Document-Level Adaptation for Neural Machine Translation

Sachith Sri Ram Kothur*
Dept. of Computer Science
Johns Hopkins University
kothursachith@gmail.com

Rebecca Knowles*
Dept. of Computer Science
Johns Hopkins University
rknowles@jhu.edu

Philipp Koehn
Dept. of Computer Science
Johns Hopkins University
phi@jhu.edu

Abstract

It is common practice to adapt machine translation systems to novel domains, but even a well-adapted system may be able to perform better on a particular document if it were to learn from a translator's corrections within the document itself. We focus on adaptation within a single document – appropriate for an interactive translation scenario where a model adapts to a human translator's input over the course of a document. We propose two methods: *single-sentence adaptation* (which performs online adaptation one sentence at a time) and *dictionary adaptation* (which specifically addresses the issue of translating novel words). Combining the two models results in improvements over both approaches individually, and over baseline systems, even on short documents. On WMT news test data, we observe an improvement of +1.8 BLEU points and +23.3% novel word translation accuracy and on EMEA data (descriptions of medications) we observe an improvement of +2.7 BLEU points and +49.2% novel word translation accuracy.

1 Introduction

The challenge of adapting to a new domain is a well-studied problem in machine translation research. But even within a particular domain, each new document may pose unique challenges due to novelty of vocabulary, word senses, style, and more.[1] It stands to reason that fine-grained adaptation using sentences from within a document (for example, as it is being translated by a human translator in a computer aided translation (CAT) environment) could provide the added benefit of a closer in-domain match than existing approaches that use data from other documents within the same domain. We propose two complementary approaches to the treatment of novel words and fine-grained document-level adaptation of machine translation systems, and show that the combination of approaches outperforms each approach individually, resulting in BLEU point improvements of +1.8 and +2.7 across two domains, in addition to demonstrating improvements in novel word translation accuracy.

As Carpuat (2009) observed, there is a tendency for translators to produce translations such that the "one translation per discourse" hypothesis holds within a particular document.[2] That is, human translators tend to prefer consistent translations of individual terms throughout a document. Other work on "translationese" has also found that translations show regularities in syntax and punctuation (Baroni and Bernardini, 2005). Thus, even expanding beyond words with multiple senses, we expect that learning from the translator's lexical, syntactic, and stylistic choices at the beginning of a document should result in a well-tailored system that is better at translating subsequent sentences. We can think of fine-grained adaptation over a document as producing a document-specific machine translation system that encodes or highlights document context.

Continued training of neural machine translation (NMT) systems has been shown to be an effective and efficient way to tune them for a specific target domain (Luong and Manning, 2015). One such technique is incremental updating – comparing the system's predicted translation of an input sentence to a reference translation and then updat-

*These authors contributed equally to this work.

[1] Carpuat et al. (2012) decompose errors into seen, sense, score, and search; the first two are most relevant to our work.

[2] This work follows from "one sense per discourse" (Gale et al., 1992), which found that the vast majority of polysemous words share only one sense within a given document.

Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 64–73
Melbourne, Australia, July 20, 2018.

Source	Reference	Baseline MT Output
Ambirix (Ambi/rix)	Ambirix (Ambi/rix)	Hampshire, Glaurix, Tandemrix, ...
Prepandemic (Prep/an/demic)	Präpandemischer (Prä/pandem/ischer)	Proteasehemmer
Cataplexy (Cat/ap/lex/y)	Kataplexie (Kat/ap/lex/ie)	Cataplexy
hormone-dependent (hormon/e-/dependent)	hormonabhängig (hormon/abhängig)	hormonell

Table 1: Examples of novel words and their mistranslations. The subword segmentation (in parentheses) is indicated by "/" for the source and reference.

ing the model parameters to improve future predictions. Though this is typically done in batches during training, a single sentence pair or even a word and its translation can be treated as a training instance.

