Domain Adaptation of Neural Machine Translation by Lexicon Induction

Junjie Hu, Mengzhou Xia, Graham Neubig, Jaime Carbonell
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
{junjieh,gneubig,jgc}@cs.cmu.edu, mengzhox@andrew.cmu.edu

Abstract

It has been previously noted that neural machine translation (NMT) is very sensitive to domain shift. In this paper, we argue that this is a dual effect of the highly lexicalized nature of NMT, resulting in failure for sentences with large numbers of unknown words, and lack of supervision for domain-specific words. To remedy this problem, we propose an unsupervised adaptation method which fine-tunes a pre-trained out-of-domain NMT model using a pseudo-in-domain corpus. Specifically, we perform lexicon induction to extract an in-domain lexicon, and construct a pseudo-parallel in-domain corpus by performing word-for-word back-translation of monolingual in-domain target sentences. In five domains over twenty pairwise adaptation settings and two model architectures, our method achieves consistent improvements without using any in-domain parallel sentences, improving up to 14 BLEU over unadapted models, and up to 2 BLEU over strong back-translation baselines.

1 Introduction

Neural machine translation (NMT) has demonstrated impressive performance when trained on large-scale corpora (Bojar et al., 2018). However, it has also been noted that NMT models trained on corpora in a particular domain tend to perform poorly when translating sentences in a significantly different domain (Chu and Wang, 2018; Koehn and Knowles, 2017). Previous work in the context of phrase-based statistical machine translation (Daumé III and Jagarlamudi, 2011) has noted that unseen (OOV) words account for a large portion of translation errors when switching to new domains. However this problem of OOV words in cross-domain transfer is under-examined

in the context of NMT, where both training methods and experimental results will differ greatly. In this paper, we try to fill this gap, examining domain adaptation methods for NMT specifically focusing on correctly translating unknown words.

As noted by Chu and Wang (2018), there are two important distinctions to make in adaptation methods for MT. The first is data requirements; *supervised* adaptation relies on in-domain parallel data, and *unsupervised* adaptation has no such requirement. There is also a distinction between *model-based* and *data-based* methods. Model-based methods make explicit changes to the model architecture such as jointly learning domain discrimination and translation (Britz et al., 2017), interpolation of language modeling and translation (Gulcehre et al., 2015; Domhan and Hieber, 2017), and domain control by adding tags and word features (Kobus et al., 2017). On the other hand, data-based methods perform adaptation either by combining in-domain and out-of-domain parallel corpora for supervised adaptation (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016) or by generating pseudo-parallel corpora from in-domain monolingual data for unsupervised adaptation (Sennrich et al., 2016a; Currey et al., 2017).

Specifically, in this paper we tackle the task of *data-based, unsupervised* adaptation, where representative methods include creation of a pseudo-parallel corpus by back-translation of in-domain monolingual target sentences (Sennrich et al., 2016a), or construction of a pseudo-parallel in-domain corpus by copying monolingual target sentences to the source side (Currey et al., 2017). However, while these methods have potential to strengthen the target-language decoder through addition of in-domain target data, they do not explicitly provide direct supervision of domain-specific words, which we argue is one of the major difficulties caused by domain shift.

To remedy this problem, we propose a new

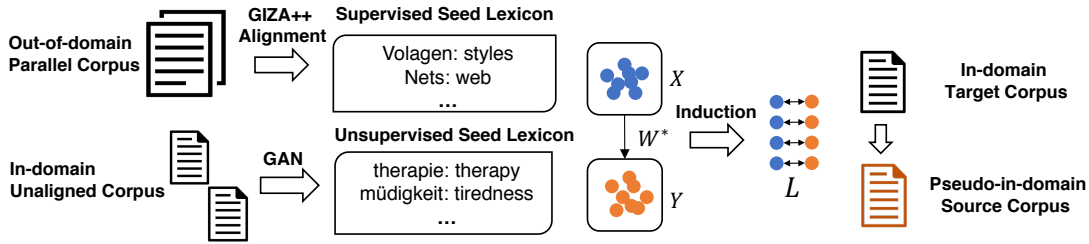Code/scripts are released at https://github.com/junjiehu/dali.

Figure 1: Work flow of domain adaptation by lexicon induction (DALI).

data-based method for unsupervised adaptation that specifically focuses the unknown word problem: **domain adaptation by lexicon induction (DALI)**. Our proposed method leverages large amounts of monolingual data to find translations of in-domain unseen words, and constructs a pseudo-parallel in-domain corpus via word-for-word back-translation of monolingual in-domain target sentences into source sentences. More specifically, we leverage existing supervised (Xing et al., 2015) and unsupervised (Conneau et al., 2018) lexicon induction methods that project source word embeddings to the target embedding space, and find translations of unseen words by their nearest neighbors. For supervised lexicon induction, we learn such a mapping function under the supervision of a seed lexicon extracted from out-of-domain parallel sentences using word alignment. For unsupervised lexicon induction, we follow Conneau et al. (2018) to infer a lexicon by adversarial training and iterative refinement.

