# Multilingual Document-Level Translation Enables Zero-Shot Transfer From Sentences to Documents

Anonymous ACL submission

## Abstract

Document-level neural machine translation (DocNMT) achieves coherent translations by incorporating cross-sentence context. However, for most language pairs there's a shortage of parallel documents, although parallel sentences are readily available. In this paper, we study whether and how contextual modeling in DocNMT is transferable via multilingual modeling. We focus on the scenario of zero-shot transfer from *teacher* languages with document level data to *student* languages with no documents but sentence level data, and for the first time treat document-level translation as a transfer learning problem. Using simple concatenation-based DocNMT, we explore the effect of 3 factors on the transfer: the number of teacher languages with document level data, the balance between document and sentence level data at training, and the data type of document level data (genuine vs. back-translated). Our experiments on Europarl-7 and IWSLT-10 show the feasibility of multilingual transfer for DocNMT, particularly on document-specific metrics. We observe that more teacher languages and adequate data schedule both contribute to better transfer quality. Surprisingly, the transfer is less sensitive to the data type, where multilingual DocNMT delivers decent performance with either back-translated or genuine document pairs.

## 1 Introduction

Recent years have witnessed a trend moving from sentence-level neural machine translation (SenNMT) to its document-level counterpart (DocNMT). SenNMT inevitably suffers from translation errors related with document phenomena (Maruf et al., 2021) and delivers obviously inferior performance when compared against human translations and evaluated at a document level (Läubli et al., 2018). Most efforts on DocNMT aim at improving contextual modeling via dedicated model architectures and/or decoding algorithms (Bawden et al., 2018; Voita et al., 2019; Chen et al., 2020) and heavily rely on large-scale parallel document resources. Nevertheless, document resources are unevenly distributed across language pairs, with most pairs having little to no such resources.[1]

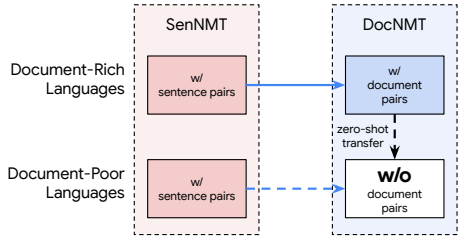


Figure 1: Overview of the focused zero-shot problem for DocNMT. We study transferring contextual modeling from document-rich (teacher) languages to document-poor (student) languages, where only sentence pairs are given for students. The transfer occurs in a multilingual setup, shown by the dashed rectangles. Dashed arrows show the transfer direction.

One promising way to accommodate languages with varied training data is multilingual modeling, as demonstrated in multilingual SenNMT (Firat et al., 2016; Johnson et al., 2017). By sharing parameters across languages, multilingual modeling encourages cross-lingual knowledge transfer, enabling performance improvement and even zero-shot transfer (Aharoni et al., 2019; Arivazhagan et al., 2019b; Zhang et al., 2020). In the context of translation, however, most studies on multilingual transfer center around SenNMT, seldom going beyond sentence-level translation. So far, the question of whether and how document-level contextual modeling can be learned cross-lingually in multilingual DocNMT is still unanswered.

In this paper, we study *zero-shot generalization* for DocNMT – the ability to attain plausible DocNMT quality for some focused (*student*) language pair(s) with only parallel sentences for the student, but parallel documents for other (*teacher*) languages in the multilingual mix. The high-level

[1] Note that we use *language* and *language pair* interchangeably since one side of our parallel data is always English.

research question we seek to answer is in Figure 1.

We resort to transfer learning via multilinguality to leverage document resources in teacher languages to help the student languages. We perform our analysis using simple concatenation based DocNMT with consecutive sentences chained into one sequence for translation. We extensively investigate three dimensions to understand the transfer in multilingual DocNMT: 1) the number of languages with document level data (teacher languages), where we simplify our transfer setup to contain either only one teacher language (with N students) or N teachers (with one student); 2) the data balance (schedule) for parallel documents, i.e. manipulating the ratio of document-level data to sentence-level data during training; and 3) the data type (condition) of parallel documents, where we adopt back-translated parallel documents when only monolingual documents are given in teacher languages or use genuine parallel documents crawled natively.

We conduct experiments on two publicly available datasets, namely Europarl-7 and IWSLT-10, covering 6 and 9 languages from/to English respectively. We analyze one-to-many (En→Xx) and many-to-one (Xx→En) translation scenarios separately. Following recent work (Ma et al., 2021), we adopt document-specific metrics for evaluation apart from BLEU and support our findings with human evaluations. We also propose a pronoun F1 metric (targeted at gendered pronouns: he/she) for Xx→En translation, and employ accuracy on contrastive test sets (Bawden et al., 2018; Müller et al., 2018) for En→Xx translation. Our main findings:

- Zero-shot transfer from sentences to documents is feasible through multilingual DocNMT modeling, particularly when evaluated with document-specific metrics. This is partially supported by human evaluation.
- Transfer quality is strongly affected by the number of teacher languages that use document level data and the data schedule for documents. Higher quality is achieved with more teacher languages and adequate document schedule, where the optimal schedule varies across scenarios.
- Surprisingly, transfer via back-translated documents performs comparable to transfer via genuine parallel documents.
- Zero-shot transfer from high-resource document level languages to low-resource sentences level ones is easier, resulting in better quality compared to other scenarios.

## 2 Related Work

**Document-level MT** Integrating document-level information meaningfully into NMT is a challenging task and has inspired research not only on exploring advanced context-aware neural architectures, including simple concatenation-based models (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Lopes et al., 2020), multi-source models (Jean et al., 2017; Bawden et al., 2018; Zhang et al., 2018), hierarchical models (Miculicich et al., 2018; Zheng et al., 2020; Chen et al., 2020), multi-pass models (Voita et al., 2019; Yu et al., 2020; Mansimov et al., 2020) and dynamic context models (Kang et al., 2020), to name a few. But it has also motivated the field to revisit the common protocols resorted for evaluation (Freitag et al., 2021). Despite the hard to measure success, all the above mentioned methods implicitly assume an abundance of document resources and overlook the data scarcity problem. In this study, we adopt the simple concatenation model as our experimental protocol, and leave the exploration of various input formatting options and modelling to future work. Considering the fast changing landscape of the (contextual) MT evaluation, we also provide multiple evaluation metrics including human evaluations, to give a full picture of the phenomena under investigation, while acknowledging the current imperfections of and disagreements on the right way of evaluating MT systems (Kocmi et al., 2021).

**Zero-Shot Transfer via Multilinguality** Multilingual modeling often clusters sentences of similar meaning from different languages within a shared semantic space (Kudugunta et al., 2019; Siddhant et al., 2020). Such representation space hypothesized to enable zero-shot transfer, delivering improved performance in many cross-lingual tasks (Eriguchi et al., 2018; Hu et al., 2020; Chi et al., 2021; Ruder et al., 2021), especially based on large-scale pretrained multilingual Transformers (Devlin et al., 2019; Conneau and Lample, 2019; Xue et al., 2021). In the context of translation, multilingual SenNMT successfully achieves zero-shot translation, transferring sentence-level generation knowledge to language pairs unseen during training (Firat et al., 2016; Johnson et al., 2017; Gu et al., 2019; Arivazhagan et al., 2019a) even in massively multilingual settings (Aharoni

et al., 2019; Arivazhagan et al., 2019b; Zhang et al., 2020). Our study extends multilingual SenNMT to multilingual DocNMT and aims at document-level knowledge transfer from languages that have document level data to languages that only have sentence level data. To the best of our knowledge, our study is the first demonstrating the emergence of document-level zero-shot transfer across languages for multilingual machine translation.

