HABLex: Human Annotated Bilingual Lexicons for Experiments in Machine Translation

Brian Thompson* **Rebecca Knowles*** **Xuan Zhang*** **Huda Khayrallah**
Kevin Duh **Philipp Koehn**
Johns Hopkins University
{brian.thompson, rknowles, xuanzhang, huda, phi}@jhu.edu
kevinduh@cs.jhu.edu

Abstract

Bilingual lexicons are valuable resources used by professional human translators. While these resources can be easily incorporated in statistical machine translation, it is unclear how to best do so in the neural framework. In this work, we present the HABLex dataset, designed to test methods for bilingual lexicon integration into neural machine translation. Our data consists of human-generated alignments of words and phrases in machine translation test sets in three language pairs (Russian→English, Chinese→English, and Korean→English), resulting in clean bilingual lexicons which are well matched to the reference. We also present two simple baselines—constrained decoding and continued training—and an improvement to continued training to address overfitting.

1 Introduction

Neural machine translation (NMT) is the current state-of-the-art. In contrast with statistical machine translation (SMT; Koehn et al., 2007), where there are several established methods of incorporating external knowledge,[1] recent work is still examining how best to incorporate bilingual lexicons into NMT systems. Bilingual lexicon integration is desirable in a number of scenarios: highly technical vocabulary (which might be rare, or require translations of a domain-specific sense), lower-resource settings (where bilingual lexicons might be a significant portion of the available parallel data), translation settings where a client requires particular terms to be used (e.g. brand names), or for improving rare word translation.

At present, there is no standard dataset for benchmarking bilingual lexicon integration, making it difficult to compare methods.

Source	на передней стенке корпуса установлены **амортизаторы** .
Lexical Entry	**амортизаторы↔shock absorbers**
Target	**shock absorbers** are mounted on a front wall of the housing .

Table 1: Example Russian source sentence, lexicon entry, and English target sentence.

We create and release[2] **H**uman **A**nnotated **B**ilingual **Lex**icons (HABLex). Our bilingual lexicons are (1) generated by bilingual experts to ensure high quality, (2) derived from the development and test references so the best-case-scenario impact on translation performance can be directly measured, (3) covering 3 language pairs, and (4) focused on challenging words.

We perform exploratory work on our development set, showing two representative baselines to compare incorporating the lexicon at training time (continued training) vs. at decoding time (constrained decoding). We examine the tradeoffs in terms of BLEU, recall, and speed.

We also present a novel application of Elastic Weight Consolidation (EWC; Kirkpatrick et al., 2017; Thompson et al., 2019) which significantly improves performance by preventing overfitting during continued training on the bilingual lexicon.

2 Related Work

We first review prior work on incorporation of bilingual lexicons into NMT and then discuss the datasets used and explain how our new dataset addresses some shortcomings.

Recent work on the incorporation of bilingual lexicons into NMT systems can be loosely clus-

*These authors contributed equally to this work.

[1] E.g. requiring a particular translation via XML markup: statmt.org/moses/?n=advanced.hybrid

[2] http://www.cs.jhu.edu/~kevinduh/a/hablex2019/

tered into two categories: incorporation at training time or during decoding. These two general approaches have different performance characteristics: incorporation at training time may slow down training, but tends not to alter inference speed, while incorporation at inference time tends to significantly slow down decoding (see subsection 4.3), but without slowing down training.

2.1 Incorporation at Training Time

Zhang and Zong (2016) and Fadaee et al. (2017) both propose using bilingual lexicons to create synthetic bitext to augment training data for NMT systems. Arthur et al. (2016) use translation probabilities from a lexicon (like SMT phrase tables) in conjunction with NMT probabilities.

2.2 Incorporation at Decode Time

Kothur et al. (2018) perform fine-grained continued training adaptation on very small, document-specific bilingual lexicons of novel words.[3]

A popular inference-time approach is constrained decoding (Anderson et al., 2016; Hokamp and Liu, 2017; Chatterjee et al., 2017; Hasler et al., 2018; Post and Vilar, 2018), which modifies beam search to require that user-specified words or phrases to be present in the output hypotheses. Constrained decoding can be used to ensure that target entries from a bilingual lexicon be present in the MT output whenever their corresponding source entries are present in the input. Constrained decoding with multiple target options (e.g. when a source word can be translated into one of several target words) is addressed in Chatterjee et al. (2017) and Hasler et al. (2018).

