCrawl	static	46.3	66.3	34.7
	on-the-fly	47.4	66.7	34.3
LLaMA	static	**51.6**	66.9	**36.1**
	on-the-fly	48.2	**68.9**	35.2

Table 8: Fusion with a small LM against a LLM, reporting the translation-based scores on the document-targeted test sets.

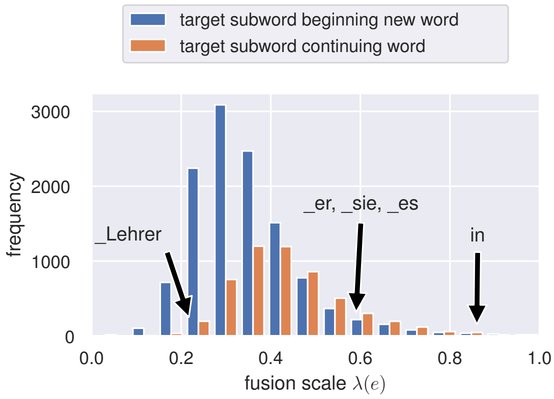

Figure 1: Distribution of the automatically learned LM fusion scales for different target-side subwords on the News task. Subwords for which document-level context is often necessary, such as the German pronouns '_er', '_sie', '_es', and the suffix 'in' marking female professions, have learned higher scales than nouns like '_Lehrer'.

tomatically learned scales for the News task. The learned LM scale of subwords that continue another subword are in general higher than the ones that begin a new word. This is intuitive as continuing a subword is an LM task while beginning a new word requires information about the source sentence.

How much data is needed for automatically learning scales? The static fusion scales are usually tuned on a small validation set via grid search. Table 7 shows that it is also possible to use automatic differentiation to learn static scales only on the validation set. The automatically learned subword-agnostic scales have similar values as the ones tuned via grid search and therefore also their translation performance is similar. Learning subword-dependent scales automatically on the validation set on the other hand improves performance on this set, but does not generalize which indicates overfitting.

7 Conclusions

This work presents multiple extensions to document-level LM fusion, a technique of utilizing document-level monolingual data for context-aware MT. In comparison to existing work, our extensions significantly improve discourse-modeling across four MT tasks and furthermore are computationally more efficient. We conduct evaluations against two baselines: document-level back-translation and a task-specific LM re-ranking method. Despite our extensions, back-translation in general still outperforms document-level LM fusion. Nevertheless back-translation can be effectively combined with LM fusion, further improving translation performance. On very specific test sets, the LM re-ranking performs best. However, our experiments also show that systems trained on document-level parallel data outperform the best systems trained with monolingual documents only.

Finally, this work is the first to explore document-level LM fusion with LLMs. First findings demonstrate that fusion with an LLM outperforms a small LM trained on in-domain data and open the path for future investigations.

Limitations

The experiments in this work were limited to four MT tasks, from which two are low-resource and three are translating from English into German. Apart from the experiments with the LLM, we did not conduct any experiments on a large-scale dataset of multi-domain monolingual documents. The LLM in our experiments only has 7B parameters, while much larger LLMs exist (e.g., Touvron et al., 2023).

Further, our work focuses only on one specific architecture for document-level MT and uses only two sentences target-side context. Various other architectures exist and may entail different properties. This work further does not investigate the behavior of larger translation models.

Another limitation lies in the evaluation of document-level MT models. The document-level targeted metrics we used are all reference-based and limited to the translation of pronouns, gender-referring professions or salutation forms. Other discourse phenomena like e.g. cohesion exist (Maruf et al., 2022) but were not studied in our work. It is unclear how well automated metrics actually correlate with the actual document-level translation quality (Currey et al., 2022), and this work did not perform any qualitative analysis.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4555–4567. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 526–532. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation, IWSLT 2017, Tokyo, Japan, December 14-15, 2017*, pages 2–14. International Workshop on Spoken Language Translation.

Linqing Chen, Junhui Li, Zhengxian Gong, Boxing Chen, Weihua Luo, Min Zhang, and Guodong Zhou. 2021a. Breaking the corpus bottleneck for context-aware neural machine translation with cross-task pre-training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2851–2861. Association for Computational Linguistics.

Linqing Chen, Junhui Li, Zhengxian Gong, Xiangyu Duan, Boxing Chen, Weihua Luo, Min Zhang, and Guodong Zhou. 2021b. Improving context-aware neural machine translation with source-side monolingual documents. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3794–3800. ijcai.org.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods* *in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4287–4299. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel L. Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, MT-Summit 2019, Dublin, Ireland, August 19-23, 2019*, pages 118–119. European Association for Machine Translation.

Thierry Etchegoyhen and Harritxu Gete. 2020. To case or not to case: Evaluating casing methods for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3752–3760. European Language Resources Association.

Eva Martínez Garcia, Carles Creus, and Cristina España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 13–23. Association for Computational Linguistics.