# 3 Zero-Shot Transfer in Multilingual DocNMT

We first formulate the *zero-shot generalization* framework explored in this paper. Given **N+1** language pairs, we assume that all of them have parallel sentences for training, but only some of them have parallel documents (teachers). Through multilingual training, we study to what degree contextual modeling in document-supervised DocNMT can be transferred to those document-poor (student) languages as in Figure 1. Any form of parallel document for student languages is disallowed at training, ensuring that the transfer is measured zero-shot.

## 3.1 Multilingual DocNMT

We employ the concatenation-based method with $D2D$ structure for DocNMT, where $D$ consecutive sentences in a document are concatenated into one sequence for translation (Junczys-Dowmunt, 2019; Sun et al., 2020). Sentence boundary is indicated by a special symbol "[SEN]". We adopt the language token method (Johnson et al., 2017) for multilingual DocNMT, using source and target language token for Xx→En and En→Xx translation respectively. Instead of appending this token to the source sequence, we add its embedding to each source word embedding to strengthen the language signal in a document translation setting.

For training, we adopt a two-stage method where we first pretrain a multilingual SenNMT followed by finetuning on document data to obtain multilingual DocNMT. Our analysis requires training a large number of DocNMT models, and the two-stage method saves substantial amounts of computation by sharing the pretrained SenNMT. For evaluation, we distinguish sentence-level (SenInfer) with document-level (DocInfer) inference. SenInfer translates sentences separately (out of context), while DocInfer translates $D$ consecutive and non-overlapping sentences in context with each other.[2]

## 3.2 Zero-Shot Setup

We explore three factors for the zero-shot transfer:

- **The number of teacher languages** The source of the transfer comes from teacher languages. Intuitively, both the number of teacher languages and their relevance to student language(s) affect the transfer result. However, exhaustively exploring all possible teacher-student combinations in a multilingual setting will lead to a large search space that expands exponentially with respect to the total number of languages involved. Instead, we simplify our study by exploring two extreme transfer settings, namely **N21** and **12N** transfer. The first setting uses N teachers that incorporate document level data with 1 student having sentence level data only, while the second setting has 1 teacher and N students. Note that in either N21 or 12N transfer, there exist N teacher-student configurations, and we report average results over them.[3]
- **The data schedule for parallel documents** When varying the number of teacher languages, the proportion of document data at training also changes. Such imbalance could deeply affect transfer (Arivazhagan et al., 2019b). To offset this effect, we include the data schedule for analysis by controlling the sampling ratio $p$ of documents from 0.1 to 0.9 with a step size of 0.1. Note $p$ is for documents in *all* teacher languages, and the relative proportion among teachers is always retained.
- **The data condition of parallel documents** We also study when teacher languages have no parallel documents but only monolingual ones. Methods utilizing monolingual documents for DocNMT vary greatly. Following recent work (Sugiyama and Yoshinaga, 2019; Huo et al., 2020), we adopt back-translation (BT) to construct pseudo parallel documents. Note that, for teacher languages, we replace all sentence training data with pseudo documents rather than mixing them (according to our empirical results in Appendix C).

# 4 Experimental Settings

**Datasets** We conduct experiments on two public datasets: Europarl-7 and IWSLT-10. Europarl-7

[2] At decoding phase, the last chunk in a source document can have $< D$ sentences for DocInfer.

[3] Note we also include transfer results to individual languages (German and French) in Appendix D.

is extracted from European Parliament (v10) and have translations between English and $N$=6 different languages, including *Czech, German, Finnish, French, Lithuanian and Polish* (Koehn, 2005). This dataset offers sentence-aligned parallel documents (0.9K∼3.7K documents, 190K∼1.9M sentences) and also monolingual documents (9.7K∼11K documents, 0.65M∼2.28M sentences) for training. For evaluation, we use the latest WMT evaluation sets (dev and test set) (Barrault et al., 2020) available for each language pair. In contrast, IWSLT-10 is collected from TED talks and covers translations between English and $N$=9 different languages, including *Arabic, German, French, Italian, Japanese, Korean, Dutch, Romanian and Chinese* (Cettolo et al., 2017). Unlike Europarl-7, the distribution of training data over languages in IWSLT-10 is much smoother (uniform). There are ∼1.9K sentence-aligned parallel documents with ∼240K sentences for each language pair. We further collected about 1K TED talks for each language pair (crawled from Feb 2018 to Jan 2021) as monolingual documents. We use IWSLT17 dev and test sets for evaluation. Detailed statistics are given in Appendix A. We preprocess all texts with the byte pair encoding (BPE) algorithm (Sennrich et al., 2016). We use the *sentencepiece* toolkit (Kudo and Richardson, 2018), and set the vocabulary size to 32K and 64K for IWSLT-10 and Europarl-7, respectively.

**Model Details** We use the Transformer-base model (Vaswani et al., 2017) for experiments with 6 encoder/decoder layers, 8 attention heads and a model dimension of 512/2048. We set $D = 5$ for DocNMT. We use Adam (Kingma and Ba, 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.98$) for parameter update with a learning rate warmup step of 4K and label smoothing rate of 0.1. We apply dropout to residual connections and attention weights with a rate of 0.5 and 0.2, respectively. Other training and decoding details are given in Appendix B.

**Back-Translation** Some of our models are trained using back-translated monolingual documents. Back-translations are obtained using bilingual SenNMT (independently for Europarl-7 and IWSLT-10). To train these models, we halve the BPE vocabulary size as well as the training steps. All other settings are kept as mentioned above.

**Evaluation** Following previous work, we use BLEU (Post, 2018)[4] to measure the general translation quality. Document-level BLEU is calculated by counting n-gram at the document level instead of at the individual sentence level (Sun et al., 2020).

Measuring improvements to document phenomena in translation automatically remains challenging and oftentimes simple surface-based metrics such as BLEU (Läubli et al., 2018) are not sensitive enough. Therefore, we evaluate our model on test sets that focus on such document phenomena. We use the contrastive test sets for En-De (Müller et al., 2018) and En-Fr (Bawden et al., 2018) which measure a model's ability to distinguish correct from incorrect anaphoric pronoun translations. We include 4 and 1 additional context sentences for En-De and En-Fr contrastive evaluation, respectively.