In the experiments on German-to-English translation across five domains (Medical, IT, Law, Subtitles, and Koran), we find that DALI improves both RNN-based (Bahdanau et al., 2015) and Transformer-based (Vaswani et al., 2017) models trained on an out-of-domain corpus with gains as high as 14 BLEU. When the proposed method is combined with back-translation, we can further improve performance by up to 4 BLEU. Further analysis shows that the areas in which gains are observed are largely orthogonal to back-translation; our method is effective in translating in-domain unseen words, while back-translation mainly improves the fluency of source sentences, which helps the training of the NMT decoder.

2 Domain Adaptation by Lexicon Induction

Our method works in two steps: (1) we use lexicon induction methods to learn an in-domain lexicon

from in-domain monolingual source data $D_{\text{src-in}}$ and target data $D_{\text{tgt-in}}$ as well as out-of-domain parallel data $D_{\text{parallel-out}}$, (2) we use this lexicon to create a pseudo-parallel corpus for MT.

2.1 Lexicon Induction

Given separate source and target word embeddings, $\mathbf{X}$, $\mathbf{Y} \in \mathbb{R}^{d \times N}$, trained on all available monolingual source and target sentences across all domains, we leverage existing lexicon induction methods that perform *supervised* (Xing et al., 2015) or *unsupervised* (Conneau et al., 2018) learning of a mapping $f(\mathbf{X}) = \mathbf{W}\mathbf{X}$ that transforms source embeddings to the target space, then selects nearest neighbors in embedding space to extract translation lexicons.

Supervised Embedding Mapping Supervised learning of the mapping function requires a seed lexicon of size n, denoted as $L = \{(s,t)_i\}_{i=1}^n$. We represent the source and target word embeddings of the i-th translation pair $(s,t)_i$ by the i-th column vectors of $\mathbf{X}^{(n)}, \mathbf{Y}^{(n)} \in \mathbb{R}^{d \times n}$ respectively. Xing et al. (2015) show that by enforcing an orthogonality constraint on $\mathbf{W} \in O_d(\mathbb{R})$, we can obtain a closed-form solution from a singular value decomposition (SVD) of $\mathbf{Y}^{(n)}\mathbf{X}^{(n)T}$:

$$\mathbf{W}^* = \arg \max_{\mathbf{W} \in O_d(\mathbb{R})} \|\mathbf{Y}^{(n)} - \mathbf{W}\mathbf{X}^{(n)}\|_F = \mathbf{U}\mathbf{V}^T$$

$$\mathbf{U}\Sigma\mathbf{V}^T = \text{SVD}(\mathbf{Y}^{(n)}\mathbf{X}^{(n)T}). \tag{1}$$

In a domain adaptation setting we have parallel out-of-domain data $D_{\text{parallel-out}}$, which can be used to extract a seed lexicon. Algorithm 1 shows the procedure of extracting this lexicon. We use the word alignment toolkit GIZA++ (Och and Ney, 2003) to extract word translation probabilities $P(t|s)$ and $P(s|t)$ in both forward and backward directions from $D_{\text{parallel-out}}$, and extract lexicons $L_{\text{fw}} = \{(s,t), \forall P(t|s) > 0\}$ and $L_{\text{bw}} =$

Algorithm 1 Supervised lexicon extraction

Input: Parallel out-of-domain data $D_{\text{parallel-out}}$
Output: Seed lexicon $L = \{(s,t)\}_{i=1}^n$

1: Run GIZA++ on $D_{\text{parallel-out}}$ to get $L_{\text{fw}}, L_{\text{bw}}$
2: $L_g = L_{\text{fw}} \cup L_{\text{bw}}$
3: Remove pairs with punctuation only in either s and t from L_g
4: Initialize a counter $C[(s,t)] = 0 \ \forall (s,t) \in L_g$
5: **for** (src, tgt) $\in D_{\text{parallel-out}}$ **do**
6:    **for** $(s,t) \in L_g$ **do**
7:       **if** $s \in$ src **and** $t \in$ tgt **then**
8:          $C[(s,t)] = C[(s,t)] + 1$
9: Sort C by its values in the descending order
10: $L = \{\}, S = \{\}, T = \{\}$
11: **for** $(s,t) \in C$ **do**
12:    **if** $s \notin S$ **and** $t \notin T$ **then**
13:       $L = L \cup \{(s,t)\}$
14:       $S = S \cup \{s\}, \ T = T \cup \{t\}$
15: **return** L

