Exploring Inline Lexicon Injection for Cross-Domain Transfer in Neural Machine Translation

Jesujoba O Alabi, Rachel Bawden

Exploring Inline Lexicon Injection for Cross-Domain Transfer in Neural Machine Translation

Jesujoba O. Alabi*
Saarland University, Germany
jalabi@lsv.uni-saarland.de

Rachel Bawden
Inria, Paris, France
rachel.bawden@inria.fr

Abstract

Domain transfer remains a challenge in machine translation (MT), particularly concerning rare or unseen words. Amongst the strategies proposed to address the issue, one of the simplest and most promising in terms of generalisation capacity is coupling the MT system with external resources such as bilingual lexicons and appending inline annotations within source sentences. This method has been shown to work well for controlled language settings, but its usability for general language (and ambiguous) MT is less certain. In this article we explore this question further, testing the strategy in a multi-domain transfer setting for German-to-English MT, using the mT5 language model fine-tuned on parallel data. We analyse the MT outputs and design evaluation strategies to understand the behaviour of such models. Our analysis using distractor annotations suggests that although improvements are not systematic according to automatic metrics, the model does learn to select appropriate translation candidates and ignore irrelevant ones, thereby exhibiting more than a systematic copying behaviour. However, we also find that the method is less successful in a higher-resource setting with a larger lexicon, suggesting that it is not a magic solution, especially when the baseline model is already exposed to a wide range of vocabulary.

1 Introduction

Data-driven machine translation (MT) models, and in particular neural MT models, have led to significant progress in the quality of automatic translation, particularly in settings where large amounts of data are available (Barrault et al., 2020; Akhbardeh et al., 2021; Saunders, 2021). However, a scenario in which MT typically struggles to perform as well is cross-domain transfer (Koehn and Knowles, 2017; Vu et al., 2021; Pham et al., 2021; Hasler et al., 2021; Bogoychev and Chen, 2021), where a model trained on one domain is adapted to a second domain, for which there typically exists less data. A major challenge is ensuring that the model is capable of handling the domain-specific vocabulary of the new domain, which may be rare or even unseen in the initial training corpus (Hu et al., 2019).

Domain adaptation for MT has benefited from pretraining via language models (Devlin et al., 2019; Lample and Conneau, 2019; Liu et al., 2020) trained on large quantities of monolingual text, therefore exposing the model to a wider vocabulary and improving cross-domain transfer (Clinchant et al., 2019; Verma et al., 2022). However, the model's capacity to exploit this underlying vocabulary is limited by the problem of catastrophic forgetting (Goodfellow et al., 2013) after fine-tuning (Hasler et al., 2021; Arthaud et al., 2021); the model becomes overly specific to the new data and loses the capacity to generalise to new domains.

A line of research with the aim of tackling this problem is the use of external resources such as bilingual lexicons and dictionaries (Tan et al., 2015; Dinu et al., 2019). These resources, comprising words or phrases and their translations (or words and their definitions in the case of dictionaries) provide a wider (and complementary) lexical coverage than the parallel training data. One aim is for the trained model to be able to exploit the external resource whenever a domain-specific or rare word appears. Different integration strategies have been proposed, including interpolation of translation probabilities and external lexicon probabilities (Arthur et al., 2016), the use of memory networks

*Work done at Inria, Paris, France

(Feng et al., 2017), constrained decoding (Hasler et al., 2018), and the inclusion of inline information such as translation candidates (Dinu et al., 2019; Pham et al., 2018; Niehues, 2021) and definitions (Zhong and Chiang, 2020).

In this work, we explore the last of these strategies: attaching additional information inline within the source sentence as a way of incorporating domain-specific translation knowledge. It is a simple and commonly used method in the literature and one that has been shown to work well in controlled language settings (i.e. where terms are known in advance and can be translated without ambiguity) (Dinu et al., 2019). Our aim is to explore how this strategy could work in a practical setting for cross-domain adaptation in the general translation setting, particularly when using pretrained language models that have seen a wider variety of vocabulary than those trained just on parallel data (the previous studies concentrate on vanilla MT). We try to gain some insights into how inline information is used, whether models are able to generalise, disambiguate between multiple candidate translations and how this can ultimately help cross-domain transfer. Our experiments on German-to-English (de→en) translation show that the use of the method in this more general (as opposed to controlled) setting is not so successful. Our results are largely negative; we can see small (although not systematic) improvements when applying a model to a new domain. However, we also analyse how the approach works; through a systematic analysis, we show that the approach is more than just a copy mechanism, as we see evidence for the inline translation candidates being used effectively by the model, even when distractor candidates are introduced. We also show that in a higher-resource setting with a more diverse training vocabulary and a larger lexicon, the method is less effective and therefore it is not a go-to method in all settings. Our code and outputs will be made publicly available.

2 Related Work

Different strategies exploiting bilingual lexicons and dictionaries have been developed in the past to handle rare words, the majority focusing on integrating bilingual lexicons containing word (or phrase) translation pairs (Song et al., 2019; Dinu et al., 2019; Duan et al., 2020). They differ from normal parallel data in that entries are shorter and they often cover domain-specific and rare vocabulary.