Gender bias in translation models has attracted much attention recently (Kuczmarski and Johnson, 2018; Saunders and Byrne, 2020). We expect that contextual information can help to alleviate it. To this end, we introduce gendered pronoun F1 based on the following precision and recall scores to evaluate English translations:

$$\begin{aligned} \text{Precision} &= \frac{\sum_{i,g\in\mathcal{G}} \min(C^g_{\mathbf{r}_i}, C^g_{\mathbf{h}_i})}{\sum_{i,g\in\mathcal{G}} C^g_{\mathbf{h}_i}}, \\ \text{Recall} &= \frac{\sum_{i,g\in\mathcal{G}} \min(C^g_{\mathbf{r}_i}, C^g_{\mathbf{h}_i})}{\sum_{i,g\in\mathcal{G}} C^g_{\mathbf{r}_i}} \end{aligned} \tag{1}$$

where $\mathbf{r}_i$ and $\mathbf{h}_i$ denotes the $i$-th gold reference and hypothesis sentence respectively, comprising the gendered pronouns of interest $\mathcal{G}$[5]. $C^g_{\mathbf{x}}$ denotes the count of pronoun $g$ in sentence $\mathbf{x}$.

Finally, we conduct human evaluation to verify the performance delivered by zero-shot transfer. For each source language, we sampled 50 test documents which were translated into the target language using the corresponding models and decoding techniques. The translated documents were presented to bilingual human raters who are native in the non-English locale. The raters were asked to evaluate translation qualities while taking the full source document context into account. The raters assign a score in a 0-6 scale to every sentence-translation pair in the document, where 0 and 6 mean nonsense and perfect translations, respectively. For each model, the scores were aggregated across the entire test corpus and the average scores were reported. To ensure a fair diversity of ratings, each rater has rated no more than 6 documents per model; an average of 18 raters evaluated each model independently.

[4]Signature: BLEU+c.mixed+#refs.1+s.exp+tok.13a+v.1.4.14

[5]*he, his, him, himself, she, her, hers, herself.*

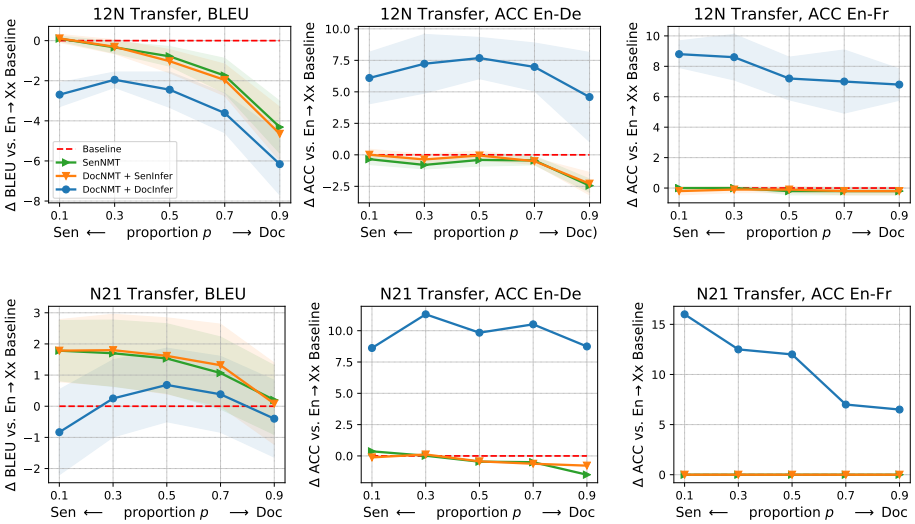


Figure 2: Average BLEU and accuracy (ACC) (En-De, En-Fr) in N21 (bottom) and 12N (top) transfer settings for En→Xx translation on Europarl-7. Shadow areas denote the standard deviation. $p$ indicates the proportion of documents, and "$p = 0$": training with student/sentence data alone (*Sen*). "*Baseline*": multilingual SenNMT with SenInfer trained with raw training data. "*SenNMT*": the same as Baseline but its training data is scheduled with $p$ at sentence level.

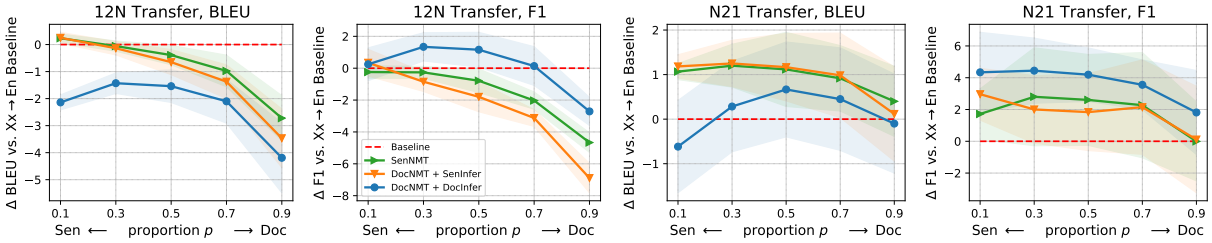


Figure 3: Performance of N21 and 12N transfer as a function of proportion $p$ for Xx→En translation on Europarl-7.

| Model | Xx→En | En→Xx |
|---|---|---|
| SenNMT w/ SenInfer | 22.40 | 18.82 |
| SenNMT w/ DocInfer ($D = 2$) | -2.98 | -4.05 |
| SenNMT w/ DocInfer ($D = 5$) | -11.7 | -13.0 |

Table 1: Average BLEU on Europarl-7 for multilingual SenNMT with SenInfer and DocInfer. Rows 2 and 3 represent deltas compared to their Row 1 counterpart. Directly applying DocInfer to SenNMT performs poorly.

## 5 Results and Analysis

**Does SenNMT have the capability of leveraging context?** *Not really!* We put our major analysis on Europarl-7 (N=6, all European languages). Before diving deep into the transfer, we start with analyzing whether SenNMT models trained on sentences alone could generalize to contextual translation. If multilingual SenNMT can be directly used for DocInfer, studying zero-shot transfer would be meaningless. Results in Table 1 challenge this possibility: SenNMT results in large quality reduction with DocInfer. We observe that SenNMT produces significantly shorter translations under DocInfer, preferring to translate the first few input sentences. We ascribe such failures to the poor generalization to documents from sentence-level training.

**Impact of the data schedule and the number of teacher languages on zero-shot transfer** Figure 2 and 3 summarize the results for En→Xx and Xx→En translation, respectively, where we report the average performance paired with the standard deviation over $N$ configurations.[6] Overall, the document-level zero-shot transfer is achievable via multilingual modeling. Transfer-based DocNMT could successfully identify and translate the correct number of input sentences for student languages.

[6]Note the average results are for transfer directions, not the supervised ones. Each experiment in N21 transfer has only one transfer direction, so we directly report the average over N settings; by contrast, in 12N transfer, we have N transfer directions, where we first perform average over these N transfer results followed by another average over N settings. Also note, the average results contains transfer from high/low and similar/distant languages.