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1296–1302. Association for Computational Linguistics.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023a. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse: CODI 2023, Toronto, Canada, July 13-14, 2023*. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023b. On search strategies for document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 604–616. Association for Computational Linguistics.

Sébastien Jean, Ankur Bapna, and Orhan Firat. 2019. Fill in the blanks: Imputing missing sentences for larger-context neural machine translation. *CoRR*, abs/1910.14075.

Sébastien Jean and Kyunghyun Cho. 2020. Log-linear reformulation of the noisy channel model for document-level neural machine translation. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP@EMNLP 2020, Online, November 20, 2020*, pages 95–101. Association for Computational Linguistics.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. Blonde: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1550–1565. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 225–233. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *CoRR*, abs/2304.03245.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Shaohui Kuang and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 607–617. Association for Computational Linguistics.

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.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

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. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

António V. Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. 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 1275–1284. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2947–2954. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.

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. *CoRR*, abs/1904.01038.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

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 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Fahimeh Saleh, Alexandre Berard, Ioan Calapodescu, and Laurent Besacier. 2019. Naver labs europe's systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 273–279. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6490–6500. Association for Computational Linguistics.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. Cued@wmt19: Ewc&lms. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 364–373. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 35–44. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2020. Context-aware decoder for neural machine translation using a target-side document-level language model. *CoRR*, abs/2010.12827.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5781–5791. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 82–92. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 877–886. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. 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 1264–1274. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *CoRR*, abs/2304.02210.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Trans. Assoc. Comput. Linguistics*, 8:346–360.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *CoRR*, abs/2301.07069.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 533–542. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendix

A.1 Model Training

Data. The *News En→De* task comprises 330k parallel sentences from NewsCommentary v14[8], which we combine with document-level monolingual data from NewsCrawl[9] (70M sentences[10]). Our *Subtitles En→De* data consists of a total of 39M monolingual movie show subtitles from OpenSubtitles, from which a subset of 22.5M sentences has been aligned to English sentences and forms our parallel training data (Lison et al., 2018). For *TED En→It* we use 230k parallel sentences from scientific TED talks released as part of the IWSLT17 multilingual task (Cettolo et al., 2017) which we combine with 2.2M sentences of talks from the European parliament (Koehn, 2005). Finally, the *e-Commerce En→De* task is about translating item descriptions from e-Commerce listings. We use 326M parallel sentences of out-of-domain parallel training data from the ParaCrawl v9 corpus (Esplà-Gomis et al., 2019) which we combine with 128k parallel sentences in-domain data. The monolingual data was sampled from item descriptions and is entirely in-domain (119M sentences).

The sizes of our training corpora are shown in Table 10.

On each task, we use a validation set for selecting the best checkpoint, tuning the fusion scales and for finding which method works best. For the final comparison in Table 1 we then report on an unseen test set of the same domain.

The News validation set is `newstest2015`, and `newstest2018` as test set. For Subtitles, our validation and test sets were sampled from the training corpus. The precise document IDs for the validation set are: 1995/254, 1997/165, 2000/313, 2002/461, 2005/441, 2007/781, 2010/273, 2012/757, 2015/1488, 2017/525 for the validation set; and for our test set: 1997/310, 2002/40, 2007/189, 2012/1085, 2017/644. The test set is the same as used in Huo et al. (2020). For TED, we concatenate `dev2010` and `tst2010` and use `tst2017.mltlng` as test set. For e-Commerce, we create the validation and test set ourselves by translating English e-Commerce item descriptions into German: Our validation set comprises 85 documents (2882 sentences) and the test set 100

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

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

[10] To reduce training time, our back-translation experiments on this task utilize only the first 2M sentences.

Data	News			Subtitles			TED		e-Commerce		
	test	pron.	proff.	test	pron.	proff.	test	proff.	test	pron.	proff.
parallel data	129.7	125.7	168.6	**26.6**	**36.0**	**103.2**	**47.6**	**114.8**	61.7	**44.2**	**52.6**
monolingual data	**97.1**	**94.9**	**161.3**	27.8	36.6	117.4	75.4	116.1	**50.8**	48.7	57.5

Table 9: Perplexities of sentence-level LMs trained on equal amount of target-side data.

Task	Data	docs	sents	words
News	parallel	8.5k	330k	7.4M
	mono.	3M	70M	1.0B
Subtitles	parallel	30k	22.5M	136M
	mono.	47k	39M	223M
TED	parallel	1.9k	230k	3.7M
	mono.	6k	2.2M	54.6M
e-Commerce	parallel	n.a.	326M	9.6B
	mono.	1.5M	119M	3.1B

Table 10: Training data statistics.

