Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation

Mozhdeh Gheini, Xiang Ren, Jonathan May
Information Sciences Institute
University of Southern California
{gheini, xiangren, jonmay}@isi.edu

Abstract

We study the power of *cross-attention* in the Transformer architecture within the context of *transfer learning* for machine translation, and extend the findings of studies into cross-attention when *training from scratch*. We conduct a series of experiments through fine-tuning a translation model on data where either the source or target language has changed. These experiments reveal that fine-tuning only the cross-attention parameters is nearly as effective as fine-tuning all parameters (i.e., the entire translation model). We provide insights into why this is the case and observe that limiting fine-tuning in this manner yields cross-lingually aligned embeddings. The implications of this finding for researchers and practitioners include a mitigation of catastrophic forgetting, the potential for zero-shot translation, and the ability to extend machine translation models to several new language pairs with reduced parameter storage overhead.[1]

1 Introduction

The Transformer (Vaswani et al., 2017) has become the de facto architecture to use across tasks with sequential data. It has been dominantly used for natural language tasks, and has more recently also pushed the state-of-the-art on vision tasks (Dosovitskiy et al., 2021). In particular, transfer learning from large pretrained Transformer-based language models has been widely adopted to train new models: adapting models such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) for encoder-only tasks and models such as BART (Lewis et al., 2020) and mBART (Liu et al., 2020) for encoder-decoder tasks like machine translation (MT). This transfer learning is predominantly performed in the form of fine-tuning: using the values of several hundred million parameters from the pretrained model to initialize a model and start training from there.

Fine-tuning pretrained models often involves updating all parameters of the model without making a distinction between them based on their importance. However, copious recent studies have looked into the relative cruciality of multi-headed self- and cross- attention layers when training an MT model *from scratch* (Voita et al., 2019; Michel et al., 2019; You et al., 2020). Cross-attention (also known as encoder-decoder attention) layers are more *important* than self-attention layers in the sense that they result in more degradation in quality when pruned, and hence, are more sensitive to pruning (Voita et al., 2019; Michel et al., 2019). Also, cross-attention cannot be replaced with hard-coded counterparts (e.g., an input-independent Gaussian distribution) without significantly hurting the performance, while self-attention can (You et al., 2020). With the ubiquity of fine-tuning as a training tool, we find a similar investigation focused on transfer learning missing. In this work, we inspect cross-attention and its importance and capabilities through the lens of transfer learning for MT.

At a high level, we look at training a model for a new language pair by transferring from a pretrained MT model built on a different language pair. Given that, our study frames and addresses three questions: 1) How *powerful* is cross-attention alone in terms of adapting to the new language pair while other modules are frozen? 2) How *crucial* are the cross-attention layers pretrained values with regard to successful adaptation to the new task? and 3) Are there any *qualitative differences* in the learned representations when cross-attention is the only module that gets updated?

To answer these questions, we compare multiple strategies of fine-tuning towards a new language pair from a pretrained translation model that shares one language with the new pair. These are depicted in Figure 1: a) Ignoring the pretrained parameters and training entirely from randomly initialized parameters (i.e. 'from scratch') b) Fine–

[1] Our code is available at `https://github.com/MGheini/xattn-transfer-for-mt`.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1754–1765
November 7–11, 2021. ©2021 Association for Computational Linguistics