These strategies include but are not limited to adding lexicons to the parallel training data (Tan et al., 2015), combining translation and external lexicon probabilities (Arthur et al., 2016), using memory networks (Feng et al., 2017), constrained decoding (Hasler et al., 2018) and infixing of translation candidates within the source sentence (Pham et al., 2018; Dinu et al., 2019; Michon et al., 2020; Niehues, 2021). In this inline approach, the idea is to either add translations inline within the source sentences or to replace the terms with their translations. It has been shown to work well with controlled and non-ambiguous settings (Dinu et al., 2019; Niehues, 2021) and when using a mechanism to encourage annotation copying (Pham et al., 2018). A similar code-switching-inspired method was introduced by Song et al. (2019), whereby terms are replaced by their translations from bilingual lexicons, and the generated examples used as extra training data. Xu and Yvon (2021) also look at code-switched data, replacing terms with their translation equivalents. Similar strategies have been used elsewhere, for example Duan et al. (2020) integrate code-switching-style replacements using the bilingual lexicon in the back-translation step of an unsupervised MT model, and Junczys-Dowmunt and Grundkiewicz (2016) and Crego et al. (2016) augment sentences with fuzzy translation matches.

A few studies have looked into the use of dictionary definitions in MT, as opposed to bilingual lexicons. Zhong and Chiang (2020) use a method similar to Dinu et al. (2019), involving appending unknown words' definitions to source sentences and indicating through positional embeddings to which words the definitions are attached. Beyond MT, the use of dictionary definitions has also been investigated for word embedding creation: Bosc and Vincent (2018) by auto-encoding and reconstructing definitions to improve word embeddings and Shi et al. (2019) by using definitions as a bridge between translations. Theoretically, there is not a clear distinction between bilingual lexicons and bilingual dictionaries in that dictionary definitions often contain synonyms (corresponding to translations in the bilingual case). However, we would expect dictionary definitions to be descriptive rather than translations.[1] In this work, we use bilingual lexicons (containing possible translate candidates)

[1] A number of works (Arthur et al., 2016; Pham et al., 2018) use automatically constructed phrase tables as lexicons, which differ in that they often contain noisy candidates and many inflections, whereas lexicons are often restricted to lemmas.

rather than dictionaries, but where several possible candidates are present for each source word.

3 Integrating Lexicon Entries

We concentrate on the use of bilingual lexicons with word-candidate pairs to improve domain transfer in MT. Some examples of the bilingual lexicon entries are given in Table 1. Many of the entries contain a single translation for each term, but some of the terms have several possible translation candidates.

German term	English translation(s)
verehren	to carry a torch for [Am.] to adore, to enshrine, to revere, to venerate
wut	angriness, furiousness, fury, irateness, rabidness, rage, wrath
wälzlager	antifriction bearing, rolling contact bearing
biologisch abbaubar	biodegradable
tuberkulös	tuberculous

Table 1: Examples of bilingual lexicon entries.

Specifically, we consider a scenario where we train MT models to translate from German to English and attempt to transfer them to new domains by incorporating bilingual lexicon entries inline within source sentences (Pham et al., 2018; Dinu et al., 2019; Zhong and Chiang, 2020; Niehues, 2021).[2] We compare this to an alternative strategy, which is to concatenate the bilingual lexicon to the training data, i.e. treating it as additional parallel data, with the advantage that the entire lexicon can be used for training (rather than only the words that appear in the training data) but with the disadvantage that the method cannot generalise to novel lexicon entries.[3] In this sense, it may be seen as a model included for results comparison, but not one which could be considered a desirable alternative.

3.1 `inline`: Infixing Lexicon Entries within Source Sentences

We use the bilingual lexicon to provide context during training and at inference time for unknown or rare words. We do this by annotating identified terms in the source sentence with their corresponding target entries. For every word in the data that appears fewer than k times in the training data (i.e. the data on which the pretrained language model is fine-tuned),[4] we search for a corresponding lexicon entry to append inline to the term. Contrarily to (Niehues, 2021) and as in (Zhong and Chiang, 2020), we choose not to disambiguate the translation candidates and simply add the raw entry inline so that the model can learn to choose the most appropriate translation, potentially more appropriate in non-controlled language setting. Entries therefore resemble dictionary entries. An example of a German source sentence augmented with lexicon entries is shown in Example 1, with two rare words (underlined) and their translations according to the lexicon added within `<def></def>` tags.[5]

(1) **German source**: Begleittherapie Timolol kann mit anderen Arzneimitteln <def>pharmacotherapy</def> wechselwirken <def>interactively</def> (siehe Abschnitt 4.5)
English reference: Concomitant therapy Timolol may interact with other medicinal products (see section 4.5).

In order to expand the lookup in the lexicon beyond exact token matches,[6] we match rare words with lexicon terms by choosing the one with the shortest normalised Levenshtein distance. Similar to (Zhong and Chiang, 2020), to make this computation more efficient (by reducing the search space over the lexicon), we use locality-sensitive hashing (LSH) by creating vectors of all the lexicon headwords using their character-level trigrams. The rare words are then queried against the lexicon using the Jaccard[7] score character-level trigram overlap. The rare words that do not meet the Jaccard threshold will have no annnotation attached to them.

Including translations inline gives the model the potential to handle new entries. However, its main disadvantage is an increase in source sentence length, which can be problematic for models whose maximum sentence length is small.

3.2 `concat`: Using Bilingual Lexicons as Parallel Training Data

We compare this to the method of mixing the bilingual lexicon into the parallel training data. We consider two versions (see examples in Table 2): (i) `concat-diff`: mixing the data sources and prefixing each training instance with a different tag

[2] Unlike Pham et al. (2018), we do not force the model to copy the annotations and instead choose to explore the scenario where the model can learn to copy if relevant.

[3] Alternative fine-tuning strategies for continual learning would have to be used (Arthaud et al., 2021).

[4] A word is defined here as a token as obtained by the Moses tokeniser (Koehn et al., 2007).