2.3 Datasets Used in Prior Studies

Prior work has used either human-generated general purpose bilingual lexicons not tailored to a test set (Arthur et al., 2016; Zhang and Zong, 2016) or automatic alignments (Fadaee et al., 2017; Hokamp and Liu, 2017; Chatterjee et al., 2017; Hasler et al., 2018; Kothur et al., 2018) which likely contain errors (especially on rare words). There exist manually word-aligned parallel corpora for Japanese-English Wikipedia data (Neubig, 2011) and for Chinese-English mixed-domain data (Liu and Sun, 2015), from which it would be possible to extract lexicons.

[3] Although this method involves adapting the model, it is done prior to translating each document, so we view it as a decoding-time modification.

Most prior work experimented in a single language pair,[4] making it difficult to predict if a method will generalize. Our work addresses these issues by providing a multi-language testbed for comparisons, focusing on a consistent and lexically-challenging domain.

3 HABLex Dataset

Our motivation is to allow straightforward evaluation of lexicon incorporation methods. By building a reference-derived bilingual lexicon, we ensure that if a method successfully learns correct translations from the lexicon while maintaining overall system performance, BLEU will increase.

We annotate the existing parallel text with alignments for certain vocabulary items, allowing us to extract bilingual lexicons that are specific to the context in which the translation should appear. This produces a cleaner and more well-tailored bilingual lexicon than would be found in most real-world scenarios. However, it could also be used as a standardized starting point to produce a noisier lexicon that more closely mimics real-world lexicons (e.g. by adding irrelevant entries or relevant morphological variants, lemmatizing entries, or subsampling).

At a high level, our lexicons are created by a two-step process: (1) identifying interesting words on the source side of the test and development sets, and (2) human annotators correcting or validating automatic alignments of the identified words.

3.1 Patent Domain & Languages

We chose the patent domain because it contains interesting technical terminology and because of the availability of the high-quality, multilingual World Intellectual Property Organization (WIPO) COPPA-V2 corpus (Junczys-Dowmunt et al., 2016). We build bilingual lexicons for three language pairs: Russian→English (Ru→En), Korean→English (Ko→En), and Chinese→English (Zh→En). We also release the development and test splits from which the annotations were produced.

3.2 Interesting Word Selection

We select interesting words (from the source side of the data) as follows: words that appear less than 5 times in WIPO data, words that appear less than

[4] Except Hokamp and Liu (2017) and Hasler et al. (2018).

	Development		Test	
	Entries	Sentences	Entries	Sentences
Ru	9040	2412	8001	2142
Ko	5593	1744	5595	1756
Zh	1773	885	2289	1025

Table 2: Number of bilingual lexicon entries and sentences containing at least one annotation for each language pair. Most source entries (between 87% and 96%) have a unique translation; the remainder have 7 or fewer target translations.

5 times in general domain training data (see section 4), and words that appear less than 5 times in both.[5] These represent three types of words that may be challenging to translate: words that have only recently been learned, words that may have been forgotten during continued training, and generally rare words.

3.3 Human Annotations

Our bilingual annotators used the LDC Word Aligner tool (Grimes, 2010) to annotate words of interest in context. Table 1 shows example lexicon entries resulting from this annotation process. In order to save annotator time and effort, we displayed highlighted automatic alignments[6] of each of the words of interest in their parallel sentence context (similar to Grimes et al. (2012)). The annotator confirms or corrects the automatic alignment, adding or removing source or target side words as needed to complete a valid alignment. Since annotation is cumbersome with very long sentences, we omit sentences of length 100 or more tokens. Filtering out numerical entries and phrases longer than 3 words[7] results in bilingual lexicons with sizes shown in Table 2. The data contains a small number of discontiguous alignments; these are so infrequent as to have a negligible effect on BLEU score results.

[5] We focus on 5-count (and rarer) words because we expect them to be particularly challenging for MT, but given time and resources there is nothing to prevent the application of the annotation protocol to other terms.