Computer aided translation provides an ideal use case for exploring model adaptation at such a fine granularity. As a human translator works, each sentence that they translate (or each novel word for which they provide a translation) can then be used as a new training example for a neural machine translation system. In an interactive translation setting or a post-editing scenario, rapid incremental updating of the neural model will allow the neural system to adapt to an individual translator, a particular new domain, or novel vocabulary over the course of a document.

In an open-vocabulary NMT system that uses byte-pair encoding (Sennrich et al., 2016b), tokens that were never seen in training data are represented as sequences of known subword units. These may sometimes be successfully translated (or copied, subword by subword, when appropriate) on the first try, but sometimes systems generate incorrect translations or even nonsensical words. Table 1 shows example mistranslations of novel words.

We test our two complementary approaches to document-level NMT adaptation (dictionary training and single-sentence adaptation) on two very different domains: news and formal descriptions of medications, each of which provide their own challenges. In our datasets, just under 80% of news documents and just over 90% of medical documents contain at least one word that was unobserved in the training data. In the news documents, 12.8% of lines contain at least one novel word, whereas in the medical data, 38.3% of lines contain at least one novel word. We show that models can learn to correctly translate novel vocabulary items and can adapt to document-specific terminology usage and style, even in short documents.

2 Related Work

This work relates closely to three lines of research on neural machine translation models: rare word translation, copying mechanisms, and domain adaptation. Concerns about rare words and copying mechanisms are closely linked; words that need to be copied (or nearly copied) are often proper names or technical vocabulary, which may be infrequent or unobserved in training data.

Arthur et al. (2016) propose to improve the translation of rare (low-frequency) content words through the use of translation probabilities from discrete lexicons. Nguyen and Chiang (2018) propose to train a feed-forward neural network to generate a target word based directly on a source word. Both then weight these probabilities using the attention mechanism and combine them with the standard translation approach. Gu et al. (2016) propose a (monolingual) sequence-to-sequence model, COPYNET, that can select input sequences to copy to the output within the course of generating a single sequence. All of these approaches require modifications to the neural network architecture. Additionally, some require knowledge of the rare words during training, meaning they are inapplicable to novel words.

By modifying the available training data rather than the neural architecture, Currey et al. (2017) find that training a neural machine translation system to do both translation and copying of target language text improves results on low-resource neural machine translation and learns to pass untranslated words through to the target. They do this by mixing monolingual target data (as source-target pairs) with parallel training data. In contrast, Khayrallah and Koehn (2018) find that this dramatically hurts performance (in a higher-resource setting). Ott et al. (2018) provide additional analysis of copying behavior. Fadaee et al. (2017) propose to learn better translations of rare words by generating new sentences that include them to add to the training data.

Domain adaptation has long been an area of in-

terest for researchers in the machine translation community and is relevant both to the translation of new words and to more general improvements in translation quality. Recent work (Freitag and Al-Onaizan, 2016; Luong and Manning, 2015) has proposed to do domain adaptation for NMT systems by training a general system then fine-tuning by continuing to train using only in-domain data (typically a smaller dataset). Wang et al. (2017) present a similar approach where they weight each source-target sentence pair during training based on scores from in-domain and out-of-domain language models. Kobus et al. (2017) use special tokens to indicate domain. Chu et al. (2017) compare the approaches. These approaches typically use larger amounts of in-domain data to do adaptation, far greater than the amounts that might be available in a CAT setting. Cettolo et al. (2014) proposed adapting statistical phrase-based machine translation systems to particular projects (multiple documents) and Peris and Casacuberta (2018) propose adapting neural machine translation systems in CAT settings. Neither explore very small amounts of data at the sub-document level.

Two recent papers have tried a domain adaptation approach using very small data sizes, ranging from 1 sentence to 128 sentences (Farajian et al., 2017; Li et al., 2016). They adapt models for new sentences by training on sentence pairs from a training corpus (or translation memory) that are similar to the new sentence, which means they cannot adapt to novel vocabulary.