[5] Note that, as shown in this example, the candidate translations do not always correspond to the reference translation. However, they may nevertheless provide lexical knowledge enabling the model to make a correct translation choice.

[6] We leave the multi-token matching to future work.

[7] We use a Jaccard similarity threshold score of 0.7.

Data	concat	concat-diff
Lexicon	**src**: transDeEn: beleuchtungstechnik **ref**: lighting technology	**src**: defDeEn: beleuchtungstechnik **ref**: lighting technology
Parallel	**src**: transDeEn: Schlucken Sie die Kapsel(n) als Ganzes mit einem Glas Wasser. **ref**: Swallow the capsule(s) whole with a glass of water.	**src**: transDeEn: Schlucken Sie die Kapsel(n) als Ganzes mit einem Glas Wasser. **ref**: Swallow the capsule(s) whole with a glass of water.

Table 2: concat strategy: mixing the two data sources (lexicon and parallel) without distinguishing their origin (concat) and with different tags indicating the data source (concat-diff).

indicating the data source and (i) concat: mixing the two data sources together without distinguishing the two sources. The hypothesis is that this could help the model distinguish the two data types as previously seen for domain labels (Kobus et al., 2017; Caswell et al., 2019) and politeness (Sennrich et al., 2016). Note that in practice, we use the prefix tranDeEn: for source sentences of all models (including inline), except for concat-diff, where the prefix defDeEn: is used for lexicon entries.

4 Experimental Setup

4.1 Data

Training Data We cover four different domains: biomedical, commerce, news and films, using data from EMEA,[8] ECB,[9] GlobalVoices,[10] and OpenSubtitles2018[11] (Lison et al., 2018) from OPUS (Tiedemann, 2012). Pre-processing includes fixing orthographic errors, removing duplicate parallel sentences, and filtering via language identification with Bifixer/Bicleaner (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020) and FastText (Joulin et al., 2016; Joulin et al., 2017). Table 3 shows the dataset sizes after pre-processing.

Validation and Test Data From each dataset, we split off distinct 2000 random sentence pairs from the pre-processed data for each of the validation and test sets. We also test on other datasets: the WMT 2018 and 2020 news test sets, the WMT 2018 biomedical test set and the different genres from the 2022 WMT General MT task: news, e-commerce, social, and chat (see Table 3).

Bilingual Lexicon We use the Stardict German-English dictionary (based on Freedict[12] and originally with 81,628 entries). We preprocess the lexicon by removing empty entries, lower-casing German headwords, concatenating multiple candidates of a same headword, and deleting bracketed descriptions. The final lexicon has 79,936 entries (German headwords) associated with one or more English translations (see examples of the preprocessed entries in Table 1). In concat approaches (treating the lexicon as parallel data), we filter out 150 lexicon entries to use as a development set.

Source	Domain	Train	Dev.	Test
GlobalVoices	news	$\sim$ 61k	2k	2k
ECB	commerce	$\sim$76k	2k	2k
EMEA	medical	$\sim$235k	2k	2k
Opensubtitles	movies	$\sim$16M	2k	2k
WMT				
News_{18}	news	–	–	2998
News_{20}	news	–	–	785
News_{22}	news	–	–	506
Medline_{20}	medical	–	–	404
eCom_{22}	commerce	–	–	501
Soc_{22}	social	–	–	515
Conv_{22}	conversation	–	–	462

Table 3: #sentences per dataset per domain.

4.2 Training Setup

We initialise all MT models using the pre-trained multilingual language model mT5-base (Xue et al., 2021), implemented in Transformers (Wolf et al., 2020).[13] We train the models for up to 40 epochs with a batch size of 10, a learning rate of 5e-5, dropout of 0.1, and a maximum source and target length of 512. For decoding, we use a beam of 10. The output of the best checkpoint (according to the training loss) is evaluated using BLEU (Papineni et al., 2002) as computed by SacreBLEU[14] (Post, 2018). We choose to use BLEU for evaluation because we observe similar trends with other metrics such as COMET (Rei et al., 2020), and BLEU has the advantage of having more easily interpretable