[6] For strong initial alignments, GIZA++ (Och and Ney, 2003) is trained on train, development, and test for all data described in section 4 as well as a TED talk corpus (Cettolo et al., 2012).

[7] While we want the dictionary to match the reference, we did not want to train on large phrases from the reference.

4 Baseline Experiments and Results

To build strong baseline systems, we first train models on general domain data. As general domain data, we use the OpenSubtitles18 (Lison et al., 2018) corpora for both Ru→En and Ko→En. For Ru→En and Zh→En we also use the parallel portion of the WMT17 news translation task (Bojar et al., 2017). We then fine-tune these general-domain models on WIPO training data (Luong and Manning, 2015), using the dev set for validation. These domain-adapted models are then used as the initial systems for our lexicon incorporation experiments.

We build the systems in Sockeye (Hieber et al., 2017), using a two-layer LSTM network with hidden unit size 512. We use an initial learning rate of 3e-4 both for training the general domain model and adapting to WIPO. We apply the Moses tokenizer (Koehn et al., 2007), lowercasing, and byte-pair encoding (BPE; Sennrich et al., 2016) with a vocabulary size of 30k. BPE is trained on the general domain corpus only, then applied to all data.

4.1 Evaluation Metrics

We evaluate lexicon incorporation approaches using two main metrics: BLEU and recall. For each annotated instance of a source-side lexical entry, we can check whether the system output contains the correct aligned target-side translation. Recall is computed as the percentage of the time that the system produces the correct output, averaged over all annotations. Note that this does not guarantee that the words are placed in a sensible location in the sentence, only that they appear.

We also consider training and decoding speed. All of these factors help determine which approaches are best given a particular use case. For example, if a user requires exact fidelity to a lexicon (e.g. branding), they may care the most about recall (while still ensuring that overall translation quality is still acceptable).

4.2 Continued Training on HABLex

We perform continued training (CT; Luong and Manning, 2015)—typically used for domain adaptation—on the bilingual lexicon data. With bilingual lexicons approximately two orders of magnitude larger than those used successfully in Kothur et al. (2018), we find that performance drops dramatically with standard CT. To address this problem, we apply EWC (Kirkpatrick et al.,

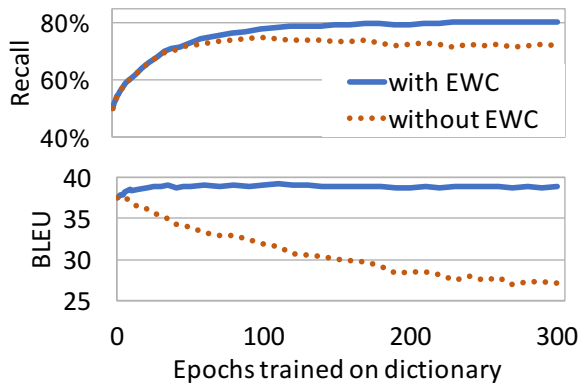

Figure 1: Ru→En recall and BLEU during continued training on bilingual lexicon for the development set.

			CT		CD	
		BL	CT	+EWC	Rand.	Oracle
Ru	bleu	37.5	27.1	38.9	40.0	40.5
	rec.	48.2	72.1	80.2	95.9	99.9
Ko	bleu	34.1	10.9	31.5	31.6	33.7
	rec.	55.0	47.3	69.7	87.6	99.0
Zh	bleu	39.9	34.8	39.0	42.2	42.5
	rec.	38.9	78.7	81.2	97.4	99.9

Table 3: BLEU and recall % (rec.) of baseline (BL), continued training (CT) with and without EWC, and oracle and random constrained decoding (CD) on development set.[8]

2017), a method for training a neural network to learn a new task without catastrophically forgetting how to perform a previously learned task. EWC has recently been shown to significantly reduce general domain performance loss during domain adaptation in NMT (Thompson et al., 2019); we apply it to retain the ability to translate full sentences while training on a bilingual lexicon.

We experimented with initial learning rates (0.001, 0.00316, 0.01, 0.0316, 0.1), all using SGD for 300 epochs, and EWC weight decay values (1e-1, 1e-2, 1e-3, 1e-4, 1e-5). We chose a learning rate 0.1 because it allowed recall to (at least approximately) converge, and a weight decay value of 1e-4 because it performed reasonably well on the dev sets for all 3 languages. All results reported are for the final checkpoint.