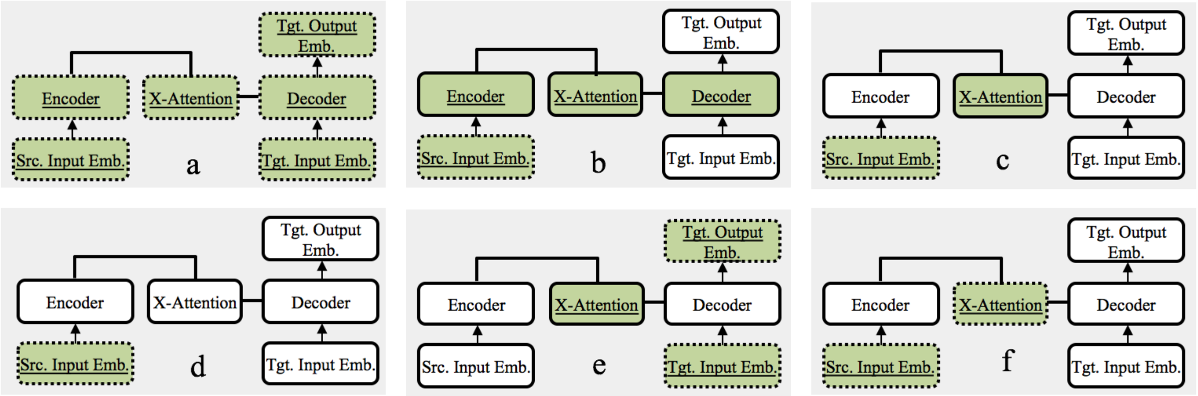

Figure 1: Overview of our transfer learning experiments, depicting (a) training from `scratch`, (b) conventional fine-tuning (`src+body`), (c) fine-tuning cross-attention (`src+xattn`), (d) fine-tuning new vocabulary (`src`), (e) fine-tuning cross-attention when transferring target language (`tgt+xattn`), (f) transfer learning with updating cross-attention from scratch (`src+randxattn`). Dotted components are initialized randomly, while solid lines are initialized with parameters from a pretrained model. Shaded, underlined components are fine-tuned, while other components are frozen.

tuning all parameters except the embeddings for the language in common,[2] (i.e. 'regular' fine-tuning, our upper bound), c) Fine-tuning solely the cross-attention layers and new embeddings, and d) Fine-tuning only the new embeddings. Here, new embeddings refer to randomly initialized embeddings corresponding to the vocabulary of the new language. In Figures 1a–1d, we assume the new language pair has a new source language and not a new target language; Figure 1e shows an example of target-side transfer. In the experiments that follow we will always train new, randomly initialized embeddings for the vocabulary of the newly introduced language. Generally, all other parameters are imported from a previously built translation model and, depending on the experiment, some will remain unchanged and others will be adjusted during training.

Our experiments and analyses show that fine-tuning the cross-attention layers while keeping the encoder and decoder fixed results in MT quality that is close to what can be obtained when fine-tuning all parameters (§4). Evidence also suggests that fine-tuning the *previously trained* cross-attention values is in fact important—if we start with randomly initialized cross-attention parameter values instead of the pretrained ones, we see a quality drop.

Furthermore, intrinsic analysis of the embeddings learned under the two scenarios reveals that full fine-tuning exhibits different behavior from cross-attention-only fine-tuning. When the encoder and decoder bodies are not fine-tuned, we show that the new language's newly-learned embeddings *align* with the corresponding embeddings in the pretrained model. That is, when we transfer from Fr–En to Ro–En for instance, the resulting Romanian embeddings are aligned with the French embeddings. However, we do not observe the same effect when fine-tuning the entire body. In §5 we see how such aligned embeddings can be useful. We specifically show they can be used to alleviate forgetting and perform zero-shot translation.

Finally, from a practical standpoint, our strategy of fine-tuning only cross-attention is also a more *lightweight* fine-tuning approach (Houlsby et al., 2019) that reduces the storage overhead for extending models to new language pairs: by fine-tuning a subset of parameters, we only need to keep a copy of those instead of a whole-model's worth of values for the new pair. We quantify this by reporting the fraction of parameters that is needed in our case relative to having to store a full new model for each adapted task.

Our **contributions** are: 1) We empirically show the competitive performance of exclusively fine-tuning the cross-attention layers when contrasted with fine-tuning the entire Transformer body; 2) We show that when fine-tuning only the cross-attention layers, the new embeddings get aligned with the respective embeddings in the pretrained model. The same effect does not hold when fine-tuning the entire Transformer body; 3) we demonstrate effective application of this aligning artifact in mitigating

[2] Freezing shared language embeddings is common practice (Zoph et al., 2016).

catastrophic forgetting (Goodfellow et al., 2014) and zero-shot translation.

2 Cross-Attention Fine-Tuning for MT

Fine-tuning pretrained Transformer models towards downstream tasks has pushed the limits of NLP, and MT has been no exception (Liu et al., 2020). Despite the prevalence of using pretrained Transformers, recent studies focus on investigating the importance of self- and cross- attention heads while training models from scratch (Voita et al., 2019; Michel et al., 2019; You et al., 2020). These studies verify the relative importance of cross-attention over self-attention heads by exploring either pruning (Voita et al., 2019; Michel et al., 2019) or hard-coding methods (You et al., 2020). Considering these results and the popularity of *pretrained* Transformers, our goal in this work is to study the significance of cross-attention while focusing on transfer learning for MT. This section formalizes our problem statement, introduces the notations we will use, and describes our setup to address the questions we raise.

2.1 Problem Formulation

In this work, we focus on investigating the effects of the cross-attention layers when fine-tuning pretrained models towards new MT tasks. Fine-tuning for MT is a transfer learning method that, in its simplest form (Zoph et al., 2016), involves training a model called the 'parent' model on a relatively high-resource language pair, and then using the obtained parameters to initialize a 'child model' when further training towards a new, potentially low-resource, language pair. Here, high-resource and low-resource refer to the amount of *parallel* data that is available for the languages. Henceforth, we use 'parent' and 'child' when referring to training components (e.g., model, data, etc.) in the pretraining and fine-tuning stages, respectively.