Approach	valid set		doc.-targeted	
	BLEU	COMET	pron.	proff.
baseline	24.5	80.9	45.1	**65.9**
LM fusion	24.8	81.2	45.0	65.8
+ (Jean, 2020; Sugiyama, 2021)[5]	24.9	81.2	46.0	65.0
+ ILM: separate	**25.8**	**82.1**	**47.3**	65.2
+ ILM: $h = 0$	25.5	82.0	43.8	64.7
+ ILM: mini self-att.	**25.8**	**82.1**	44.6	65.1

Table 11: Document-level LM fusion (a) without subtracting any LM, (b) subtracting the sentence-level probabilities of the external LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021), and (c) subtracting different approximations of the internal LM (ILM) learned by the MT model, reported on the News task.

documents (2520 sentences).

As the pronouns test set (Müller et al., 2018) was extracted from the OpenSubtitles corpus, we remove these sentences from the Subtitles training data. The professions test set (Currey et al., 2022) was curated from Wikipedia articles and is not part of our training corpora.

Models. We train the News, Subtitles and TED models with a shared embedding and projection matrix. Th resulting MT models for News and Subtitles have 60M parameters, 51M parameters for TED and 90M for e-Commerce. For model training we use eight Tesla V100-SXM2-32GB GPUs. Training the baselines takes approximately 7h for News, 21h for Subtitles, 5h for TED, and 30h for e-commerce. Due to resource constraints, we report only a single run for each experiment.

Optimization. For optimization we use Adam (Kingma and Ba, 2015) and a batch size of 22k subwords. The low-resource MT models (News, TED) are trained for 100k update steps with 30 % dropout, 20 % label smoothing and weight decay, while the high-resource models (Subtitles, e-Commerce) are trained for 300k updates with 10 % dropout, 10 % label smoothing and no weight decay.

A.2 Domain Effects

In an effort to estimate how well the domain of the training data matches the test sets, we train LMs on the target-side part of the parallel and the monolingual training data. Within each task, the LMs are trained with the same parameters and the same vocabulary. We then report the perplexities on the task-specific test sets and the document-targeted challenge sets in Table 9.

For News, the monolingual data is more in-domain for all test sets. Similarly the domain of the e-Commerce monolingual data is closer to the task-specific test set. For Subtitles, the domains of parallel and monolingual data are more or less equal and on TED, the monolingual data is slightly out-of-domain.

This domain effect explains the improvements in BLEU and COMET on the task-specific test sets that we reported in Table 1 on News and on e-Commerce.

A.3 Comparing Internal Language Model Estimations

Herold et al. (2023) propose several ways of approximating the internal LM learned implicitly by the MT model in the context of sentence-level MT. We evaluate three of their approaches for document-level LM fusion and compare them against the existing document-level LM fusion approach that subtracts the sentence-level probabilities of the external LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). Table 11 shows the results: Subtracting the internal LM substantially improves LM fusion over existing work. Estimating it by training a separate LM on the same data as the MT model works best.

Fusion Scales		valid set		doc.-targeted	
Approach	Restriction	BLEU	COMET	pron.	proff.
none	-	24.5	80.9	45.1	65.9
static	-	**25.8**	**82.1**	47.3	65.2
	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.4	81.8	46.5	65.1
on-the-fly	-	22.3	78.3	43.4	69.3
	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.6	81.8	**48.0**	65.5
auto.	-	24.8	80.7	44.7	**69.4**
learned	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.3	81.5	46.7	64.9

Table 12: LM fusion with an imposed restriction on the search space of the fusion scales $\lambda_0, \lambda_1, \lambda_2$, reported on the News task.

Data		valid set		doc.-targeted	
parallel	mono.	BLEU	COMET	pron.	proff.
sent.	-	24.5	80.9	45.1	65.9
sent.	sent.	27.0	**83.2**	46.7	65.7
sent.	doc.	26.9	82.4	47.8	**80.7**
pseudo-doc.	doc.	**27.1**	83.0	**48.7**	80.5

Table 13: Effect of back-translation on the News task.

A.4 Fusion Scale Restrictions

The three LM fusion scales $\lambda_0, \lambda_1, \lambda_2$ in Equation 1 balance the contribution of the MT model and the two LMs. In our experiments the optimal scales usually lie at $\lambda_0 \approx 1$ and $\lambda_1 \approx \lambda_2$. This is plausible as the internal LM (λ_2) should neutralize the external LM (λ_1) to the same degree. For the non-static fusion scales however, we find that searching over the three-dimensional search space of independent $\lambda_0, \lambda_1, \lambda_2$ finds unintuitive scale combinations and that this causes bad performance. Therefore in our experiments we restrict the search space of fusion scales to the one-dimensional slice where $\lambda_0 = 1$ and $\lambda_1 = \lambda_2$. Table 12 gives a direct comparison.

A.5 Document-level Back-translation

In Table 13 we compare back-translation using document-level data against sentence-level back-translation.

Sentence- and document-level back-translation gives the same performance improvements in BLEU and COMET, however only back-translation on document-level improves the document-targeted metrics. For document-level back-translation we find that creating pseudo-documents from the parallel data is necessary to achieve the same BLEU and COMET scores as sentence-level back-translation.