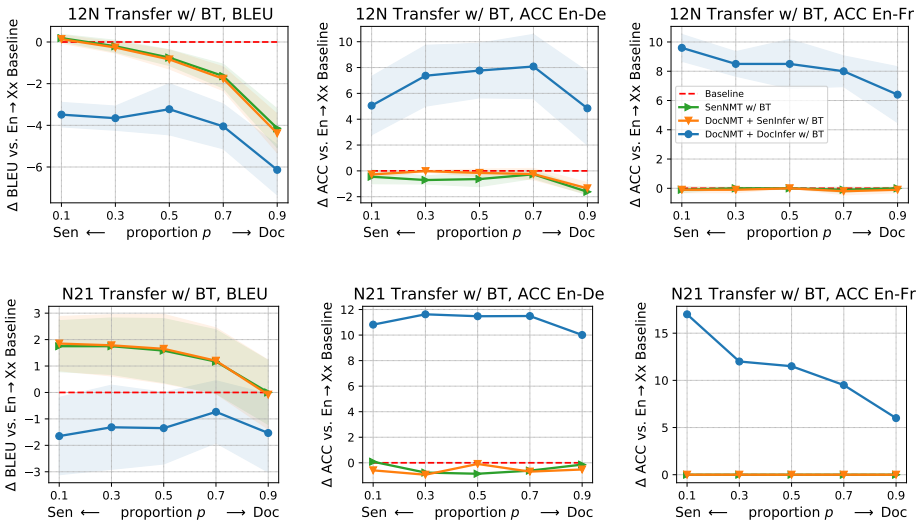


Figure 4: Performance of N21 and 12N transfer with back-translated (BT) documents for En→Xx translation on Europarl-7.

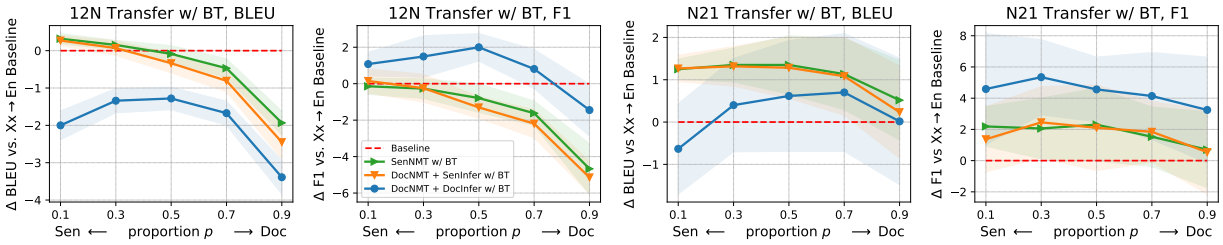


Figure 5: Performance of N21 and 12N transfer with back-translated (BT) documents for Xx→En translation on Europarl-7.

With a proper sampling ratio for document-level data, student DocNMT yields better performance than its SenNMT counterpart, especially shown by document-specific evaluations (F1 and ACC).

*Increasing teacher languages improves transfer.* In En→Xx and Xx→En translation, we find that N21 transfer performs consistently better than 12N transfer on all metrics. This is reasonable since N21 transfer has $N$ teacher languages, offering richer and more informative sources for transfer.

*Balancing between document and sentence data matters for transfer.* We also observe that performance changes over the document proportion on all metrics in both 12N and N21 transfer. Applying more or fewer documents during training often hurts zero-shot transfer, indicating a trade-off. Roughly, setting $p$ to 30%∼50% delivers good performance (Figure 2 and 3), although the optimal proportion depends.

*SenInfer underperforms DocInfer on document-specific metrics.* DocNMT w/ SenInfer performs similarly to SenNMT, and better than DocInfer on BLEU. When evaluating document phenomena, however, SenInfer shows clear insufficiency. This resonates with the findings of Ma et al. (2021).

**Can we achieve zero-shot transfer with monolingual documents?** *Yes.* We next repeat our experiments with BT document pairs. Figure 4 and 5 show that BT performs surprisingly well on document-level zero-shot transfer. We observe almost the same performance pattern compared to training with genuine documents in all settings (En→Xx and Xx→En, N21 and 12N transfer and different metrics), although BLEU scores become worse and the optimal proportion also changes. We argue that the target-side genuine context information in BT documents helps contextual modeling (Ma et al., 2021). These results are promising, encouraging further research on exploring monolingual documents for multilingual DocNMT.

**Impact of high/low-resource languages on zero-shot transfer.** The data distribution of Europarl-7 is highly skewed over languages, with *Cs, Lt, Pl* being relatively low-resource languages while *De, Fi, Fr* being high-resource ones. Studies on multilingual SenNMT have witnessed the transfer from high-resource to low-resource languages (Aharoni

| Xx→En | BLEU | | | | F1 | | | |
|---|---|---|---|---|---|---|---|---|
| | High→ | Low→ | →High | →Low | High→ | Low→ | →High | →Low |
| DocNMT + 12N transfer | **-1.07** | -1.79 | -1.71 | **-1.15** | **+1.73** | +0.95 | +0.73 | **+1.95** |
| w/ BT | **-1.19** | -1.37 | -1.59 | **-0.97** | +1.61 | **+2.38** | +1.36 | **+2.64** |

| En→Xx | BLEU | | | | ACC En-De | | ACC En-Fr | |
|---|---|---|---|---|---|---|---|---|
| | High→ | Low→ | →High | →Low | High→ | Low→ | High→ | Low→ |
| DocNMT + 12N transfer | **-1.85** | -2.05 | -2.03 | **-1.87** | **+8.67** | +6.27 | **+10.25** | +7.50 |
| w/ BT | **-3.29** | -4.03 | -4.39 | **-2.93** | **+8.55** | +6.57 | **+6.61** | +6.19 |

Table 2: Relative performance to multilingual SenNMT baseline when transferring from and into high-resource (High) and low-resource (Low) languages for En→Xx and Xx→En translation on Europarl-7. Results are for DocNMT with DocInfer under 12N transfer; "High/Low→": average results for transferring from high-/low-resource teacher languages to all student languages; "→High/Low": average results for transferring from all teacher languages to high-/low-resource student languages. Transferring from high-resource teacher languages (High→) and transferring into low-resource student languages (→Low) show better performance, highlighted in **bold**.

| Dataset | Models | En→Xx | | | Xx→En | |
|---|---|---|---|---|---|---|
| | | BLEU | ACC En-De | ACC En-Fr | BLEU | F1 |
| Europarl-7 | SenNMT (Baseline) | 18.82 | 52.14 | 50.00 | 22.40 | 53.81 |
| | DocNMT ‡ w/ SenInfer | -0.07 | -0.15 | 0.00 | -0.20 | +0.67 |
| | DocNMT ‡ w/ DocInfer | +0.38 | +14.74 | +14.50 | +0.15 | +2.35 |
| | N21 Transfer ($p = 0.3$) w/ DocInfer | +0.25 | +11.31 | +12.50 | +0.28 | +4.44 |
| | 12N Transfer ($p = 0.3$) w/ DocInfer | -1.95 | +7.23 | +8.60 | -1.43 | +1.34 |
| | N21 Transfer + BT ($p = 0.3$) w/ DocInfer | -1.32 | +11.62 | +12.00 | +0.40 | +5.35 |
| | 12N Transfer + BT ($p = 0.5$) w/ DocInfer | -3.23 | +7.77 | +8.50 | -1.28 | +2.00 |
| IWSLT-10 | SenNMT (Baseline) | 25.39 | 40.39 | 50.00 | 29.41 | 65.37 |
| | DocNMT ‡ w/ SenInfer | -0.77 | +3.19 | 0.00 | +0.26 | +2.42 |
| | DocNMT ‡ w/ DocInfer | -0.09 | +15.34 | +8.00 | +0.51 | +4.52 |
| | N21 Transfer ($p = 0.3$) w/ DocInfer | +0.01 | +15.74 | +15.00 | +1.10 | +4.40 |
| | 12N Transfer ($p = 0.5$) w/ DocInfer | -3.43 | +5.56 | +5.13 | -1.14 | +1.67 |
| | N21 Transfer + BT ($p = 0.3$) w/ DocInfer | -1.11 | +13.28 | +18.00 | +1.53 | +3.85 |
| | 12N Transfer + BT ($p = 0.5$) w/ DocInfer | -5.32 | +4.19 | +4.88 | -1.71 | +1.58 |