[8] European Medicines Agency: https://www.ema.europa.eu

[9] European Central Bank: https://www.ecb.europa.eu

[10] GlobalVoices: https://globalvoices.org

[11] https://www.opensubtitles.org

[12] https://freedict.org/

[13] https://github.com/huggingface/transformers

[14] case:mixed|eff:no| tok:13a|smooth:exp|v:2.3.1

Setup	EMEA	ECB	GV	News18	News20	Med20	News22	eCom22	Soc22	Conv22
Trained on Globalvoices										
Baseline	**21.1**	**19.2**	32.0	33.5	23.4	24.5	**22.3**	22.3	**23.1**	23.8
`concat-diff`	20.6	18.9	31.7	33.5	23.2	**24.6**	21.5	22.8	21.8	**24.0**
`concat`	20.6	19.1	31.6	33.1	23.4	24.2	21.5	22.8	22.0	23.9
`inline`	20.9	18.8	**32.1**	**33.7**	**24.1**	24.5	21.8	22.7	22.8	23.8
`inline+concat-diff`	20.5	18.6	31.7	33.2	23.2	24.2	21.5	**23.1**	21.9	23.8
`inline+concat`	20.6	19.1	31.9	33.5	23.6	20.6	21.8	22.6	21.7	22.8
Trained on ECB										
Baseline	16.8	52.2	21.1	24.6	**18.7**	20.9	17.7	19.7	17.2	19.8
`concat-diff`	19.6	**52.9**	21.5	25.6	18.3	**21.9**	18.5	20.7	18.1	**21.1**
`concat`	19.4	52.6	21.6	25.7	18.7	22.8	18.3	21.4	18.4	20.6
`inline`	19.1	52.2	21.3	25.4	18.1	20.3	18.3	20.4	17.2	19.3
`inline+concat-diff`	**20.6**	52.6	**21.8**	26.1	18.3	21.4	**19.0**	21.2	18.3	18.8
`inline+concat`	20.3	52.4	21.7	**26.2**	17.3	21.5	18.7	**21.4**	**18.5**	18.2
Trained on EMEA										
Baseline	64.7	18.2	15.9	19.2	12.2	28.2	14.4	17.8	12.9	15.7
`concat-diff`	65.1	18.7	17.2	21.1	13.9	28.1	15.3	**19.0**	14.8	17.1
`concat`	**65.2**	18.8	17.1	20.8	12.1	27.8	15.9	18.9	14.8	17.1
`inline`	64.9	18.2	16.6	19.5	**13.2**	28.3	15.0	17.6	13.5	15.6
`inline+concat-diff`	64.9	**19.0**	17.4	**21.4**	11.8	**28.4**	**16.4**	18.4	**15.8**	**17.6**
`inline+concat`	64.9	18.9	**17.4**	21.4	12.0	**28.4**	16.3	18.3	15.1	17.1

Table 4: BLEU scores of each domain-specific model on each of the test sets. The coloured cells indicate that the training and test data are from a similar domain. The highest BLEU score for each model on each test set is marked in bold.

	Real Definition	Fake definition
Source	Sie haben zur Befestigung ein 16mm Hülse als Anschluß, damit können Sie direkt an Ihr **Fotostativ** **<def>a photo tripod</def>**.	Sie haben zur Befestigung ein 16mm Hülse als Anschluß, damit können Sie direkt an Ihr **Fotostativ** **<def>green box</def>**.
ECB	You have a 16 mm sleeve for attaching it so you can attach it directly to your photo tripod.	You have a 16 mm sleeve for attaching it so you can attach it directly to your photo stative.
EMEA	You have a 16 mm needle attached to it so that you can directly attach it to your photogravure.	You have a 16 mm needle attached to it so that you can directly attach it to your photogravure.
GlobalVoices	They have a 16mm housing so you can hang it directly on your photo tripod.	They have a 16mm housing so you can hang directly on your photo stative.
Source	Immer neue **Omikron-Fälle** **<def>a variant of corona virus</def>** besorgen Politik und Wissenschaft in Großbritannien.	Immer neue **Omikron-Fälle** **<def>green box</def>** besorgen Politik und Wissenschaft in Großbritannien.
ECB	Policy and science in the UK are providing every new case of Omikron.	Each new case of Omikron provides policy and science in the UK.
EMEA	Manage new cases of Omicron in the UK, policy and science in the UK.	Manage new cases of Omicron in the context of policy and science in the UK.
GlobalVoices	New cases of Omicron are increasingly affecting Britain's politics and science.	New micron cases are increasingly creating a boost to Britain's politics and science.

Table 5: Examples of `inline` outputs created during our manual analysis of actual and fake annotations.

absolute scores. We train on each of the training sets in Table 3 and evaluate each one on all test sets.

5 Results

To test `inline`'s ability to transfer to new domains, we train one model per training dataset and evaluate on all test sets. We compare to `concat` approaches and to a baseline that does not use the lexicon, trained and evaluated in the same way. Given that `inline` only sees the lexicon words seen in the training data, we also test a hybrid approach involving training on the concatenation of both data sources and then fine-tuning using the inline approach. We compare a total of five models:

- `baseline` (no lexicon)
- `concat-diff`: concatenate lexicon and parallel data, with different prefixes
- `concat`: concatenate lexicon and parallel data
- `inline`: target lexicon entries are inserted inline into the source sentence
- `inline+concat-diff` and `inline+concat`: combinations of `inline` and either `concat-diff` or `concat`.

Results are shown in Table 4.[15] As expected, all baseline models perform well on data from the same

[15] Similar trends were seen using COMET (Rei et al., 2020) and

domain as the training data and struggle when tested on data from different domains. For example, the EMEA model has scores of 64.70 and 28.18 on the EMEA and Med20 test sets respectively, whereas it obtained less than 20 BLEU points on the other test sets. This supports the idea that NMT models are sensitive to out-of-domain data, as previously seen (Koehn and Knowles, 2017).

Compared to the baseline, both `concat` approaches improve the EMEA and ECB models' performance by at least 1 BLEU on a majority of the test sets from different domains. However, they do not provide any gains to the GlobalVoices model's performance on other domains. This may be because of the small size of the GlobalVoices training data (the bilingual lexicon contains 30k more examples and so possibly outweighs it). The `inline` model trained on GlobalVoices does not show improved performance on most of the test sets either. However, similar to the `concat` models' results, there was at least +0.5 BLEU when transferring from ECB→{EMEA,News18,News20,eComm22} and EMEA→{GV,News20,News22,Soc22}.

These results indicate that there is some evidence for cross-domain transfer for both approaches, which show small improvements for the ECB and EMEA models when evaluated on a different domain (although GlobalVoices models show little improvement, possibly due to the small dataset size). However, there is little improvement when these models are tested on the data from the same source as the training data (e.g. EMEA→EMEA and ECB→ECB). The hybrid approaches show some benefits over the individual methods in several cases especially for `inline+concat-diff`.