3 Approaches

We propose two complementary approaches for adapting an NMT model over the course of a single document's translation and the combination of the two. For each approach, adaptation is done at a document level and the model is reset to baseline between documents.[3]

3.1 Single-Sentence Adaptation

In this approach, the model is iteratively adapted over the previous translated sentence (and its reference), then the updated model is used to translate the next sentence. Thus, line n of the document is translated by a model which has been incrementally adapted to all previous lines (1 through $n-1$) of the document. See Algorithm 1 for details. Such an approach could be applied in a computer aided translation tool, which would allow the machine translation system to adapt to translator corrections as produced by post-editing or through an interactive translation prediction interface (Wuebker et al., 2016; Knowles and Koehn, 2016). Single-sentence adaptation allows the model to learn the translator's preferred translations, which may be specific to the particular document. For example, the system might initially produce a valid translation for a word in the document, while the translator prefers an alternate translation; after single-sentence adaptation, the system can learn to produce the translator's preferred translation in future sentences.

Algorithm 1 Single-Sentence Adaptation

M : Baseline Model
D : Set of Documents
for $d \in D$ **do**
  ▷ ref : reference translation of d
  ▷ m_i : model trained through i^{th} sentence
  ▷ d_i : i^{th} line in d
  ▷ ref_i : i^{th} line in ref
  $result \leftarrow \{\}$
  $m_0 \leftarrow M$
  for $i \leftarrow 1, \text{NUMLINES}(d)$ **do**
    $result_i \leftarrow \text{INFER}(m_{i-1}, d_i)$
    $m_i \leftarrow \text{ADAPT}(m_{i-1}, (d_i, ref_i))$
  end for
  $baseHyp \leftarrow \text{INFER}(M, d)$
  $baseScore \leftarrow \text{BLEU}(baseHyp, ref)$
  $adaptScore \leftarrow \text{BLEU}(result, ref)$
end for
▷ We compare $baseScore$ and $adaptScore$

3.2 Dictionary Training

This approach aims to adapt models with the specific goal of better translating novel words. Given a new document to translate, we identify words that are novel (have not appeared in any training or adaptation data). Next, we obtain a single translation for each of these words (in a computer aided translation setting, this might consist of asking a human translator to provide translations; along the lines of terminology curation). In this work, we simulate the collection of such dictionaries (or terminology banks) using the reference. We then treat the list of novel words and their respective translations as bitext and continue model training,

[3] In cases where the domain is fairly homogeneous, it may be beneficial *not* to reset the model between documents, while in heterogeneous domains it may be desirable to do so always. We leave this issue to future work.

producing a model specifically adapted to this document's novel vocabulary, which we can then use to decode the complete document. Note that this is a very small bitext to train on, and each line of the bitext contains a single word (segmented into multiple tokens by byte-pair encoding).

To simulate a translator-produced dictionary, we build a dictionary of novel word translations from the source and reference. First we run fast-align (Dyer et al., 2013) over the byte-pair encoded representations of the source and reference sentences.[4] The target-side token whose subword segments most frequently align to the subword segments of the source-side token is selected as a candidate translation, and a single final translation is selected based on the most common candidate translation within the document.[5]

3.3 Single-Sentence Adaptation with Dictionary Training

Dictionary training and sentence adaptation offer distinct benefits when adapting over a document. Dictionary training helps the model learn the right translations for novel words and single-sentence adaptation can provide a more general adaptation. The latter can also learn correct translations of repeated novel words, but may require multiple instances to do so. Doing dictionary adaptation beforehand could ensure that the novel terminology is correctly and consistently translated from the beginning of the document, which could eliminate a pain point for human translators. In this combined approach, we begin with the document's dictionary trained model and use that as the initial model for single-sentence adaptation.