Formal Definition. Consider a model f_θ trained on the parent dataset, where each training instance (x_{s_p}, y_{t_p}) is a pair of source and target sentences in the parent language pair s_p–t_p. Then fine-tuning is the practice of taking the model's parameters θ from the model f_θ to initialize another model g_θ. g_θ is then further optimized on a dataset of (x_{s_c}, y_{t_c}) instances in the child language pair s_c–t_c until it converges to g_ϕ. We assume either $s_c = s_p$ or $t_c = t_p$ (i.e., child and parent language pairs share one of the source or target sides).

Granular Notations. It is common practice for fine-tuning to further update all parent parameters θ on the child data without making any distinction between them. We instead consider θ at a more granular level, namely as:

$$\theta = \bigcup\{\theta_{\text{src}}, \theta_{\text{tgt}}, \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_{\text{xattn}}\}$$

where θ_{src} includes source-language token embeddings, source positional embeddings, and source embeddings layer norm parameters; θ_{tgt} similarly includes target-language (tied) input and output token embeddings, target positional embeddings, and target embeddings layer norm parameters; θ_{enc} includes self-attention, layer norm, and feed-forward parameters in the encoder stack; θ_{dec} includes self-attention, layer norm, and feed-forward parameters in the decoder stack; and θ_{xattn} includes cross-attention and corresponding layer norm parameters.

2.2 Analysis Setup

Inspections like ours into individual modules of Transformer often rely on introducing some constraints in order to understand the module better. These constraints come in the form of full removal or pruning (Tang et al., 2019; Voita et al., 2019), hard-coding (You et al., 2020), and freezing (Bogoychev, 2020). We rely on freezing. We proceed by taking pretrained models, freezing certain parts, and recording the effect on performance, measured by BLEU.

Within the framework of our problem, to address the questions raised in §1, our analysis compares full and partially-frozen fine-tuning for MT under several settings, which we summarize here:

Cross-attention fine-tuning & embedding fine-tuning comparative performance. This is to realize how much fine-tuning the cross-attention layers helps in addition to fine-tuning respective embeddings alone.

Cross-attention fine-tuning & full fine-tuning comparative performance. We wish to find out where fine-tuning cross-attention stands relative to fine-tuning the entire body. This is to confirm whether or not cross-attention alone can adapt to the child language pair while the encoder and decoder layers are frozen.

Pretrained cross-attention layers & random cross-attention layers. We wish to understand how important a role cross-attention's pretrained values play when single-handedly adapting to a new language pair. This determines if the knowledge encoded in cross-attention itself has a part in its power.

Translation cross-attention & language modelling cross-attention. Finally, we contrast the knowledge encoded in cross-attention learned by different pretraining objectives. This is to evaluate if the knowledge brought about by a different pretraining objective affects the patterns observed from a cross-attention pretrained on MT while fine-tuning for MT.

3 Experimental Setup

In this section, we describe our experiments and the data and model that we use to materialize the analysis outlined in §2.2.

3.1 Methods

We first provide the details of our transfer setup, and then describe the specific fine-tuning baselines and variants used in our experiments.

General Setup. An important concern when transferring is initializing the embeddings of the new language. When initializing parameters in the child model, there are several ways to address the vocabulary mismatch between the parent and the child model: frequency-based assignment, random assignment (Zoph et al., 2016), joint (shared) vocabularies (Nguyen and Chiang, 2017; Kocmi and Bojar, 2018; Neubig and Hu, 2018; Gheini and May, 2019; Liu et al., 2020), and no assignment at all, which results in training randomly initialized embeddings (Aji et al., 2020). In our experiments, we choose to always use new random initialization for the new embeddings (including token embeddings, positional embeddings, and corresponding layer norm parameters). This decision is made to later let us study what happens to embeddings under each of the settings, independent of any pretraining artifacts that exist in them. For instance, when transferring from Fr–En to {Ro–En, Fr–Es}, respectively, all parameters are reused except for $\{\theta_{\text{src}}, \theta_{\text{tgt}}\}$,[3] which get re-initialized given the new {source, target} language. The side that remains the same (e.g., En when going from Fr–En to Ro–En), uses the parent vocabulary and keeps the corresponding embeddings frozen during fine-tuning.[4]

	Train Corpus (Sent. Count)	Test Corpus	Vocab. Size
Ro–En	WMT16 (612.4 K)	newstest2016	16 K / reuse tgt
Ja–En	IWSLT17 (223.1 K)	IWSLT17	8 K / reuse tgt
De–En	IWSLT16 (196.9 K)	IWSLT16	8 K / reuse tgt
Ha–En	ParaCrawl v8 (159.0 K)	newsdev2021	8 K / reuse tgt
Fr–Es	News Comm. v15 (283.5 K)	newstest2013	reuse src / 8 K
Fr–De	News Comm. v15 (284.1 K)	newstest2020	reuse src / 8 K

Table 1: Data sources and statistics for each of the child language pairs.