6 Going Further: When are Inline Definitions Used?

These results show that the inline approach leads to slight improvements in translation performance in some cases and does not improve in others. Examples 2-4 from the EMEA test sets (using the ECB-trained model) illustrate how attaching the candidates inline can sometimes be effectively used in the generated hypothesis and sometimes not. The models fail to use the annotations in Example 2, while they are partially and fully used in Examples 3 and 4 respectively.

(2) **Source:** transDeEn: - können Sie schwere *Migräne* <def>migraine</def> bekommen.
Target: - you may develop a severe migraine.
Baseline: - you can be vulnerable to severe crises.
`inline`: - you can become vulnerable to severe migration.

(3) **Source:** transDeEn: NovoMix 70 Penfill Patronen dürfen nicht wieder *aufgefüllt* <def>filled up, refilled, replenished</def> werden.
Target: Do not refill NovoMix 70 Penfill cartridges.
Baseline: Novo mix 70 penfill patrones must not be re-filled.
`inline`: Novo mix 70 penfills cannot be refilled.

(4) **Source:** transDeEn: Es enthält den *Wirkstoff* <def>active agent</def> Docetaxel.
Target: It contains the active substance docetaxel.
Baseline: contains Docetaxel.
`inline`: It contains the active agent Docetaxel.

We did some initial experimentation with the inline models by manually sampling examples from the test sets and creating hypothetical test examples (either with manually created correct translations or invented (incorrect) translations). A few such examples are shown in Table 5, whereby the fake candidate translations are simply composed of the word "green box". This preliminary analysis shows that rather than blindly copying, the models seem to make selective use of the definitions, which leads us to conduct a more systematic analysis.

6.1 Experimental Settings

We provide a more systematic analysis by creating artificial test cases, where we modify the inline translation candidates either by (i) replacing them with random translation candidates and (ii) prepending or appending the random candidates to the true ones. We show results for `inline` trained on ECB data and testing on EMEA, although we see similar results across the other models and test sets.

Rather than taking truly random contrastive translation candidates, we select random candidates amongst those whose headword matches the part of speech (POS) tag of the annotated source word.[16] To ensure the definitions are not too long, we only prepend/append alternative candidates containing a maximum of 4 tokens.

The four setups are illustrated in Examples 5-8:

(5) Original (green):
Source: transDeEn: Was Xagrid enthält Der *Wirkstoff*

will include these results in the appendix. We report BLEU instead of COMET since the conclusions are the same for the two metrics. COMET is better correlated with human judgments and is recommended by (Alam et al., 2021) for evaluation terminology translation, but BLEU is more tangible, so readers familiar with MT can get a better appreciation of absolute quality.

[16] In practice, we apply the POS tagger to the training data to determine the POS tag of potential headwords.

<def>active agent</def> ist Anagrelid.
Target: What Xagrid contains The active substance is anagrelide.

(6) Random replacement (underlined, red):
Source: transDeEn: Was Xagrid enthält Der *Wirkstoff* <def>economics</def> ist Anagrelid.

(7) Random prepended (underlined, red):
Source: transDeEn: Was Xagrid enthält Der *Wirkstoff* <def>veep, vice president, active agent</def> ist Anagrelid.

(8) Random appended (underlined, red):
Source: transDeEn: Was Xagrid enthält Der *Wirkstoff* <def>active agent, veep, vice president</def> ist Anagrelid.

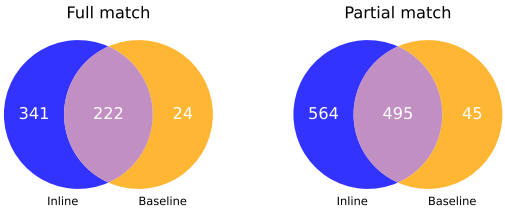
In order to approximate whether the model is using the candidate translations in the inline annotations, for each annotated source word, we count the number of times the candidate annotation appears in the resulting translation outputs. We acknowledge the limitations of this approach: (i) we may get false positives when the candidate term appears elsewhere in the translation (and not as a translation of the annotated word), but these instances should be few given the rarity of the words in question, and (ii) as shown in Example 1, there are cases where the candidates do not appear in the reference at all. Nevertheless, this method gives us a way of getting a global picture of what is going on, particularly when it comes to copying behaviour. Since there can be multiple candidates, as well as multi-word candidates, we count the number of exact matches (the whole annotation appears) and partial matches, i.e. where one of the (comma-separated) candidates exists.[17]

6.2 Analysis Results

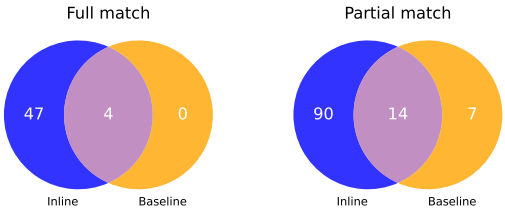
Do the models make use of the definitions? From our analysis using the manually and systematically created examples we found that these models make use of the definitions attached to unknown and rare words. However, we also found that the models use definitions that do not fit into the context of the input sentences rarely, at least far more frequently than for real definitions.

How often are the translation candidates used? Figure 1a shows how often the original candidate translations are used, either fully or partially. The full annotations appeared 563 times in the output, of which 222 were also in the baseline output. Importantly, a far higher number of candidates (341) only

[17]For partial matches, we remove stopwords such as *the* and *to* from definitions.



(a) Translation candidates from the lexicon



(b) Randomly selected translation candidates

Figure 1: Frequency of translation candidates used in the `inline` outputs compared to the baseline.

appeared in the `inline` outputs compared to the baseline ones (24), showing that `inline` is learning to copy the candidates. We see the same pattern for partial matches (one of the multiple candidates), again with the `inline` outputs containing far more candidate translations. This trend was consistent across the models and test sets that we analysed.