4 Data and Models

We use two distinct datasets and baseline models to evaluate our approaches, translating from English into German. We evaluate on WMT news data and EMEA medical data using baseline WMT and EMEA domain adapted models, respectively. The different domains (news vs. medical) allow us to evaluate our approaches in different scenarios.

[4]The fast-align model is trained over the byte-pair encoded representations of the full training data: WMT data, backtranslations released by Sennrich et al. (2016b), and EMEA data used for adaptation.

[5]Note that, particularly for words with morphological variants in the target language, there may have been more than one correct translation. We account for this in evaluation, but only train on one translation option.

4.1 WMT

WMT Data: We test on the full WMT 2017 news translation test set, splitting it into 130 unique documents (derived from the document splits in the original SGM file). Each document is a short news story. These stories are drawn from a number of news sources, covering a wide range of topics. While all documents are in the "news" domain, this is a fairly heterogeneous dataset. The documents range in length from 2 to 64 lines, with an average length of 22.1 lines (median 20).

We used the first 20 documents from the 2016 WMT news translation test set as a development set for selecting training parameters for dictionary training experiments, and a subset of 8 of these documents for selecting parameters for the single-sentence training experiments. The development set documents had a similar range of lengths (3 lines to 62 lines, with an average of 19.0).

The number of novel word types per document in our test set ranged from 0 (no novel words; no dictionary adaptation) to 15 novel words. There are 295 novel types (across all documents combined) and 442 novel tokens. Across the test set, 12.8% of lines contain at least one novel word. In some cases, up to 75% of the lines within a single document contain at least one novel word.

WMT Baseline Model: We use a publicly available English-German model.[6] The model is trained using Nematus (Sennrich et al., 2017) on the WMT parallel text, supplemented by synthetic back-translated data as described in Edinburgh's WMT 2016 submission (Sennrich et al., 2016a). They use byte-pair encoding (Sennrich et al., 2016b) to allow for (near) open-vocabulary NMT. The model uses 512 length word-embeddings with an hidden layer size of 1024. As this was trained for the 2016 WMT evaluation, both the 2016 and 2017 test sets can be safely used for development and testing, respectively, as they were not included in training data.

4.2 EMEA

EMEA Data: We use a subset[7] of the European Medicines Agency (EMEA) parallel corpus.[8] It consists of sentence-aligned documents focusing

[6]`data.statmt.org/rsennrich/wmt16_systems`

[7]We select only those documents labeled as "humandocs" and filter out documents that contain only or primarily highly-repetitive dosage information.

[8]`http://opus.lingfil.uu.se/EMEA.php`

on medical products (Tiedemann, 2009). The corpus contains high levels of domain-specific terminology and repetition, making it appropriate for this task. Each document describes a new medication, meaning that new documents contain novel vocabulary. The medication name is typically repeated frequently within the document. Other novel vocabulary items include highly-specific medical terminology; these tend to appear fewer times within the document.

We divide the documents into training, development, and test sets such that all documents about a particular medication are in the same set. Thus most novel medication names in the development and test data will have been unobserved in the training data. We use four splits of the data: 500 document pairs (375K sentence pairs) for training a baseline EMEA-adapted model, 22 document pairs (5K sentence pairs) as validation for that training, 5 document pairs (285 sentence pairs) for a small grid search over parameters, and 47 documents (2,755 sentence pairs) for testing.

Test documents ranged in length from 48 lines to 95 lines. In general, the EMEA documents have a greater variation in length than this (with some having 1000 or more lines). For data with 200 or more lines, considerable BLEU improvements have been documented with online adaptation and continued training. However, we seek to demonstrate that adaptation can be done with even shorter documents, and so focus this test set on documents with fewer than 100 lines.

The number of novel types per document in our test set ranged from 0 (no novel words; no dictionary adaptation) to 10 novel words. There are a total of 151 novel types (all documents combined) and 1,129 novel tokens. Across the test set, 38.3% of lines contain at least one novel word. In some cases, up to 63.5% of the lines within a single document contain at least one novel word. Some novel word types occurred more than 30 times within a single document.