Fine-tuning Settings. With the general transfer setup, we employ different settings in our experiments to address the points in §2.2. Each fine-tuning method is clarified based on our notations in §2.1 : **1)** `{src,tgt}` only updates the embeddings $\{\theta_{\text{src}}, \theta_{\text{tgt}}\}$ (Figure 1d). **2)** `{src,tgt}+body` additionally updates the entire Transformer body ($\{\theta_{\text{src}}, \theta_{\text{tgt}}\} + \theta_{\text{enc}} + \theta_{\text{dec}} + \theta_{\text{xattn}}$) (Figure 1b). **3)** `{src,tgt}+xattn` only updates the cross-attention layers in addition to the first baseline ($\{\theta_{\text{src}}, \theta_{\text{tgt}}\} + \theta_{\text{xattn}}$), and keeps the encoder and decoder stacks frozen (Figure 1c, 1e). These collectively address the first and second settings in §2.2. **4)** `{src,tgt}+randxattn` similarly only updates the cross-attention layers in addition to embeddings, but uses randomly initialized values instead of pretrained values (Figure 1f). This addresses the third setting in §2.2.

For all transfer experiments, we also conduct the `scratch` variant (Figure 1 a), where we train a model from scratch on the child dataset. This is to confirm the effectiveness of transfer under each setting. We conduct all the above experiments using a French–English translation model as parent and transferring to six different child language pairs. In §4.1 we conduct an ablation that substitutes mBART (Liu et al., 2020) as a parent. mBART is trained with denoising objective in a self-supervised manner. In contrast to a translation model, the cross-attention layers in mBART have thus not been learned using any parallel data. This enables us to distinguish between different pretraining objectives, addressing the fourth setting in §2.2.

[3] We drop the "respectively" henceforth and use {...} throughout to indicate alternation.

[4] Preliminary ablations fine-tuning all embeddings did not change the outcome or conclusions of our experiments.

3.2 Data and Model Details

Dataset. For the choice of language pairs and datasets, we mostly follow You et al. (2020) (Fr–En, Ro–En, Ja–En, De–En) and additionally include Ha–En, Fr–Es, and Fr–De. We designate Fr–En as the parent language pair and Ro–En, Ja–En, De–En, Ha–En (new source), Fr–Es, Fr–De (new target) as child language pairs. Our Fr–En parent model is trained on the Europarl + Common Crawl subset of WMT14 Fr–En,[5] which comprises 5,251,875 sentences. Details and statistics of the data for the child language pairs are provided in Table 1.

Model Details. We use the Transformer base architecture (6 layers of encoder and decoder with model dimension of 512 and 8 attention heads) for all models, (Vaswani et al., 2017) and the Fairseq (Ott et al., 2019) toolkit for all our experiments.

All models rely on BPE subword vocabularies (Sennrich et al., 2016) processed through the SentencePiece (Kudo and Richardson, 2018) BPE implementation. The vocabulary for the parent model consists of 32K French subwords on the source side, and 32K English subwords on the target side. The sizes of the vocabularies for child models are also reported in Table 1. We follow the advice from Gowda and May (2020) when deciding what vocabulary size to choose, i.e., we choose the maximum number of operations to ensure a minimum of 100 tokens per type.

4 Results and Analysis

Our preliminary empirical results consist of five experiments for each of the child language pairs based on methods described in §3.1: `scratch`, `{src,tgt}`, `{src,tgt}+body`, `{src,tgt}+xattn`, and `{src,tgt}+randxattn`. Our core results, which rely on transferring from the Fr–En parent under each setting, are reported in Table 2. All scores are detokenized cased BLEU computed using SacreBLEU (Post, 2018).[6]

4.1 Cross-attention's *Power* and *Importance*

Translation Quality. Table 2 shows that `{src,tgt}+xattn` substantially improves upon `{src,tgt}` in all but one case (Ha–En), especially when transferring to a pair with a new target language, and is competitive with `{src,tgt}+body` across all six language pairs, suggesting that cross-attention is capable of taking advantage of encoded generic translation knowledge in the Transformer body to adapt to each child task. Performance gain from `{src,tgt}` and drop from `{src,tgt}+body` when changing the target language (i.e., Fr–Es and Fr–De) are more pronounced than when transferring the source. This is expected—when changing the target, two out of three cross-attention matrices (key and value matrices) are now exposed to a new language. When transferring source, only the query matrix is exposed to the new language.