$\{(s,t), \ \forall P(s|t) > 0\}$. We take the union of the lexicons in both directions and further prune out translation pairs containing punctuation that is non-identical. To avoid multiple translations of either a source or target word, we find the most common translation pairs in $D_{\text{parallel-out}}$, sorting translation pairs by the number of times they occur in $D_{\text{parallel-out}}$ in descending order, and keeping those pairs with highest frequency in $D_{\text{parallel-out}}$.

Unsupervised Embedding Mapping For unsupervised training, we follow Conneau et al. (2018) in mapping source word embeddings to the target word embedding space through adversarial training. Details can be found in the reference, but briefly a discriminator is trained to distinguish between an embedding sampled from $\mathbf{WX}$ and $\mathbf{Y}$, and $\mathbf{W}$ is trained to prevent the discriminator from identifying the origin of an embedding by making $\mathbf{WX}$ and $\mathbf{Y}$ as close as possible.

Induction Once we obtain the matrix $\mathbf{W}$ either from supervised or unsupervised training, we map all the possible in-domain source words to the target embedding space. We compute the nearest neighbors of an embedding by a distance metric, Cross-Domain Similarity Local Scaling (CSLS; Conneau et al. (2018)):

$$\text{CSLS}(\mathbf{Wx}, \mathbf{y}) = 2\cos(\mathbf{Wx}, \mathbf{y}) - r_T(\mathbf{Wx}) - r_S(\mathbf{y})$$

$$r_T(\mathbf{Wx}) = \frac{1}{K} \sum_{\mathbf{y}' \in \mathcal{N}_T(\mathbf{Wx})} \cos(\mathbf{Wx}, \mathbf{y}')$$

where $r_T(\mathbf{Wx})$ and $r_S(\mathbf{y})$ measure the average cosine similarity between their K nearest neighbors in the source and target spaces respectively.

To ensure the quality of the extracted lexicons, we only consider mutual nearest neighbors, i.e., pairs of words that are mutually nearest neighbors of each other according to CSLS. This significantly decreases the size of the extracted lexicon, but improves the reliability.

2.2 NMT Data Generation and Training

Finally, we use this lexicon to create pseudo-parallel in-domain data to train NMT models. Specifically, we follow Sennrich et al. (2016a) in back-translating the in-domain monolingual target sentences to the source language, but instead of using a pre-trained target-to-source NMT system, we simply perform word-for-word translation using the induced lexicon L. Each target word in the target side of L can be deterministically back-translated to a source word, since we take the nearest neighbor of a target word as its translation according to CSLS. If a target word is not mutually nearest to any source word, we cannot find a translation in L and we simply copy this target word to the source side. We find that more than 80% of the words can be translated by the induced lexicons. We denote the constructed pseudo-parallel in-domain corpus as $D_{\text{pseudo-parallel-in}}$.

During training, we first pre-train an NMT system on an out-of-domain parallel corpus $D_{\text{parallel-out}}$, and then fine tune the NMT model on a constructed parallel corpus. More specifically, to avoid overfitting to the extracted lexicons, we sample an equal number of sentences from $D_{\text{parallel-out}}$, and get a fixed subset $D'_{\text{parallel-out}}$, where $|D'_{\text{parallel-out}}| = |D_{\text{pseudo-parallel-in}}|$. We concatenate $D'_{\text{parallel-out}}$ with $D_{\text{pseudo-parallel-in}}$, and fine-tune the NMT model on the combined corpus.

3 Experimental Results

3.1 Data

We follow the same setup and train/dev/test splits of Koehn and Knowles (2017), using a German-to-English parallel corpus that covers five different domains. Data statistics are shown in Table 2.