Table 3: Performance of different models on Europarl-7 and IWSLT-10. ‡: multilingual DocNMT trained on parallel documents from all language pairs. For 12N and N21 transfer, we report one group of results under the approximately optimal proportion $p$. Notice that the results for transfer settings are averaged over different teacher-student configurations, while those for DocNMT ‡ are for one model. We report absolute scores for SenNMT but relative scores for the others.

et al., 2019; Zhang et al., 2020). We next analyze how this data scale difference affects document-level zero-shot transfer. We mainly explore 12N transfer because of the single transfer source, avoiding interference from other teacher languages.

Table 2 lists the results. Regardless of the data condition (genuine or BT document pairs), transferring from high-resource teacher languages often outperforms that from low-resource ones. Besides, transferring into low-resource student languages delivers better transfer than into high-resource ones. These suggest that increasing the document data for teacher languages benefits zero-shot transfer.

Note we also provide transfer results from individual languages to De and Fr in Appendix D.

**Performance on Europarl-7 and IWSLT-10** We summarize the main results on both datasets in Table 3. Although IWSLT-10 (N=9) includes more (distant) languages and distributes quite differently over languages, the results on IWSLT-10 resemble those on Europarl-7. On both datasets, we observe that transfer, both 12N and N21, yields very positive results, particularly with document-specific metrics. Unlike Europarl-7, BT-based transfer performs much worse than models trained on genuine document pairs on IWSLT-10. We ascribe this to the data scarcity, where only very small-scale monolingual documents are used for BT in IWSLT-10. This also reinforces our observation that more document resources benefits zero-shot transfer.

## 6 Discussion

Apart from automatic evaluation, we also offer human evaluation on En-De. We choose En-De as

| Models | Human Rating (↑) |
|---|---|
| Reference | **4.96** |
| SenNMT (Baseline) | 3.31 |
| DocNMT ‡ w/ SenInfer | 3.60 |
| DocNMT ‡ w/ DocInfer | **3.84** |
| N21 Transfer w/ DocInfer | 3.46 |
| 12N Transfer w/ DocInfer | 2.78 |
| N21 Transfer + BT w/ DocInfer | 3.18 |
| 12N Transfer + BT w/ DocInfer | 2.72 |

Table 4: Document-level human ratings (↑) for En-De on Europarl-7. We evaluate the best system indicated by document-level metrics (ACC En-De) for 12N and N21 transfer. We randomly sample 50 documents for human evaluation. Ratings are on a 0-6 scale; higher scores mean better quality.

| Models | ACC En-Fr | ACC En-De |
|---|---|---|
| SenNMT w/ SenInfer | 50.00 | 52.00 |
| SenNMT w/ DocInfer | **58.50**$^\star$ | 50.80 |
| DocNMT w/ SenInfer | 50.00 | 51.90 |
| DocNMT w/ DocInfer | **64.50**$^\dagger$ | **66.80**$^\dagger$ |

Table 5: Applying DocInfer and SenInfer to DocNMT and SenNMT for contrastive evaluation. Models are trained on Europarl-7. $^\star/^\dagger$: significant at $p < 0.05/0.01$.

its WMT20 test set is intentionally constructed for DocNMT evaluation. Table 4 lists the results.

We observe that zero-shot transfer matches and even surpasses SenNMT through N21 transfer, but fails with 12N transfer, although accuracy improvements on contrastive test sets show that both transfers are better than SenNMT. We conjecture that these contrastive test sets only target a limited number of document phenomena and thus can't fully reflect the overall translation quality and represent human preference. These numbers verify the feasibility of document-level zero-shot transfer through multilinguality. Besides, we find that genuine parallel documents benefit the transfer slightly more than BT-based pseudo ones, and that the supervised DocNMT reaches the best result under DocInfer.

We surprisingly find that DocNMT with SenInfer yields very competitive performance, although no contextual information is used for decoding. We also observe that such decoding tends to produce longer translations than SenNMT despite using the same decoding hyperparameters. This behaviour should be shaped by the fact that DocNMT is biased towards long concatenated target references. This partially agrees with the recent argument that context improves DocNMT with some sort of regularization rather than teaching the model to deal with context (Kim et al., 2019). On the other hand, this challenges how to properly evaluate DocNMT.

Another observation is that applying DocInfer to SenNMT delivers a significant accuracy improvement on En-Fr contrastive test set (+8.5%, Table 5), but slightly worse results on En-De. To accurately recognize the correct translation in these test sets, models need to leverage context. Such improvement might suggest that SenNMT has some limited capability of contextual modeling, but might just reflect the instability of small-scale test sets (only 200 cases in En-Fr test set, indicating a radius of around 7% for the 95% confidence interval). To some extent, this devalues the improvement achieved by 12N transfer as shown in Table 3, but strengthens the success of N21 transfer (often >9% gains).

## 7 Conclusion and Future Work

This paper studies the variables playing role in achieving zero-shot document-level translation capability for languages that only have sentence level data (students), through multilingual transfer from languages that have access to document level data (teachers). We make the first step in this direction by extensively exploring properties of transfer by investigating three different variables. Our experiments on Europarl-7 and IWSLT-10 confirm the feasibility, where we discover that increasing document-supervised teacher languages thereby increasing the document training data size, adequately balancing between document and sentence data at training, and leveraging monolingual documents via back-translation all benefit zero-shot transfer in varying degrees. The transferability of contextual modeling in DocNMT demonstrates the potential of delivering multilingual DocNMT with limited document resources.

Along with the success of document-level zero-shot transfer, problems with accurately estimating the document-level translation become challenging. BLEU often fails to capture document phenomena, while contrastive test sets only cover few document-level aspects. Neither perfectly correlates with human evaluation. Besides, whether the gains really come from contextual modeling is unclear. Our human evaluation shows some preference to DocNMT with SenInfer where context is not used for decoding at all. Designing better evaluation protocols (either automatic or human) is again confirmed to be critical. Besides, performing analysis beyond 12N and N21 transfer deserves more effort and it is an interesting and plausible future direction to analyze how language similarity affects the transfer.

## References


Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation.

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

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. 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, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stüker, K. Sudoh, K. Yoshino, and C. Federmann. 2017. The iwslt 2017 evaluation campaign. In *The International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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*, pages 604–616, Online. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context?