Storage. We also report the fraction of the parameters that need to be updated in each case. This is equivalent to the storage overhead that the training process incurs, as the updated parameters need to be stored to be used later. However, the parameters that are reused are only stored once. The number of parameters updated is dependent on the size of the vocabulary in each experiment, since embeddings for a new vocabulary are included. Hence, the single number reported for each fine-tuning strategy is the average across the six language pairs. *Extending* to new language pairs following `{src,tgt}+xattn` is much more efficient in this regard, as expected. We concretely calculate the number of parameters that need to be stored combined for the six new language pairs: `{src,tgt}+xattn` stores only 124,430,336 parameters compared to `{src,tgt}+body`'s 313,583,616.

Pretrained and Random Values. Finally, `{src,tgt}+randxattn` experiments also offer perspective on the importance of translation knowledge encoded in cross-attention itself. Not only does randomly initialized cross-attention fail to perform as well as pretrained cross-attention when being transferred, but in two cases, it even falls behind training from scratch.

Our results from transferring mBART (Liu et al., 2020) to the child language pairs also emphatically illustrate the importance of the type of knowledge encoded in cross-attention. mBART is a 12-layer Transformer pretrained with a denoising objective in a self-supervised manner using span masking and sentence permutation noising functions. Hence, its cross-attention does not have any *translation* knowledge *a priori*, in contrast with the French–English MT parent model. We transfer mBART to the same language pairs as in Table 2 and pro-

[5] `http://statmt.org/wmt14/translation-task.html`

[6] Signature: BLEU+case.mixed+numrefs.1+smooth.exp +tok.13a+version.1.4.8.

	Ro–En	Ja–En	De–En	Ha–En	Fr–Es	Fr–De
`scratch` (100%)	29.0	9.2	30.8	5.4	24.4	18.5
`{src,tgt}` (8%)	29.8	8.7	32.4	**8.6**	21.6	11.6
`{src,tgt}+body` (75%)	**31.0**	**11.8**	**36.2**	**8.8**	**27.3**	**21.4**
`{src,tgt}+xattn` (17%)	**(-0.1) 30.9**	**(-2.0) 9.8**	**(-1.2) 35.0**	(-0.4) 8.4	**(-0.8) 26.5**	**(-1.8) 19.6**
`{src,tgt}+randxattn` (17%)	27.9	8.4	33.3	7.0	26.0	18.8

Table 2: BLEU scores for each of the five experiments across six language pairs. Bold numbers indicate the top two scoring approaches. Percentages in parentheses next to fine-tuning strategy is the fraction of parameters that had to be updated and hence stored as new values for future use. Numbers in parentheses next to `{src,tgt}+xattn` scores show the difference from `{src,tgt}+body`.

vide the results in Figure 2. Since mBART uses a shared vocabulary and tied embeddings between the encoder and decoder, in Figure 2 we use `embed` in experiments' names to signify all embeddings get updated in the case of mBART ($\theta_{\text{src}} + \theta_{\text{tgt}}$).

mBART is a larger model than our Fr–En parent, both in terms of architecture and training data. So a higher range of scores is expected. While the same patterns hold across `embed+{body,xattn,randxatnn}` fine-tuning, the crux of the matter is that `embed` fine-tuning fails in contrast to the comparable `{src, tgt}` fine-tuning setting of the translation parent. `src` fine-tuning has higher BLEU than `scratch` in three cases (Ro–En, De–En, Ha–En). However, `embed` fine-tuning has higher BLEU than the `scratch` baseline only in the Ja–En case, and even then, very slightly so (only by 0.1 BLEU). This shows that absence of translation knowledge in mBART's pretrained cross-attention leads to its fine-tuning being more crucial in mBART's functionality for translation adaptation: exclusively fine-tuning embeddings in mBART simply fails, while doing the same with a translation parent model is more successful.

4.2 Learned Representations Properties

Given that besides cross-attention, embeddings are the only parameters that get updated in both `{src,tgt}+body` and `{src,tgt}+xattn` settings, we take a closer look at them. We want to know how embeddings change under each setting.

To probe the relationship between embeddings learned as a result of different kinds of fine-tuning, we examine the quality of induced[7] bilingual lexicons, a common practice in cross-lingual embeddings literature (Artetxe et al., 2017) but incidentally learned in this case.