Domain	Method		Medical	IT	Subtitles	Law	Koran	Avg.	Gain
Medical	LSTM	Unadapted	46.19	4.62	2.54	7.05	1.25	3.87	+4.31
		DALI	-	11.32	7.79	9.72	3.85	**8.17**	
	XFMR	Unadapted	49.66	4.54	2.39	7.77	0.93	3.91	+4.79
		DALI	-	10.99	8.25	11.32	4.22	**8.70**	
IT	LSTM	Unadapted	7.43	57.79	5.49	4.10	2.52	4.89	+5.98
		DALI	20.44	-	9.53	8.63	4.85	**10.86**	
	XFMR	Unadapted	6.96	60.43	6.42	4.50	2.45	5.08	+5.76
		DALI	19.49	-	10.49	8.75	4.62	**10.84**	
Subtitles	LSTM	Unadapted	11.36	12.27	27.29	10.95	10.57	11.29	+2.79
		DALI	21.63	12.99	-	11.50	10.17	**16.57**	
	XFMR	Unadapted	16.51	14.46	30.71	11.55	12.96	13.87	+3.85
		DALI	26.17	17.56	-	13.96	13.18	**17.72**	
Law	LSTM	Unadapted	15.91	6.28	4.52	40.52	2.37	7.27	+4.85
		DALI	24.57	10.07	9.11	-	4.72	**12.12**	
	XFMR	Unadapted	16.35	5.52	4.57	46.59	1.82	7.07	+6.17
		DALI	26.98	11.65	9.14	-	5.15	**13.23**	
Koran	LSTM	Unadapted	0.63	0.45	2.47	0.67	19.40	1.06	+6.56
		DALI	12.90	5.25	7.49	4.80	-	**7.61**	
	XFMR	Unadapted	0.00	0.44	2.58	0.29	15.53	0.83	+7.54
		DALI	14.27	5.24	9.01	4.94	-	**8.37**	

Table 1: BLEU scores of LSTM based and Transformer (XFMR) based NMT models when trained on one domain (columns), and tested on another domain (rows). The last two columns show the average performance of unadapted baselines and DALI, and the average gains.

Note that these domains are very distant from each other. Following Koehn and Knowles (2017), we process all the data with byte-pair encoding (Sennrich et al., 2016b) to construct a vocabulary of 50K subwords. To build an unaligned monolingual corpus for each domain, we randomly shuffle the parallel corpus and split the corpus into two parts with equal numbers of parallel sentences. We use the target and source sentences of the first and second halves respectively. We combine all the unaligned monolingual source and target sentences on all five domains to train a skip-gram model using *fasttext* (Bojanowski et al., 2017). We obtain source and target word embeddings in 512 dimensions by running 10 epochs with a context window of 10, and 10 negative samples.

Corpus	Words	Sentences	W/S
Medical	12,867,326	1,094,667	11.76
IT	2,777,136	333,745	8.32
Subtitles	106,919,386	13,869,396	7.71
Law	15,417,835	707,630	21.80
Koran	9,598,717	478,721	20.05

Table 2: Corpus statistics over five domains.

3.2 Main Results

We first compare DALI with other adaptation strategies on both RNN-based and Transformer-based NMT models.

Table 1 shows the performance of the two models when trained on one domain (columns) and tested on another domain (rows). We fine-tune the unadapted baselines using pseudo-parallel data created by DALI. We use the unsupervised lexicon here for all settings, and leave a comparison across lexicon creation methods to Table 3. Based on the last two columns in Table 1, DALI substantially improves both NMT models with average gains of 2.79-7.54 BLEU over the unadapted baselines.

We further compare DALI with two popular data-based unsupervised adaptation methods that leverage in-domain monolingual target sentences: (1) a method that copies target sentences to the source side (Copy; Currey et al. (2017)) and (2) back-translation (BT; Sennrich et al. (2016a)), which translates target sentences to the source language using a backward NMT model. We compare DALI with supervised (DALI-S) and unsupervised (DALI-U) lexicon induction. Finally, we

	Medical	Subtitles	Law	Koran
Unadapted	7.43	5.49	4.10	2.52
Copy	13.28	6.68	5.32	3.22
BT	18.51	11.25	11.55	**8.18**
DALI-U	20.44	9.53	8.63	4.90
DALI-S	19.03	9.80	8.64	4.91
DALI-U+BT	**24.34**	**13.35**	**13.74**	8.11
DALI-GIZA++	28.39	9.37	11.45	8.09
In-domain	46.19	27.29	40.52	19.40

Table 3: Comparison among different methods on adapting NMT from IT to {Medical, Subtitles, Law, Koran} domains, along with two oracle results

(1) experiment with when we directly extract a lexicon from an in-domain corpus using GIZA++ (DALI-GIZA++) and Algorithm 1, and (2) list scores for when systems are trained directly on in-domain data (In-domain). For simplicity, we test the adaptation performance of the LSTM-based NMT model, and train a LSTM-based NMT with the same architecture on out-of-domain corpus for English-to-German back-translation.

First, DALI is competitive with BT, outperforming it on the medical domain, and underperforming it on the other three domains. Second, the gain from DALI is orthogonal to that from BT – when combining the pseudo-parallel in-domain corpus obtained from DALI-U with that from BT, we can further improve by 2-5 BLEU points on three of four domains. Second, the gains through usage of both DALI-U and DALI-S are surprisingly similar, although the lexicons induced by these two methods have only about 50% overlap. Detailed analysis of two lexicons can be found in Section 3.5.