EMEA Baseline Model: The WMT model is trained on data which is significantly different from the EMEA data's medical domain. We see considerable differences including vocabulary and sentence lengths. If we were to use the unadapted WMT model as our baseline, we might expect high gains from very small amounts of data due to the domain differences. Instead, in order to determine what marginal gains are possible in a real-life use scenario where a client already has access to a domain-specific model, we first adapt the WMT model on the EMEA train data so that it is familiar with the general style and vocabulary of the new dataset. Thus, improvements are attributable to document-specific adaptation rather than general domain adaptation.

We use the 375K sentence pair training set, validating on the 5K sentence pair development set, to perform continued training (Freitag and Al-Onaizan, 2016; Luong and Manning, 2015). We use the same subword vocabulary and preprocessing pipeline as the WMT model. We clip sentence lengths to 50 tokens and train with a batch size of 80 over 15 epochs. We use a learning rate of 0.001 with the Adam optimizer (Kingma and Ba, 2014).

While training, external validation is done every 1,000 batches and models are saved accordingly. We choose the model that gives the best validation score over the development set. Results are consistent with prior work: performance on the new domain peaks around the first few epochs and then tails off (Freitag and Al-Onaizan, 2016; Luong and Manning, 2015).

The performance of the baseline WMT model on the EMEA development set gives a BLEU score of 18.2. Our best adapted model gives a BLEU of 51.5. With over 30 points increase in BLEU, the adapted model is well-tuned to the EMEA corpus. We use this adapted model as the baseline for further document-level adaptation.

5 Experiments

The two domains and their respective baseline models provide us two distinct scenarios to evaluate our methodology. Both simulate a relatively data-rich realistic setting in which translators have completed translations of in-domain data and continue to work on new documents (with novel terminology) within that domain. Each domain provides its own challenges: the WMT data covers a wide range of topics and sources of news stories, while the EMEA data includes highly technical medical vocabulary, presented in fairly consistent ways. Due to the way our EMEA data splits were produced, this in particular means that the new EMEA documents will likely contain novel vocabulary (such as names of medications and other specific terminology). Similarly, we expect news stories to cover new names, locations, and more as news breaks over time.

Source	Breast-feeding should be stopped while taking Siklos .
Reference	Das Stillen sollte während der Behandlung mit Siklos eingestellt werden .
Baseline	Während der Einnahme von Xenlos sollte abgestillt werden .
Dict.-Adapt.	Während der Einnahme von Siklos sollte abgestillt werden .
Single-Sent.-Adapt.	Während der Behandlung mit Ivlos sollte abgestillt werden .
Dict.+Single-Sent.-Adapt.	Während der Behandlung mit Siklos sollte abgestillt werden .

Table 2: Complementary nature of two approaches: single-sentence approach learns the preferred translation of "while taking" ("Während der Behandlung"), but mistranslates *Siklos* as *Ivlos*. Dictionary training produces *Siklos* correctly, but makes no other changes. Combined, the overall translation is improved, though it would still require post-editing for correctness.

5.1 Single-Sentence Adaptation Experiments

For hyperparameter optimization, we did a complete grid search over a span of learning rates (0.1, 0.01, 0.001, 0.0001, 0.00001), train epochs (1, 5, 10, 20), and optimizers (*Adam*, *SGD*) on WMT data and a partial search on EMEA data. We use BLEU (Papineni et al. (2002)) to measure the effect of adaptation. We found the optimum configurations (*optim*, *lr*, *epochs*) of (*SGD*, 0.01, 5) for EMEA[9] and (*SGD*, 0.1, 20) for WMT. The difference in optimum configurations can be partly attributed to the different domains of the two datasets. We note that the best EMEA configuration matched the second-best WMT one.