We use the bilingual dictionaries released as a resource in the MUSE (Lample et al., 2018) repository.[8] For instance, to compare the German embeddings from each of the `src+body` and `src+xattn` De–En models to the French embeddings learned in the parent model, we use the De–Fr dictionary. We filter our learned embeddings (which are, in general, of subwords) to be compatible with the MUSE vocabulary. Of the 8,000 German subwords in the vocabulary, 2,025 are found in MUSE. For each of these, we find the closest French embedding by cosine similarity; if the resulting (German, French) pair is in MUSE, we consider this a match. Via this method, we find the accuracy of the bilingual lexicon induction through the embeddings of `src+xattn` model is 55%. However, the accuracy through the embeddings of `src+body` is much lower at 19.7%. Due to only considering the exact matches against the gold dictionary, this is a very strict evaluation. We also manually look at a sample of 40 words from the German set and check for the correctness of retrieved pairs for those using an automatic translator: while `src+xattn` scores in the range of 80%, `src+body` scores in the range of 30%. Details of this manual inspection are provided in Table 4 of the appendix. We further report the accuracy of the bilingual dictionaries of three other pairs learned under the two fine-tuning settings for which gold dictionaries are available in Figure 3. We don't limit ourselves to child-parent dictionary induction; we also consider child-child dictionary induction (e.g., De–Es) which essentially relies on both languages being aligned with the parent (i.e., En).

Overall, these results confirm that embeddings learned under `{src,tgt}+xattn` effectively get aligned with corresponding parent embeddings. However, this is not the case with embeddings

[7] via nearest neighbor retrieval

[8] https://github.com/facebookresearch/MUSE

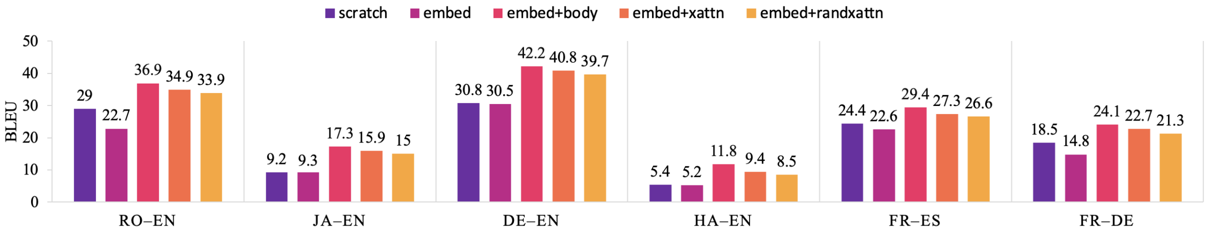

Figure 2: BLEU scores across different transfer settings using mBART as parent. Exclusive fine-tuning of embeddings (embed) is not effective at all due to lack of translation knowledge in the cross-attention layers.

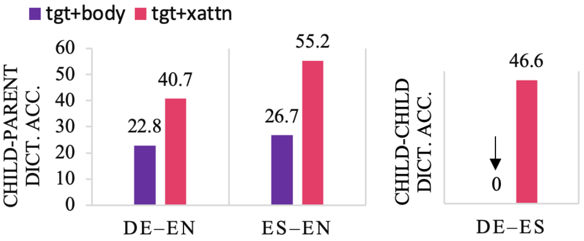

Figure 3: Accuracy of bilingual dictionaries induced through embeddings learned under tgt+body and tgt+xattn settings. De and Es effectively get aligned with En under tgt+xattn (left). As they are both aligned to En, we can also indirectly obtain a De–Es dictionary (right). Similar practice completely fails under tgt+body.

learned under {src,tgt}+body. This suggests such effect is not the default pattern in translation models, but rather an artifact of the freezing choices made in {src,tgt}+xattn.

5 Utilities of Aligned Embeddings

We saw how fine-tuning only cross-attention results in cross-lingual embeddings with respect to parent embeddings. That is how cross-attention is able to use the baked-in knowledge in the encoder and decoder without any further updates to them. In this section, we discuss two areas where this can be turned to our advantage: mitigating forgetting and performing zero-shot translation.

5.1 Mitigating Forgetting

One area where the discovery of §4.2 can be taken advantage of is mitigating catastrophic forgetting. Catastrophic forgetting refers to the loss of previously acquired knowledge in the model during transfer to a new task. To the best of our knowledge, catastrophic forgetting in MT models has only been studied within the context of inter-domain adaptation (Thompson et al., 2019; Gu and Feng, 2020), and not inter-lingual adaptation.

The effectiveness of the cross-lingual embeddings learned under the {src,tgt}+xattn setting at mitigating forgetting is evident from the results provided in Figure 4. Here we take three of the transferred models, plug back in the appropriate embeddings in them, and compare their performance **on the original language pair** against the parent model. Specifically, we take the De–En, Ro–En, and Fr–Es models transferred from Fr–En under each of the two {src,tgt}+xattn and {src,tgt}+body settings, plug in back the original {Fr, En} embeddings, and evaluate performance on the Fr–En test set. This score is then compared against the Fr–En parent model performance on Fr–En test set, which scores 35.0 BLEU. While being comparable in terms of performance on the child task as reported in Table 2, {src,tgt}+xattn constantly outperforms {src,tgt}+body on Fr–En. Compared to the original Fr–En model, the source-transferred models (De–En, Ro–En) outperform the target-transferred model (Fr–Es). However, tgt+xattn is much more robust against forgetting compared to tgt+body, which remembers close to nothing (0.2 BLEU).