3.3 Word-level Translation Accuracy

Since our proposed method focuses on leveraging word-for-word translation for data augmentation, we analyze the word-for-word translation accuracy for unseen in-domain words. A source word is considered as an unseen in-domain word when it never appears in the out-of-domain corpus. We examine two question: (1) How much does each adaptation method improve the translation accuracy of unseen in-domain words? (2) How does the frequency of the in-domain word affect its translation accuracy?

To fairly compare various methods, we use a lexicon extracted from the in-domain parallel data with the GIZA++ alignment toolkit as a reference lexicon L_g. For each unseen in-domain source word in the test file, when the corresponding target
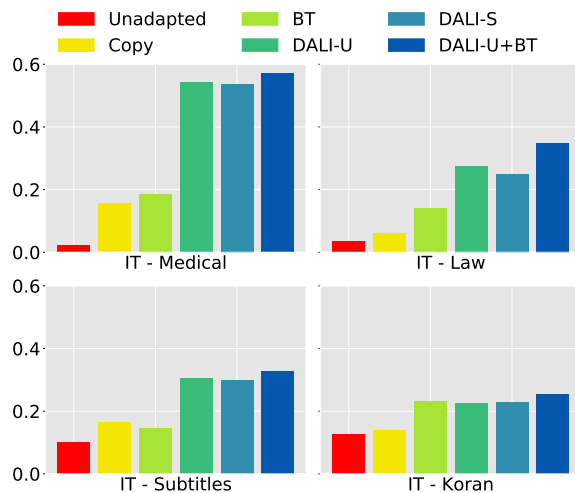

Figure 2: Translation accuracy of in-domain words of the test set on several data augmentation baseline and our proposed method with IT as the out domain

word in L_g occurs in the output, we consider it as a "hit" for the word pair.

First, we compare the percentage of successful in-domain word translations across all adaptation methods. Specifically, we scan the source and reference of the test set to count the number of valid hits C, then scan the output file to get the count C_t in the same way. Finally, the hit percentage is calculated as $\frac{C_t}{C}$. The results on experiments adapting IT to other domains are shown in Figure 2. The hit percentage of the unadapted output is extremely low, which confirms our assumption that in-domain word translation poses a major challenge in adaptation scenarios. We also find that all augmentation methods can improve the translation accuracy of unseen in-domain words but our proposed method can outperform all others in most cases. The unseen in-domain word translation accuracy is quantitatively correlated with the BLEU scores, which shows that correctly translating in-domain unseen words is a major factor contributing to the improvements seen by these methods.

Second, to investigate the effect of frequency of word-for-word translation, we bucket the unseen in-domain words by their frequency percentile in the pseudo-in-domain training dataset, and calculate calculate the average translation accuracy of unseen in-domain words within each bucket. The results are plotted in Figure 3 in which the x-axis represents each bucket within a range of frequency percentile, and the y-axis represents the average translation accuracy. With increasing frequency of words in the pseudo-in-domain data, the trans-
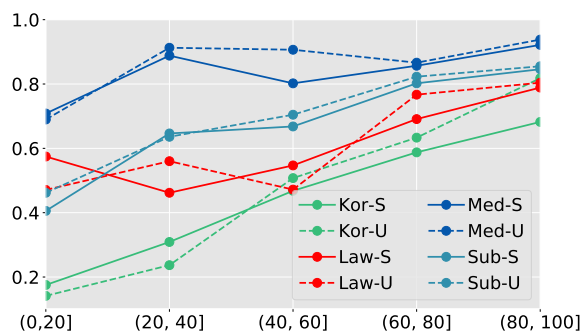
Figure 3: Translation accuracy of in-domain unseen words in the test set with regards to the frequency percentile of lexicon words inserted in the pseudo-in-domain training corpus.

lation accuracy also increases, which is consistent with our intuition that the neural network would be able to remember high frequency tokens better. Since the absolute value of the occurrences are different among all domains, the numerical values of accuracy within each bucket vary across domains, but all lines follow the ascending pattern.

3.4 When do Copy, BT and DALI Work?

From Figure 2, we can see that Copy, BT and DALI all improve the translation accuracy of in-domain unseen words. In this section, we explore exactly what types of words each method improves on. We randomly pick some in-domain unseen word pairs which are translated 100% correctly in the translation outputs of systems trained with each method. We also count these word pairs' occurrences in the pseudo-in-domain training set. The examples are demonstrated in Table 5.