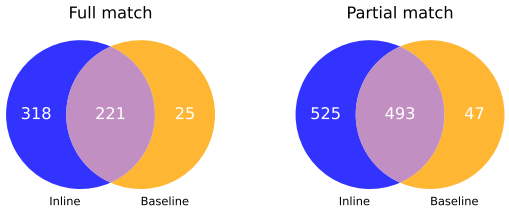
Figure 1b shows the number of candidates in the outputs when they are replaced by random (incorrect) annotations. The results indicate that the models rarely employ the incorrect definitions (i.e. they learn to discriminate between useful and irrelevant annotations). In fact, only 51 (for exact annotations) and 104 (for partial annotations) instances were detected in the `inline` translation outputs.

Can the model avoid distractor annotations? Instead of just replacing the annotation with a random replacement, we also analyse the setup where we combine the original annotations with the random ones (by prepending or appending). The results being very similar for the two cases, we only show results for the case of appending. Figure 2a shows the number of times the original annotations appear in the model outputs and 2b the number of times the distractor annotation appears. The pattern is the same as in Figure 1; the models rarely use the distractor annotations and although the number of true translation candidates decreases a little when distractors are used, the models is largely able to select and use the true annotations.

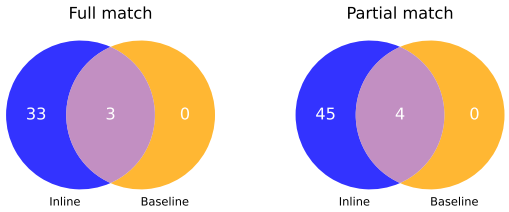
Evaluation in a higher-resource setting We also evaluate the methods in a higher-resource setting

Setup	OpenSubs	EMEA	ECB	GV	News18	News20	Med20	News22	eCom22	Soc22	Conv22
Trained on Multi-domain/high-resource											
Zero-shot	**33.6**	18.0	16.3	28.2	36.4	24.9	20.8	21.8	23.4	22.1	20.5
Baseline	32.6	50.1	**42.2**	32.2	38.2	**29.1**	31.4	**24.4**	24.9	**24.9**	**23.8**
`concat-diff`	32.4	50.2	42.1	32.2	**38.6**	29.0	31.1	24.3	24.8	24.7	23.6
`concat`	32.5	50.1	42.1	32.2	38.4	29.0	31.0	24.3	24.6	24.8	23.0
`inline`	32.6	**50.3**	42.1	**32.3**	38.4	28.6	**33.1**	24.3	**25.4**	**24.9**	23.4
`inline+concat-diff`	32.4	50.2	42.1	**32.3**	38.4	28.8	32.4	**24.4**	25.3	24.8	23.7
`inline+concat`	32.4	**50.3**	**42.2**	**32.3**	38.5	28.9	32.5	24.0	25.3	24.8	23.3

Table 6: BLEU scores of the general/multi domain model on each of the test sets. The highest BLEU score for each model on each test set is marked in bold.



(a) Translation candidates from the lexicon



(b) Randomly selected translation candidates

Figure 2: Frequency of translation candidates used in the `inline` outputs (vs. baseline outputs) when appending random candidates.

with access to a wider vocabulary and from a multi-domain setting. Instead of just fine-tuning on the domain-specific training sets, we fine-tune mt5 in several steps: (i) firstly on data from OpenSubtitles2018[18] (Lison et al., 2018) for one epoch (due to its substantial size) and then (ii) on a combination of the EMEA, ECB, and GlobalVoices datasets and 250k randomly sampled parallel sentences from OpenSubtitles to avoid overfitting. We also use a larger lexicon; we extracted and cleaned a bilingual lexicon from Wiktionary[19] and merged it with Freedict.[20] For `inline`, words are considered unknown if they appear fewer than 20 times in the combined training data (from OpenSubtitles, EMEA, ECB, and GlobalVoices). Similar to the previous experiments, we created LSH using a threshold 0.6.

As previously, we report automatic scores (see Table 6) and our automatic analysis of matching words (see Figure 4). None of the methods outperform the baseline model. However the count statistics show that these models still use relevant entries and ignore irrelevant ones, but to a lesser extent than in the lower-resourced setting.

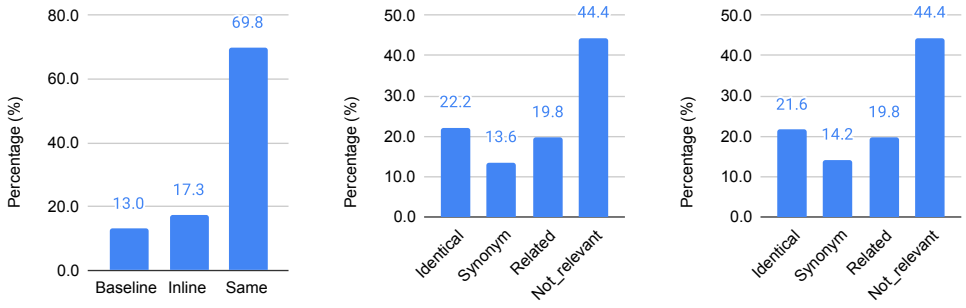
We also conducted a human evaluation involving two annotators with the aim of answering three questions: firstly, to confirm our automatic analysis, (i) which model output is better between the baseline and `inline`? and (ii) are the terms present in the source side of the `inline` model more present in the outputs than in the baseline? and finally, (iii) what sort of errors can we see? We focused on examples from the Med20 dataset where the `inline` appeared to exhibit better performance than the baseline. We selected all sentences with a single annotation, resulting in 81 distinct examples. We see (Figure 3) that a majority of translations were of the same quality, with a slight preference for `inline` (+4.32% over the baseline). We also observed a similar trend in how `inline` translations related to `inline` outputs compared to the baseline (despite the baseline not having access to them), suggesting that the information is rarely being used in this higher-resource setting, given the similarity in the behaviour of the two models.