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat,

Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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 (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. 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, Hong Kong, China. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

James Kuczmarski and Melvin Johnson. 2018. Gender-aware natural language translation.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

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*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

António 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*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Zhiyi Ma, Sergey Edunov, and Michael Auli. 2021. A comparison of approaches to document-level machine translation.

Elman Mansimov, Gábor Melis, and Lei Yu. 2020. Capturing document context inside sentence-level neural machine translation models with self-training.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

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*, pages 2947–2954, Brussels, Belgium. 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*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

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

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8854–8861.

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 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *arXiv preprint arXiv:2010.08961*.

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

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

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*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization. Main track.

## A Data Statistics

Table 6 shows the statistics for Europarl-7 and IWSLT-10. Compared to IWSLT-10, Europarl-7 includes fewer languages, but with higher quantity and more uneven distribution.

## B Model Training and Decoding Settings

We pretrain multilingual SenNMT for 100K and 300K steps on IWSLT-10 and Europarl-7 respectively, and adopt extra 20K finetuning steps for multilingual DocNMT. We train all models (SenNMT & DocNMT) with a fixed batch size of 1280 samples, and schedule the training data distribution over language pairs according to the sentence-level statistics (without oversampling, and this also applies to DocNMT). All such measures aim to ensure a fair comparison between SenNMT and DocNMT. For training, we truncate sequences with length limit of 100 and 512 for SenNMT and DocNMT separately. We average last 5 checkpoints for evaluation. Beam search is used for decoding with a beam size of 4 and length penalty of 0.6. During decoding, we disable the generation of the end-of-sentence symbol for DocInfer until the model outputs the correct number of target translations.

## C Impact of Back-Translated Documents on Translation

The back-translated documents belong to extra training data. How to mix them with the genuine sentence pairs during training is questionable. Before further study, we first explore the impact of these documents on translation.

Specifically, we sample $p\%$ BT documents for **each language** during training with the rest $(1 - p\%)$ being the original sentence pairs to testify the sensitivity of translation performance to $p$. Note the proportion $p$ here differs from the one used in our main paper (where $p$ denotes the proportion of parallel documents in all teacher languages to parallel sentences in student languages).

Figure 6 shows that larger $p$ generally yields better performance over all settings, similar to the results on genuine parallel documents as in Figure 7. Therefore, we replace all sentence pairs in teacher languages with the corresponding BT documents in our analysis.

## D Transfer Results From Individual Languages to De/Fr

We mainly report average results over all transfer directions in the paper. Below we also show the transfer from individual languages to De and Fr on Europarl-7. Note the performance at language level is much noisy. We observe that different teacher languages yield slightly different transfer behavior and transferring to Fr looks more promising.

| Language (Pair) | Train-Para | | Train-Mono | | Dev | | Test | |
|---|---|---|---|---|---|---|---|---|
| | #Sent | #Doc | #Sent | #Doc | Source | #Doc | Source | #Doc |
| Europarl-7 | | | | | | | | |
| En-Cs | 192K | 901 | 658K | 9759 | WMT19 | 123 | WMT20 | 130 (102) |
| En-De | 1.81M | 3394 | 2.10M | 10155 | WMT19 | 123 (145) | WMT20 | 130 (118) |
| En-Fi | 1.82M | 3569 | 2.00M | 10161 | WMT18 | 132 | WMT19 | 123 (134) |
| En-Fr | 1.90M | 3676 | 2.07M | 10304 | WMT13 | 52 | WMT14 | 176 |
| En-Lt | 189K | 901 | 668K | 9740 | WMT19 | 130 | WMT19 | 62 (76) |
| En-Pl | 191K | 901 | 694K | 9775 | WMT20 | 128 | WMT20 | 63 (62) |
| En | - | - | 2.28M | 11109 | - | - | - | - |
| IWSLT-10 | | | | | | | | |
| En-Ar | 232K | 1907 | 107K | 1316 | IWSLT17 Dev10 | 19 | IWSLT17 Tst17 | 12 |
| En-De | 206K | 1705 | 41K | 466 | | 19 | | 10 |
| En-Fr | 233K | 1914 | 119K | 1300 | | 19 | | 12 |
| En-It | 249K | 1902 | 89K | 942 | | 19 | | 10 |
| En-Ja | 223K | 1863 | 26K | 1037 | | 19 | | 12 |
| En-Ko | 230K | 1920 | 132K | 1153 | | 19 | | 12 |
| En-Nl | 253K | 1805 | 52K | 501 | | 19 | | 10 |
| En-Ro | 237K | 1812 | 64K | 761 | | 19 | | 10 |
| En-Zh | 231K | 1906 | 108K | 1283 | | 19 | | 12 |
| En | - | - | 136K | 1445 | - | - | - | - |

Table 6: Statistics of train, dev and test data for Europarl-7 and IWSLT-10. Numbers in the bracket are for the reversed evaluation direction. "*Para*": parallel corpus; "*Mono*": monolingual corpus; "*#Sent/#Doc*": number of sentences/documents.

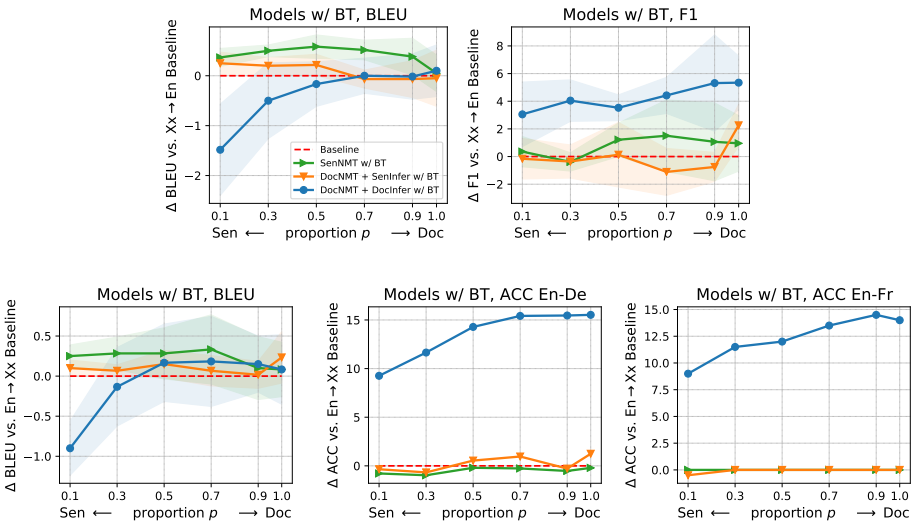


Figure 6: Translation performance as a function of proportion $p$ with **back-translated documents** for En→Xx (bottom) and Xx→En (top) translation on Europarl-7. This is fully supervised multilingual DocNMT, where pseudo documents are used for all languages. *Also, note p denotes the proportion of documents for each language, rather than teacher languages.*