5.2 Dictionary Training Experiments

For the EMEA dictionary experiments, we completed a grid search over number of epochs (1, 2, 5, 10) and learning rate (0.1, 0.5, 1.0) using SGD as the optimizer.[10] Finding consistent results, we ran a smaller grid search (epochs: 2 and 5 and learning rates 0.1, 0.5, and 1.0) over a development set of the first 20 documents from WMT 2016. Setting the learning rate and/or number of epochs too low resulted in minimal changes, while setting them too high resulted in pathological overfitting (loops of repeated tokens, etc.). Based on these initial experiments, we set a learning rate of 0.5 for both data sets, with 5 epochs for EMEA data and 2 epochs for WMT data. The parameters chosen were those that maximized BLEU score on the development sets.

5.3 Lexically Constrained Decoding Experiments

We compare our dictionary training approach against an approach that uses the same dictionaries and enforces a lexical constraint: if one of the dictionary entries appears in the source, its translation (acquired as described in Section 3.2) must appear in the translated output. We do this using the grid beam search approach described in Hokamp and Liu (2017). Rather than adapting the underlying machine translation model, this approach constrains the search space to translations containing specified sub-sequences (in this case, the byte-pair encoded representations of the translation of any words from the dictionary which appears in the source sentence). We use the publicly released implementation for Nematus, with a beam size of 12.

[9]During hyperparameter selection, document lengths were clipped to the first 60 lines.

[10]We also considered lower learning rates (0.01, 0.001, 0.0001), but found that they did not result in much, if any, change to the model.

Model	BLEU	Nov. Acc.
EMEA-Adapt. Baseline	51.1	39.9%
Single-Sent. Adapt.	52.8	62.3%
Lex. Const. Decoding	50.4	86.5%
Dictionary Training	53.3	87.9%
Dict. + Single-Sent.	53.8	89.1%

Table 3: Results of baseline and dictionary training across the full set of EMEA test documents. Accuracy is computed for novel words only.

5.4 Single Sentence Adaptation with Dictionary Training Experiments

Here we combine the approaches: for every document, we first do dictionary training. Using that as the starting point, we perform single sentence adaptation. We use the best hyperparameters obtained from the grid search for the individual methods.

6 Results & Analysis

We evaluate on two metrics. First, we compute BLEU over the full set of test documents and compare against the baseline translations. Across both domains, single-sentence adaptation provides consistent improvements in BLEU score (1.6 BLEU points on WMT data and 1.7 BLEU points on EMEA data). The dictionary training approach

Model	BLEU	Nov. Acc.
WMT Baseline	25.1	48.9%
Single-Sent. Adapt.	26.7	58.4%
Lex. Const. Decoding	25.0	76.9%
Dictionary Training	25.1	71.7%
Dict. + Single-Sent.	26.9	72.2%

Table 4: Results of baseline and dictionary training across the full set of WMT test documents. Accuracy is computed for novel words only.

has more varied results. We see no clear improvement on the WMT data, but training on these small dictionaries does not *hurt* BLEU score overall. However, for the EMEA data, dictionary training produces a 2.2 BLEU point improvement. This gain can be primarily attributed to producing correct translations of the novel vocabulary, which can make a large difference in n-gram matches.[11] The lexically constrained decoding approach results in a decrease in BLEU score on both domains. Combining both dictionary training and single-sentence adaptation results in modest improvements (0.2 on WMT and 0.5 on EMEA) over the best single approach for each domain. Full results are shown in Tables 3 and 4. The combined approach produces improvements over the baseline for 79.2% of the WMT documents and 83.0% of the EMEA documents.

Figure 1 shows difference in BLEU produced by single-sentence adaptation as compared to the baseline on EMEA data. The overall trend is a net improvement in BLEU which shows up as early as 10 sentences from the start.