Finally, we observed some limitations in the LSH method, whereby a large number of term translations were incorrect with respect to the annotated term ("Not related" category). This is likely to be exaggerated with respect to our main results using Freedict due to the lexicon being larger and less clean. This highlights an interesting point: the inaccuracy of LSH matching, which is likely to be a reason for the model learning to copy in some instances and not in others (i.e. the behaviour seen in our main results), is likely to lead to term translations not being used when the effect is too great. Neither baseline nor `inline` translations were per-

[18] `http://www.opensubtitles.org`

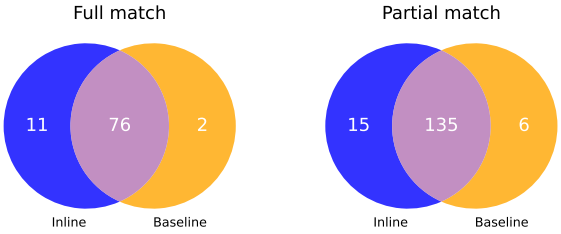
[19] Using the procedure described at `http://en.wiktionary.org/wiki/User:Matthias_Buchmeier`.

[20] We omitted Wiktionary in our main experiments due to its comparatively noisy nature compared to Freedict.



(a) Which system is better? (b) Relation (Baseline output) (c) Relation (`inline` output)

Figure 3: Human evaluation results of Med20 test set translations using the higher-resource multi-domain models. Figures 3b and 3c refer to the relation between annotated source words and their translations.



(a) Translation candidates from the lexicon

Figure 4: Frequency of translation candidates used in the `inline` outputs (vs. baseline outputs) in a multi-domain and high-resource setting.

fect, with many remaining term problems, so it appears that there is still research to be done on improving the approach.

7 Conclusions

Our study focuses on a simple method of incorporating lexical knowledge from bilingual lexicons into NMT models for cross-domain transfer: infixing translation candidates to rare terms within source sentences. We compare to using lexicon entries as additional parallel training data. We show that lexicons can sometimes help cross-domain transfer, but the gains seen (according to automatic metrics) are limited and appear to diminish in higher-resource scenarios. This is in contrast to its previous successful use in controlled language settings, showing that it is not such a promising approach in the general translation setting. Our analysis of the model outputs using distractor term translations showed that, despite the small difference in scores, the models make use of these definitions and they importantly can learn to ignore irrelevant definitions rather than blindly copying entries. However, the method is far from being as successful for this cross-domain setup as in the controlled language settings in which the method was developed, and experiments on a higher-resource language setting show that the approach does not have a huge effect to performance compared to a strong baseline.

Ethical Considerations and Limitations

There are several limitations of this work and directions for future research. Firstly, we focus on one particular language pair and leave testing in a multilingual setting to future work. In terms of the bilingual lexicons we used, we were limited to a lexicon containing fewer than 150,000 entries, along with some inherent noise in its contents. We hope that future research efforts will focus on expanding bilingual lexicon resources for a wider range of languages, particularly those with limited linguistic resources, and we see promise for studying these strategies in lower-resource scenarios. Also in this work, we associated unknown words with candidate translations using the previously proposed LSH method without any contextual information with the aim of seeing how this method could work in our domain transfer setting. We have shown that this method is insufficient and most likely led to an excess of noise in the annotations for the higher-resource scenario. In future work we could also focus on better methods for annotating the data.

Acknowledgements

Both authors' contributions were funded by R. Bawden's Emergence project, DadaNMT, funded by Sorbonne Université. R. Bawden was also funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Akhbardeh, Farhad, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November. Association for Computational Linguistics.

Alam, Md Mahfuz Ibn, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online, November. Association for Computational Linguistics.

Arthaud, Farid, Rachel Bawden, and Alexandra Birch. 2021. Few-shot learning through contextual data augmentation. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1049–1062, Online, April. Association for Computational Linguistics.

Arthur, Philip, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In Su, Jian, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November. Association for Computational Linguistics.

Barrault, Loïc, 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 Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Bogoychev, Nikolay and Pinzhen Chen. 2021. The highs and lows of simple lexical domain adaptation approaches for neural machine translation. In Sedoc, João, Anna Rogers, Anna Rumshisky, and Shabnam Tafreshi, editors, *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 74–80, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Bosc, Tom and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium, October-November. Association for Computational Linguistics.

Caswell, Isaac, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August. Association for Computational Linguistics.

Clinchant, Stephane, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In Birch, Alexandra, Andrew Finch, Hiroaki Hayashi, Ioannis Konstas, Thang Luong, Graham Neubig, Yusuke Oda, and Katsuhito Sudoh, editors, *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong, November. Association for Computational Linguistics.

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan

Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran's pure neural machine translation systems.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *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, June. Association for Computational Linguistics.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July. Association for Computational Linguistics.

Duan, Xiangyu, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual dictionary based neural machine translation without using parallel sentences. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Online, July. Association for Computational Linguistics.

Feng, Yang, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In Palmer, Martha, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark, September. Association for Computational Linguistics.