4.3 Constrained Decoding

We employ *dynamic beam allocation* (DBA; Post and Vilar, 2018) for constrained decoding (CD). At each time step of decoding, DBA groups hypotheses into *banks* based on the number of translated constraints, and a fixed-size beam is dynamically allocated across the banks. While DBA has a time complexity constant in the number of constraints, in practice we find it to be approximately an order of magnitude slower than regular decoding, primarily because Sockeye does not currently support batched DBA.[9]

One limitation of DBA is that it only works on constraints that have only one translation. In this work, we report **oracle** choice (use the right sense for the specific test sentence) as an upper bound on performance and **random** choice (choose a random possible translation) as another baseline. Our dictionary may contain more than one possible translation for a given target word. For example, we observe the following three English translations for the Russian word арматурного: *rebar, reinforced,* and *reinforcement*. These translations are appropriate in different contexts and are not interchangeable, so the oracle always selects the translation that is appropriate for the given sentence (according to its reference translation). The random approach selects one of the translations uniformly at random; if there is only one translation, random is identical to oracle.

	bleu		recall (%)	
	Baseline	Random	Baseline	Random
Ru	37.5	40.3	46.7	96.5
Ko	34.5	32.7	54.7	89.3
Zh	39.0	40.9	36.3	96.2

Table 4: Baseline and constrained decoding (random) on test set.

4.4 Results & Discussion

Table 3 summarizes key exploratory results for the baseline, CT, and CD approaches on the development set. Table 4 shows baseline and random CD benchmarks on the test set, which we otherwise reserve and release for future evaluation. We confine the remainder of our discussion to the experiments we performed on the development set.

EWC noticeably improves BLEU performance as compared to standard CT, while also increasing recall. (See Figure 1 for example performance as the model is trained on the bilingual lexicon and Table 3 for results on all three language pairs.) CD outperforms CT in terms of both BLEU score and

[8]The recall of the oracle method would be 100%, if there were no out-of-vocabulary subwords in the bilingual lexicon.

[9]Constrained decoding with DBA takes 3.1 seconds per sentence on average on a Tesla K80 GPU.

recall, at the expense of decoding speed. Random CD nears oracle CD performance because most source-side entries have only one translation. We would not necessarily expect such high random CD performance if the bilingual lexicon averaged more senses per source entry.

The use case for bilingual lexicon incorporation should also influence user decisions about what approaches to use. If exact translations (e.g. brand names or highly technical translations) are of the utmost importance, a CD approach might be preferred even if it is slower. In the case that a general lexicon has been provided but there is more flexibility in terminology, it may be better to perform CT, or perhaps even a combination of the two.

5 Conclusions

Bilingual lexicons are important resources in translation, but it is not clear how to best incorporate them in NMT. To address this challenge, we present the HABLex dataset, a multi-language, reference-derived development and test set that facilitates the evaluation of lexicon incorporation. We compare two baselines, based on incorporation at training time or at decode time, in terms of BLEU, recall, and speed. We also present a novel application of EWC to continued training which addresses lexicon overfitting.

Acknowledgments

We thank the annotators who worked on building this dataset, as well as the reviewers for their helpful comments. Brian Thompson is supported through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

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.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168. 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.

Stephen Grimes. 2010. LDC word aligner, version 1.20.

Stephen Grimes, Katherine Peterson, and Xuansong Li. 2012. Automatic word alignment tools to scale production of manually aligned parallel texts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*.

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

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

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.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe Mazenc. 2016. Coppa v2. 0: Corpus of parallel patent applications building large parallel corpora with gnu make. In *4th Workshop on Challenges in the Management of Large Corpora Workshop Programme*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic. Association for Computational Linguistics.

Sachith Sri Ram Kothur, Rebecca Knowles, and Philipp Koehn. 2018. Document-level adaptation for neural machine translation. In *Proceedings of the Second Workshop on Neural Machine Translation and Generation*, Melbourne. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2295–2301. AAAI Press.

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*.

Graham Neubig. 2011. The Kyoto free translation task.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Volume 2 (Short Papers)*. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. *CoRR*, abs/1610.07272.