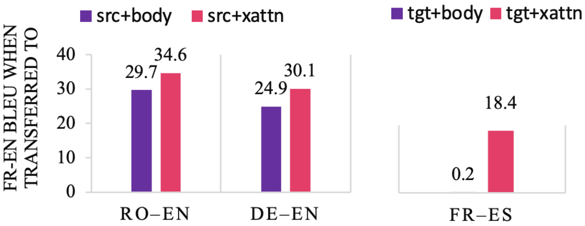

Figure 4: Performance on the original language pair after transfer. The original Fr–En parent model scores 35.0 BLEU on the Fr–En test set. {src,tgt}+xattn outperforms {src,tgt}+body on the parent task.

5.2 Zero-Shot Translation

Another area where well-aligned embeddings from the {src,tgt}+xattn setting can come in handy is zero-shot translation. Since the source embeddings are aligned, we, for instance, can replace the French embeddings in the Fr–Es model learned via tgt+xattn with German embeddings from the De–En model learned via src+xattn and form a De–Es translation model with no De–Es training or direct De–Fr alignment. We additionally build two more zero-shot systems in the same manner: Ro–Es (using transferred Ro–En and Fr–Es models) and Ro–De (using transferred Ro–En and Fr–De models). To put zero-shot scores in context, for each pair we also train a model from scratch: for De–Es using 294,216-sentence News Commentary v14 corpus, and for Ro–Es and Ro–De using 387,653-sentence and 385,663-sentence Europarl corpora respectively. All scores are provided in Table 3.

	De–Es	**Ro–Es**	**Ro–De**
Zero-shot BLEU	9.2	14.7	9.8
Supervised BLEU	18.3	18.6	13.4

Table 3: Performance of zero-shot systems for three language pairs. De–Es is evaluated on newstest2013 test set. Ro–Es and Ro–De are evaluated on respective TED talks corpus test sets (Qi et al., 2018).

In the case of De–Es, we train two additional models from scratch on 50,000- and 100,000- sentence subsets of the training corpus. These respectively score 7.2 and 12.0 BLEU on the newstest2013 De–Es test set (v.s. zero-shot performance of 9.2). Taken together, these results show that the zero-shot methods we obtain from cross-attention-based transfer can yield reasonable translation models in the absence of parallel data.

6 Related Work

Studying Cross-attention. Several recent works consider the importance of self- and cross-attention heads in the Transformer architecture (Voita et al., 2019; Michel et al., 2019; You et al., 2020). The consensus among these works is that cross-attention heads are relatively more important than self-attention heads when it comes to introducing restrictions in terms of pruning and hard-coding.

Module Freezing. In terms of restrictions introduced, our work is related to a group of recent works that freeze certain modules while fine-tuning (Zoph et al., 2016; Artetxe et al., 2020; Lu et al., 2021). Artetxe et al. (2020) conduct their study on an encoder-only architecture. They show that by freezing a pretrained English Transformer language model body and only *lexically* (embedding layers) transferring it to another language, they can later plug in those embeddings into a fine-tuned downstream English model, achieving zero-shot transfer on the downstream task in the other language. Lu et al. (2021) also work with a decoder-only architecture. They show that by only fine-tuning the input layer, output layer, positional embeddings, and layer norm parameters of an otherwise frozen Transformer language model, they can match the performance of a model fully trained on the downstream task in several modalities.

Lightweight Fine-tuning. Houlsby et al. (2019) reduce the number of parameters to be updated by inserting adapter modules in every layer of the Transformer model. Then during fine-tuning, they update the adapter parameters from scratch and fine-tune layer norm parameters while keeping the rest of the parameters frozen. Since adapters are only inserted and initialized at the time of fine-tuning, they are not able to reveal anything about the importance of pretrained modules. Our approach, however, enables highlighting the crucial role of the encoded translation knowledge by contrasting {src,tgt}+xattn and {src,tgt}+randxattn. Bapna and Firat (2019) devise adapters for MT by inserting language pair-specific adapter parameters in the Transformer architecture. In the multilingual setting, they show that by fine-tuning adapters in a shared pretrained multilingual model, they can compensate for the performance drop of high-resource languages incurred by shared training. Philip et al. (2020) replace language pair-specific adapters with monolingual adapters, which enables adapting under the zero-shot setting.