We also observe qualitative results that suggest that single-sentence adaptation is performing as expected, learning document- or translator-specific translations. For example, the baseline WMT system initially translates the English bigram "delicate operation" as "delikater Betrieb" while the reference translation prefers "heikle Tätigkeit" as the translation. In the next sentence in which "delicate operation" is observed, the sentence-adapted model successfully translates it as "heikle Tätigkeit" instead. Table 2 shows another example in which the two approaches combine to produce improvements.

We also compute accuracy for the translations of novel words. To compute accuracy, we first run

[11]Consider the case of the baseline translation *Was ist AFluntis ?* and the (correct) dictionary-adapted version *Was ist Aflunov ?* – the former contains no 4-gram matches.

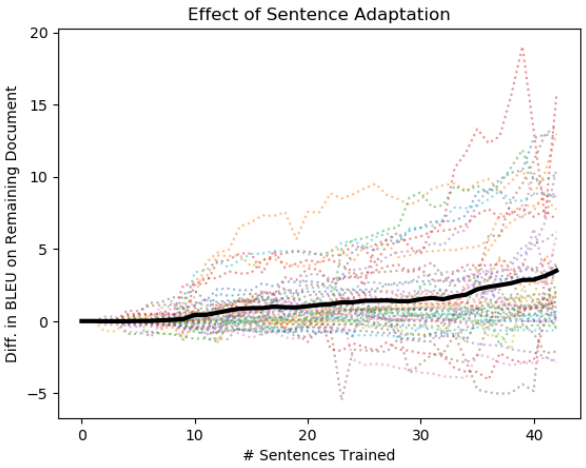

Figure 1: The X-axis shows the number of sentences to which the model has been adapted. The Y-axis shows the difference in BLEU score between this adapted model and the baseline on the document's remaining lines. Dotted lines represent individual documents; the average trend is shown in bold.

a trained fast-align model over the byte-pair encoded source and the byte-pair encoded reference. We use this alignment to map full tokens from the source to full tokens in the reference (as was done for producing the dictionaries). We then align the source sentence and the machine translation output the same way. For each instance of a novel word, we score its aligned machine translated token as correct if it matches the aligned reference token. The dictionary training approach shows, as expected, a major jump in translation accuracy. The single-sentence adaptation approach shows results between the baseline and the dictionary approach. Lexically constrained decoding underperforms dictionary training on EMEA data (in part because it sometimes produces medication names that are concatenated with other subwords, or produces the medication name more times than required), while it outperforms other methods on the WMT data (at a cost to the overall BLEU score, whereas all other methods produce improvements in BLEU). Table 3 shows that EMEA improves from a baseline accuracy of 39.9% to an accuracy of 87.9% after dictionary training, and Table 4 shows a slightly smaller jump from 48.9% to 71.7% for WMT. Both show slight improvements after combining single-sentence adaptation and dictionary training.

With this increase in accuracy comes an increase in consistency of translating the novel

Model	WMT		EMEA	
	Copy	*Trans.*	*Copy*	*Trans.*
Baseline	80.8%	11.3%	41.9%	28.4%
S.-Sent.	87.9%	23.6%	67.2%	32.7%
Dict.	92.5%	47.3%	92.5%	60.5%
Dict. + S.	94.6%	45.8%	92.9%	66.7%

Table 5: Novel word accuracy divided into tokens to be copied (*Copy*) vs. translated (*Trans.*).

words. In the baseline EMEA-adapted model, the average type-token ratio[12] for translations of novel words that occur at least 3 times (in the source text) is 0.29. With dictionary adaptation, this drops to 0.14 – lower than the reference type-token ration of 0.16 – meaning that the new model produces the exact translation from the dictionary even when a variant (e.g. different case ending) may be appropriate. As we use only one translation per novel source token in the dictionaries used for training, the model overfits slightly. This issue could potentially be alleviated by training on multiple translation options, at the risk of introducing errors from incorrect alignments.