We find that in the case of Copy, over 80% of the successful word translation pairs have the same spelling format for both source and target words, and almost all of the rest of the pairs share subword components. In short, and as expected, Copy excels on improving accuracy of words that have identical forms on the source and target sides.

As expected, our proposed method mainly increases the translation accuracy of the pairs in our induced lexicon. It also leverages the subword components to successfully translate compound words. For example, "monotherapie" does not occur in our induced lexicon, but the model is still able to translate it correctly based on its subwords "mono@@" and "therapie" by leveraging the successfully induced pair "therapie" and "therapy".

It is more surprising to find that adding a back translated corpus significantly improves the

model's ability to translate in-domain unseen words correctly, even if the source word never occurs in the pseudo-in-domain corpus. Even more surprisingly, we find that the majority of the correctly translated source words are not segmented at all, which means that the model does not leverage the subword components to make correct translations. In fact, for most of the correctly translated in-domain word pairs, the source words are never seen during training. To further analyze this, we use our BT model to do word-for-word translation for these individual words without any other context, and the results turn out to be extremely bad, indicating that the model does not actually find the correspondence of these word pairs. Rather, it rely solely on the decoder to make the correct translation on the target side for test sentences with related target sentences in the training set. To verify this, Table 4 demonstrates an example extracted from the pseudo-in-domain training set. BT-T shows a monolingual in-domain target sentence and BT-S is the back-translated source sentence. Though the back translation fails to generate any in-domain words and the meaning is unfaithful, it succeeds to generate a similar sentence pattern as the correct source sentence, which is "... ist eine (ein) ... , die (das) ... enthält .". The model can easily detect the pattern through the attention mechanism and translate the highly related word "medicine" correctly.

From the above analysis, it can be seen that the improvement brought by the augmentation of BT and DALI are largely orthogonal. The former utilizes the highly related contexts to translate unseen in-domain words while the latter directly injects reliable word translation pairs to the training corpus. This explains why we get further improvements over either single method alone.

3.5 Lexicon Coverage

Intuitively, with a larger lexicon, we would expect a better adaptation performance. In order to examine this hypothesis, we do experiments using pseudo-in-domain training sets generated by our induced lexicon with various coverage levels. Specifically, we split the lexicon into 5 folds randomly and use a portion of it comprising folds 1 through 5, which correspond to 20%, 40%, 60%, 80% and 100% of the original data. We calculate the coverage of the words in the Medical test set comparing with each pseudo-in-domain train-

BT-S	es ist eine Nachricht , die die aktive Substanz **enthält** .	BT-T	Invirase is a **medicine** containing the active substance saquinavir .
Test-S	ABILIFY ist ein **Arzneimittel** , das den Wirkstoff Aripiprazol **enthält** .	Test-T	Prevenar is a **medicine** containing the design of Arixtra .

Table 4: An example that shows why BT could translate the OOV word "Arzneimittel" correctly into "medicine". "enthált" corresponds to the English word "contain". Though BT can't translate a correct source sentence for augmentation, it generates sentences with certain patterns that could be identified by the model, which helps translate in-domain unseen words.

Type	Word Pair	Count
Copy	(tremor, tremor)	452
	(347, 347)	18
BT	(ausschuss, committee)	0
	(apotheker, pharmacist)	0
	(toxizität, toxicity)	0
DALI	(müdigkeit, tiredness)	444
	(therapie, therapy)	9535
	(monotherapie, monotherapy)	0

Table 5: 100% successful word translation examples from the output of the IT to Medical adaptation task. The Count column shows the number of occurrences of word pairs in the pseudo-in-domain training set.
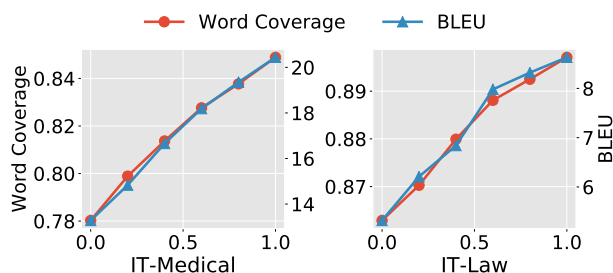

Figure 4: Word coverage and BLEU score of the Medical test set when the pseudo-in-domain training set is constructed with different level of lexicon coverage.

ing set. We use each training set to train a model and get its corresponding BLEU score. From Figure 4, we find that the proportion of the used lexicon is highly correlated with both the known word coverage in the test set and its BLEU score, indicating that by inducing a larger and more accurate lexicon, further improvements can likely be made.