Goodfellow, Ian J., Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Hasler, Eva, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In Walker, Marilyn, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hasler, Eva, Tobias Domhan, Jonay Trenous, Ke Tran, Bill Byrne, and Felix Hieber. 2021. Improving the quality trade-off for neural machine translation multi-domain adaptation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8470–8477, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Hu, Junjie, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy, July. Association for Computational Linguistics.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jörg Tiedemann, and Marco Turchi, editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany, August. Association for Computational Linguistics.

Kobus, Catherine, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In Mitkov, Ruslan and Galia Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria, September. INCOMA Ltd.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Luong, Thang, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Ananiadou, Sophia, editor, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Lample, Guillaume and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, January.

Lison, Pierre, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Liu, Yinhan, 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.

Michon, Elise, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In Scott, Donia, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Niehues, Jan. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In Merlo, Paola, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online, April. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pham, Ngoc-Quan, Jan Niehues, and Alexander Waibel. 2018. Towards one-shot learning for rare-word translation with external experts. In Birch, Alexandra, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 100–109, Melbourne, Australia, July. Association for Computational Linguistics.

Pham, Minhquang, Josep Maria Crego, and François Yvon. 2021. Revisiting multi-domain machine translation. *Transactions of the Association for Computational Linguistics*, 9:17–35.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Ramírez-Sánchez, Gema, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal, November. European Association for Machine Translation.

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

Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels, October. Association for Computational Linguistics.

Saunders, Danielle. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *CoRR*, abs/2104.06951.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In Knight, Kevin, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Shi, Weijia, Muhao Chen, Yingtao Tian, and Kai-Wei Chang. 2019. Learning bilingual word embeddings using lexical definitions. In Augenstein, Isabelle, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei, editors, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 142–147, Florence, Italy, August. Association for Computational Linguistics.

Song, Kai, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *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 449–459, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tan, Liling, Josef van Genabith, and Francis Bond. 2015. Passive and pervasive use of bilingual dictionary in statistical machine translation. In Babych, Bogdan, Kurt Eberle, Patrik Lambert, Reinhard Rapp, Rafael E. Banchs, and Marta R. Costa-jussà, editors, *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, pages 30–34, Beijing, July. Association for Computational Linguistics.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Verma, Neha, Kenton Murray, and Kevin Duh. 2022. Strategies for adapting multilingual pre-training for domain-specific machine translation. In Duh, Kevin and Francisco Guzmán, editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA, September. Association for Machine Translation in the Americas.

Vu, Thuy-Trang, Xuanli He, Dinh Phung, and Gholamreza Haffari. 2021. Generalised unsupervised domain adaptation of neural machine translation with cross-lingual data selection. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3335–3346, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Liu, Qun and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Xu, Jitao and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In Solorio, Thamar, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors, *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online, June. Association for Computational Linguistics.

Xue, Linting, 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. In Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Zhong, Xing Jie and David Chiang. 2020. Look it up: Bilingual dictionaries improve neural machine translation.

Setup	EMEA	ECB	GV	News18	News20	Med20	News22	eCom22	Soc22	Conv22
Trained on Globalvoices										
Baseline	0.756	0.759	0.846	0.818	0.778	0.772	0.781	0.771	0.778	0.801
`concat-diff`	0.758	0.757	0.848	0.819	0.777	0.776	0.778	0.775	0.776	0.801
`concat`	0.756	0.758	0.846	0.819	0.776	0.764	0.778	0.773	0.776	0.799
`inline`	0.761	0.761	0.848	0.821	0.784	0.781	0.781	0.774	0.782	0.797
`inline+concat-diff`	0.760	0.759	0.847	0.820	0.779	0.781	0.779	0.772	0.775	0.791
`inline+concat`	0.759	0.760	0.847	0.820	0.777	0.760	0.777	0.775	0.774	0.795
Trained on ECB										
Baseline	0.709	0.843	0.772	0.759	0.727	0.750	0.727	0.745	0.711	0.755
`concat-diff`	0.733	0.844	0.782	0.774	0.730	0.776	0.741	0.762	0.731	0.765
`concat`	0.732	0.843	0.785	0.775	0.737	0.773	0.739	0.762	0.731	0.761
`inline`	0.721	0.843	0.778	0.769	0.729	0.750	0.732	0.749	0.716	0.755
`inline+concat-diff`	0.738	0.843	0.786	0.780	0.734	0.766	0.744	0.759	0.733	0.756
`inline+concat`	0.739	0.843	0.788	0.780	0.733	0.771	0.747	0.757	0.735	0.764
Trained on EMEA										
Baseline	0.877	0.717	0.696	0.687	0.636	0.774	0.658	0.726	0.649	0.671
`concat-diff`	0.878	0.730	0.722	0.718	0.656	0.775	0.688	0.737	0.684	0.714
`concat`	0.878	0.729	0.724	0.716	0.653	0.775	0.687	0.741	0.685	0.715
`inline`	0.878	0.721	0.712	0.702	0.651	0.793	0.673	0.729	0.665	0.681
`inline+concat-diff`	0.878	0.734	0.733	0.727	0.659	0.777	0.696	0.743	0.695	0.721
`inline+concat`	0.878	0.730	0.735	0.727	0.659	0.781	0.699	0.747	0.696	0.724

Table 7: COMET scores of each domain-specific model on each of the test sets. The coloured cells indicate that the training and test data are from a similar domain.

A COMET scores for main results

Table 7 shows results using the COMET metric (Using the default model `Unbabel/wmt22-comet-da`.) (Rei et al., 2020) for the main results shown in Table 4. The trends we see are the same between the BLEU and COMET scores.