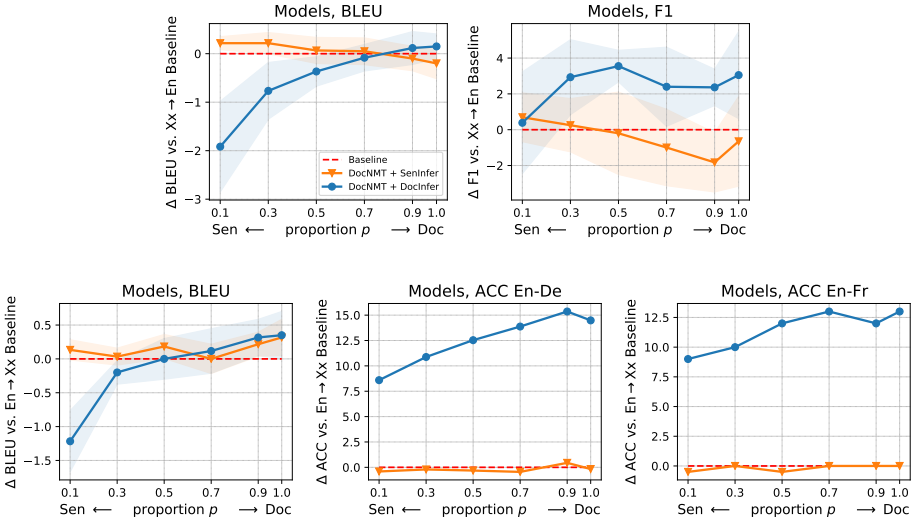


Figure 7: Translation performance as a function of proportion $p$ with **genuine parallel documents** for En→Xx (bottom) and Xx→En (top) translation on Europarl-7. Other settings follow Figure 6.

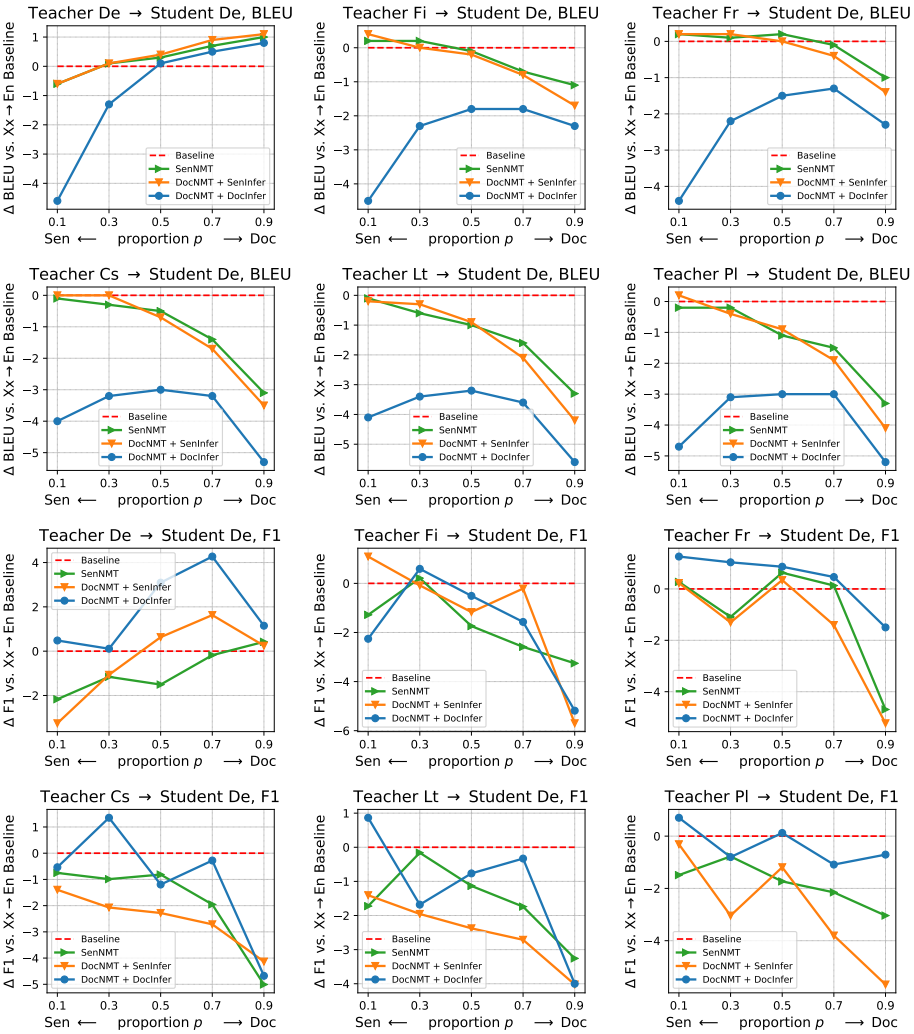


Figure 8: Transfer performance from individual languages to De as a function of proportion $p$ with genuine parallel documents for **Xx→En** translation on Europarl-7.

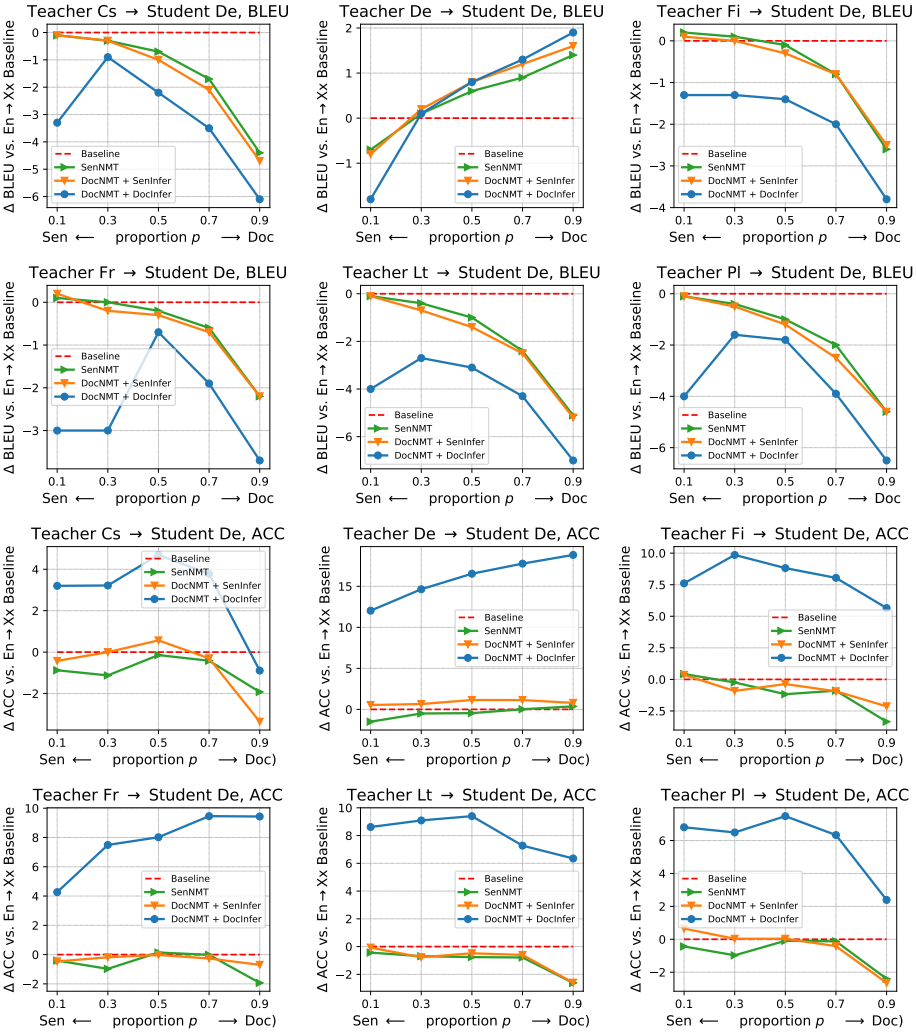


Figure 9: Transfer performance from individual languages to De as a function of proportion $p$ with genuine parallel documents for **En**→**Xx** translation on Europarl-7.