3.6 Semi-supervised Adaptation

Although we target *unsupervised* domain adaptation, it is also common to have a limited amount of in-domain parallel sentences in a semi-supervised adaptation setting. To measure efficacy of DALI in this setting, we first pre-train an NMT model on a parallel corpus in the IT domain, and adapt it

to the medical domain. The pre-trained NMT obtains 7.43 BLEU scores on the medical test set. During fine-tuning, we sample 330,278 out-of-domain parallel sentences, and concatenate them with 547,325 pseudo-in-domain sentences generated by DALI and the real in-domain sentences. We also compare the performance of fine-tuning on the combination of the out-of-domain parallel sentences with only real in-domain sentences. We vary the number of real in-domain sentences in the range of [20K, 40K, 80K, 160K, 320K, 480K]. In Figure 5(a), semi-supervised adaptation outperforms unsupervised adaptation after we add more than 20K real in-domain sentences. As the number of real in-domain sentences increases, the BLEU scores on the in-domain test set improve, and fine-tuning on both the pseudo and real in-domain sentences further improves over fine-tuning sorely on the real in-domain sentences. In other words, given a reasonable number of real in-domain sentences in a common semi-supervised adaptation setting, DALI is still helpful in leveraging a large number of monolingual in-domain sentences.

3.7 Effect of Out-of-Domain Corpus

The size of data that we use to train the unadapted NMT and BT NMT models varies from hundreds of thousands to millions, and covers a wide range of popular domains. Nonetheless, the unadapted NMT and BT NMT models can both benefit from training on a large out-of-domain corpus. We examine the question: how does fine-tuning on weak and strong unadapted NMT models affect the adaptation performance? To this end, we compare DALI and BT on adapting from subtitles to medical domains, where the two largest corpus in subtitles and medical domains have 13.9 and 1.3 million sentences. We vary the size of out-of-domain corpus in a range of [0.5, 1, 2, 4, 13.9] million, and fix the number of in-domain target sentences to 0.6 million. In Figure 5(b), as the size of out-of-domain parallel sentences increases,

Source	ABILIFY ist ein Arzneimittel , das den Wirkstoff Aripiprazol enthlt .	BLEU
Reference	abilify is a medicine containing the active substance aripiprazole .	1.000
Unadapted	the time is a figure that corresponds to the formula of a formula .	0.204
Copy	abilify is a casular and the raw piprexpression offers .	0.334
BT	prevenar is a medicine containing the design of arixtra .	0.524
DALI	abilify is a arzneimittel that corresponds to the substance aripirazole .	0.588
DALI+BT	abilify is a arzneimittel , which contains the substance aripiprazole .	0.693

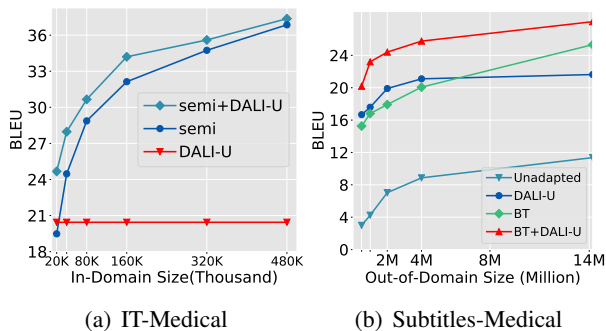Table 6: Translation outputs from various data augmentation method and our method for IT→Medical adaptation.

Figure 5: Effect of training on increasing number of in-domain (a) and out-of-domain (b) parallel sentences

Domain	Base	DALI	Word	Subword
Medical	28.94	30.06	44.1%	69.1%
IT	18.27	23.88	45.1%	77.4%
Subtitles	22.59	22.71	35.9%	62.5%
Law	24.26	24.55	59.0%	73.7%
Koran	11.64	12.19	83.1%	74.5%

Table 7: BLEU scores of LSTM based NMT models when trained on WMT14 De-En data (Base), and adapted to one domain (DALI). The last two columns show the percentage of source word/subword overlap between the training data on the WMT domain and other five domains.

we have a stronger upadapted NMT which consistently improves the BLEU score of the in-domain test set. Both DALI and BT also benefit from adapting a stronger NMT model to the new domain. Combining DALI with BT further improves the performance, which again confirms our finding that the gains from DALI and BT are orthogonal to each other. Having a stronger BT model improves the quality of synthetic data, while DALI aims at improving the translation accuracy of OOV words by explicitly injecting their translations.