Another family of lightweight fine-tuning approaches (Li and Liang, 2021; Hambardzumyan et al., 2021; Lester et al., 2021), inspired by prompt tuning (Brown et al., 2020), also relies on updating a set of additional new parameters from scratch towards each downstream task. Such sets of parameters equal a very small fraction of the total parameters in the pretrained model. By contrast, our approach updates a subset of the model's own parameters instead of adding new ones. We leave a comparison of the relative advantages and disadvantages of these approaches to future work.

Cross-lingual Embeddings. Finally, while we were able to obtain cross-lingual embeddings through our transfer learning approach without using any dictionaries or direct parallel corpora, Wada et al. (2020) use a direct parallel corpus and a shared LSTM model that does translation and reconstruction at the same time to obtain aligned embeddings. Given tremendously large monolingual corpora for embedding construction, cross-lingual embeddings can also be obtained by applying a linear transformation on one language's embedding space to map it to the second one in a way that minimizes the distance between equivalents in the shared space according to a dictionary (Mikolov et al., 2013; Xing et al., 2015; Artetxe et al., 2016). These works specifically targeted the parallel dictionary reconstruction task, while we used the task incidentally, to intrinsically evaluate the parameters learned by our methods.

7 Conclusion

We look at how powerful cross-attention can be under constrained transfer learning setups. We empirically show that cross-attention can single-handedly result in comparable performance with fine-tuning the entire Transformer body, and it is through no magic: it relies on translation knowledge in the pretrained values to do so and has new embeddings align with corresponding parent language embeddings. We furthermore show that such aligned embeddings can be used towards catastrophic forgetting mitigation and zero-shot transfer. We hope this investigative study encourages more analyses in the same spirit towards more insights into the inner workings of different modules and how they can be put to good use.

Acknowledgements

The authors would like to thank Kushal Chawla, Muhao Chen, Katy Felkner, Thamme Gowda, Xuezhe Ma, Meryem M'hamdi, and Xusen Yin for their helpful feedback on the pre-submission draft of this work. This research is based in part on research sponsored by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract FA8650-17-C-9116 and in part on research sponsored by Air Force Research Laboratory (AFRL) under agreement number FA8750-19-1-1000. Ren's research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007, the Defense Advanced Research Projects Agency with award W911NF-19-20271, NSF IIS 2048211, and NSF SMA 182926. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, Air Force Laboratory, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation 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)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Nikolay Bogoychev. 2020. Not all parameters are born equal: Attention is mostly what you need. *CoRR*, abs/2010.11859.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Mozhdeh Gheini and Jonathan May. 2019. A universal parent model for low-resource neural machine translation transfer. *CoRR*, abs/1909.06516.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2014. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *In Proceedings of International Conference on Learning Representations (ICLR)*.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2020. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4921–4933, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

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.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*. OpenReview.net.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *CoRR*, abs/2103.05247.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. 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, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. 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.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Understanding neural machine translation by simplification: The case of encoder-free models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1186–1193, Varna, Bulgaria. INCOMA Ltd.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Takashi Wada, Tomoharu Iwata, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2020. Learning contextualised cross-lingual word embeddings for extremely low-resource languages using parallel corpora. *CoRR*, abs/2010.14649.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Weiqiu You, Simeng Sun, and Mohit Iyyer. 2020. Hard-coded Gaussian attention for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7689–7700, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Manual Bilingual Dictionary Evaluation

German Word	src+xattn French Equivalent	src+body French Equivalent
Entdeckung	découverte	amende
Feind	ennemi	ennemi
Architekten	architectes	architecture
gibt	existe	jette
erforschen	explorer	sond
Philosoph	philosophie	philosophie
Cent	centi	centaines
formen	forme	forme
lassen	laissez	PCP
Nummer	numéro	Key
können	puissent	puisse
dasselbe	mêmes	lourds
gelöst	résoud	résoud
wenig	peu	peu
zerstört	détruit	dévas
Bericht	reportage	témoin
Mark	Mark	trailer
Brief	lettre	lettres
Linien	lignes	lignes
entworfen	conçus	monté
Dunkelheit	ténèbres	obscur
Kreis	cercle	rond
Haie	requins	Hun
spielt	joue	tragédie
Elektrizität	électricité	électriques
Solar	solaire	Arabes
Flügel	ailes	avion
Konzept	concept	alliance
Strukturen	structures	définit
will	veut	voulons
Hier	Ici	Vous
verlieren	perdent	perdent
unterstützen	soutien	appui
Planet	planète	planète
buchstäblich	littéralement	multimédia
Schuld	blâ	génére
dass	que	toi
plötzlich	soudainement	risques
Kann	Pouvez	ciel
Ball	ballon	ballon

Table 4: Sampled German words and their equivalents based on the embeddings learned by each of the models. The correct translations are highlighted. Each pair was manually checked for correctness using an automatic translator.