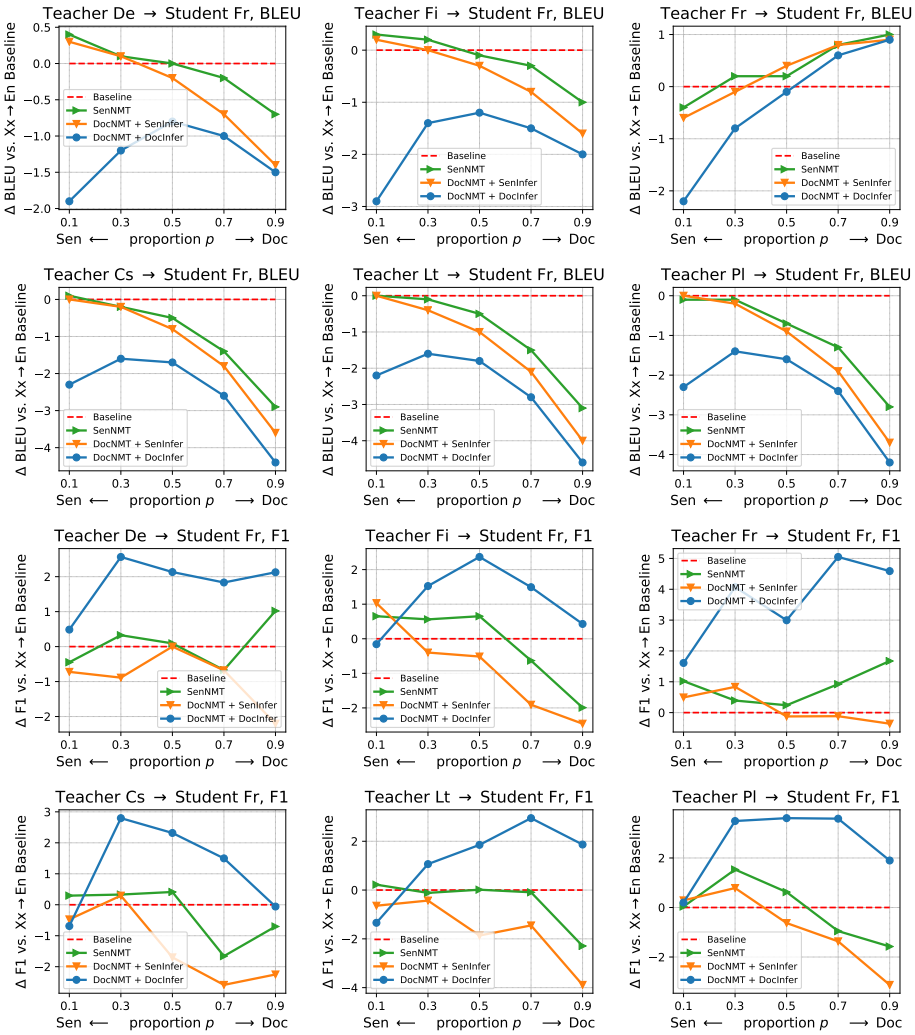


Figure 10: Transfer performance from individual languages to Fr as a function of proportion $p$ with genuine parallel documents for **Xx**→**En** translation on Europarl-7.

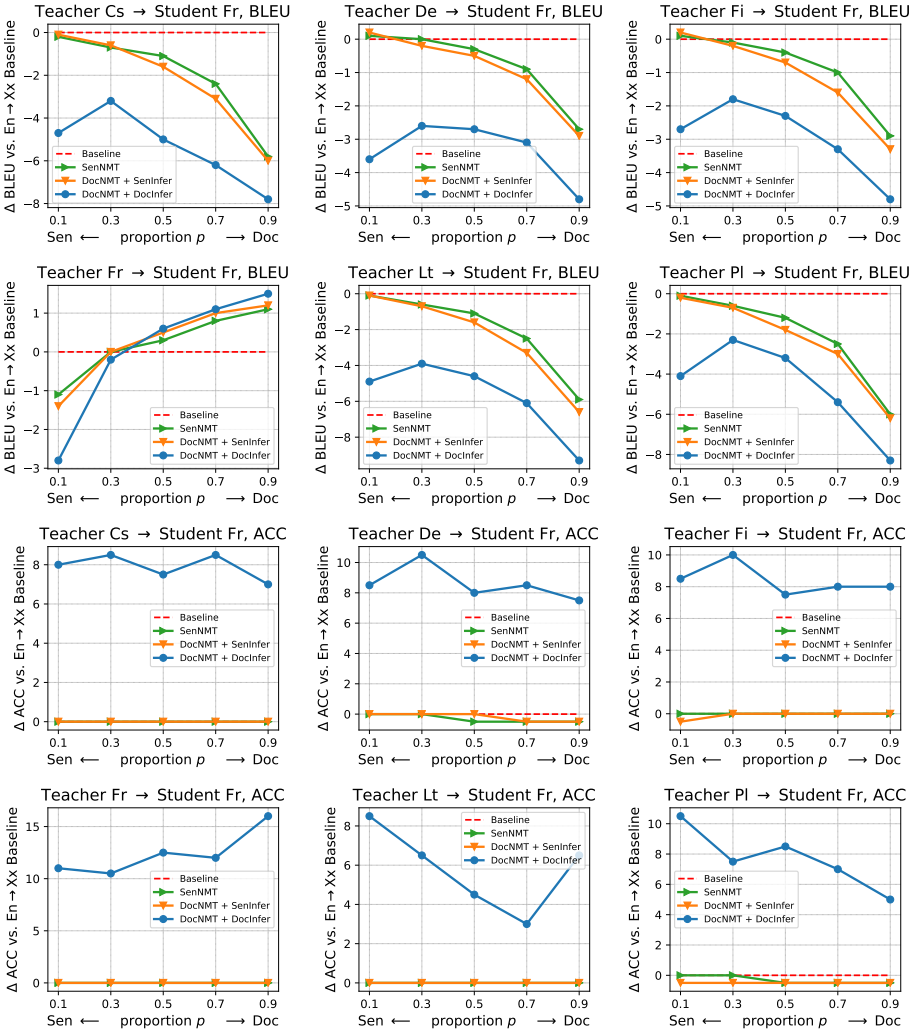


Figure 11: Transfer performance from individual languages to Fr as a function of proportion $p$ with genuine parallel documents for **En**→**Xx** translation on Europarl-7.