We perform more detailed analysis across two kinds of novel words: those which should simply be copied from source to target (e.g. medication names) and those which must be translated. Table 5 shows results for the baseline and our approaches. WMT data is almost evenly split between these: 46.8% of novel types (54.1% of tokens) must by copied, while EMEA data is skewed towards words that should be copied, with 51.7% of novel types (85.7% of tokens). On WMT data, baseline accuracy of terms to be copied is already quite high, but accuracy of terms to be translated is very low. The EMEA baseline has a much harder time with tokens that should be copied, but does better on non-copied terms. We hypothesize that this may have to do with differences in the morphological attributes of the novel tokens in the different datasets (WMT contains many names of people or places, while EMEA contains many drug names, which tend to contain character sequences not frequent in either source or target language) or with the contexts in which they appear. We observe that for many of the medication names, it takes 10 or more instances of the name being observed for the single-sentence adaptation approach alone to successfully learn to copy the word (if ever). Though there remains a gap between novel word accuracy on tokens that should be copied and those that should be translated, our approaches demonstrate improvements for *both* types of novel words.

A concern with training on a dictionary as bitext is that the model may overfit to the sentence length; we do not find that to be the case here, as the difference between the full hypothesis lengths is 48,641 tokens for the EMEA-adapted data compared to 48,627 for the dictionary-trained models. However, this is dependent on choosing the correct learning rate and number of epochs. Similarly, there's a potential concern that single-sentence training on the previous sentence may cause some type of overfitting (memorization of the sentence, etc.). We do not observe that to be the case either.

7 Conclusions and Future Work

We propose two approaches to document-level adaptation of NMT systems (single-sentence adaptation, dictionary training) and their combination, which can be effectively used to improve performance, both in terms of BLEU score and in the translation of novel words. Both approaches have minimal training data requirements, can be effective applied with an existing NMT architecture, and show considerable improvements even for short documents.

One area meriting further study is dynamic adaptation of hyper-parameters based on document length or content. During our development and test-runs, we found correlations between hyper-parameter configurations and document lengths with some learning rates and train epochs working better for shorter documents while some working better for longer ones. We could foresee dynamically adapting the hyperparameters based on the overlap between the current sentence being translated and the remainder of the document as a possible area of future study. Additionally, it would be useful to explore these methods in a user-study, to better determine the trade-off between improvement and user input required (such as for dictionary creation).

Acknowledgments

Rebecca Knowles was partially supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1232825.

[12]The number of different machine translation outputs for the source type, divided by the number of times that source type appears.

References

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

Marco Baroni and Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marine Carpuat, Hal Daumé III, Alexander Fraser, Chris Quirk, Fabienne Braune, Ann Clifton, Ann Irvine, Jagadeesh Jagarlamudi, John Morgan, Majid Razmara, Aleš Tamchyna, Katharine Henry, and Rachel Rudinger. 2012. Domain adaptation in machine translation: Final report. In *2012 Johns Hopkins Summer Workshop Final Report*.

Mauro Cettolo, Nicola Bertoldi, Marcello Federico, Holger Schwenk, Loïc Barrault, and Christophe Servan. 2014. Translation project adaptation for mt-enhanced computer assisted translation. *Machine Translation*, 28(2):127–150.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 233–237, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the Second Workshop on Neural Machine Translation and Generation*, Melbourne. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

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

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. One sentence one model for neural machine translation. *CoRR*, abs/1609.06490.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Toan Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proc. NAACL HLT*. To appear.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. *CoRR*, abs/1803.00047.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Álvaro Peris and Francisco Casacuberta. 2018. Online learning for effort reduction in interactive neural machine translation. *CoRR*, abs/1802.03594.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation (WMT)*.

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

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Joern Wuebker, Spence Green, John DeNero, Sasa Hasan, and Minh-Thang Luong. 2016. Models and inference for prefix-constrained machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Berlin, Germany. Association for Computational Linguistics.