3.8 Effect of Domain Coverage

We further test the adaptation performance of DALI when we train our base NMT model on the WMT14 German-English parallel corpus. The corpus is a combination of Europarl v7, Common Crawl corpus and News Commentary, and consists of 4,520,620 parallel sentences from a wider range of domains. In Table 7, we compare the BLEU scores of the test sets between the unadapted NMT and the adapted NMT using DALI-U. We also show the percentage of source words or subwords in the training corpus of five domains being covered by the WMT14 corpus. Although the unadapted NMT system trained on the WMT14 corpus obtains higher scores than that trained on the corpus of each individual domain, DALI still im-

proves the adaptation performance over the unadapted NMT system by up to 5 BLEU score.

3.9 Qualitative Examples

Finally, we show outputs generated by various data augmentation methods. Starting with the unadapted output, we can see that the output is totally unrelated with the reference. By adding the copied corpus, words that have the same spelling in the source and target languages e.g. "abilify" are correctly translated. With back translation, the output is more fluent; though keywords like "abilify" are not well translated, in-domain words that are highly related with the context like "medicine" are correctly translated. DALI manages to translate in-domain words like "abilify" and "substance", which are added by DALI using the induced lexicon. By combining both BT and DALI, the output becomes fluent and also contains correctly translated in-domain keywords of the sentence.

4 Related Work

There is much work on *supervised* domain adaptation setting where we have large out-of-domain parallel data and much smaller in-domain parallel data. Luong and Manning (2015) propose training a model on an out-of-domain corpus and do fine-tuning with small sized in-domain parallel data

to mitigate the domain shift problem. Instead of naively mixing out-of-domain and in-domain data, Britz et al. (2017) circumvent the domain shift problem by jointly learning domain discrimination and the translation. Joty et al. (2015) and Wang et al. (2017) address the domain adaptation problem by assigning higher weight to out-of-domain parallel sentences that are close to the in-domain corpus. Our proposed method focuses on solving the adaptation problem with no in-domain parallel sentences, a strict unsupervised setting.

Prior work on using monolingual data to do data augmentation could be easily adapted to the domain adaptation setting. Early studies on data-based methods such as self-enhancing (Schwenk, 2008; Lambert et al., 2011) translate monolingual source sentences by a statistical machine translation system, and continue training the system on the synthetic parallel data. Recent data-based methods such as back-translation (Sennrich et al., 2016a) and copy-based methods (Currey et al., 2017) mainly focus on improving fluency of the output sentences and translation of identical words, while our method targets OOV word translation. In addition, there have been several attempts to do data augmentation using monolingual source sentences (Zhang and Zong, 2016; Chinea-Rios et al., 2017). Besides, model-based methods change model architectures to leverage monolingual corpus by introducing an extra learning objective, such as auto-encoder objective (Cheng et al., 2016) and language modeling objective (Ramachandran et al., 2017). Another line of research on using monolingual data is unsupervised machine translation (Artetxe et al., 2018; Lample et al., 2018b,a; Yang et al., 2018). These methods use word-for-word translation as a component, but require a careful design of model architectures, and do not explicitly tackle the domain adaptation problem. Our proposed data-based method does not depend on model architectures, which makes it orthogonal to these model-based methods.

Our work shows that apart from strengthening the target-side decoder, direct supervision over the in-domain unseen words is essential for domain adaptation. Similar to this, a variety of methods focus on solving OOV problems in translation. Daumé III and Jagarlamudi (2011) induce lexicons for unseen words and construct phrase tables for statistical machine translation. However, it is non-trivial to integrate lexicon into NMT models that lack explicit use of phrase tables. With regard to NMT, Arthur et al. (2016) use a lexicon to bias the probability of the NMT system and show promising improvements. Luong and Manning (2015) propose to emit OOV target words by their corresponding source words and do post-translation for those OOV words with a dictionary. Fadaee et al. (2017) propose an effective data augmentation method that generates sentence pairs containing rare words in synthetically created contexts, but this requires parallel training data not available in the fully unsupervised adaptation setting. Arcan and Buitelaar (2017) leverage a domain-specific lexicon to replace unknown words after decoding. Zhao et al. (2018) design a contextual memory module in an NMT system to memorize translations of rare words. Kothur et al. (2018) treats an annotated lexicon as parallel sentences and continues training the NMT system on the lexicon. Though all these works leverage a lexicon to address the problem of OOV words, none specifically target translating in-domain OOV words under a domain adaptation setting.

5 Conclusion

In this paper, we propose a *data-based, unsupervised adaptation* method that focuses on domain adaption by lexicon induction (DALI) for mitigating unknown word problems in NMT. We conduct extensive experiments to show consistent improvements of two popular NMT models through the usage of our proposed method. Further analysis show that our method is effective in fine-tuning a pre-trained NMT model to correctly translate unknown words when switching to new domains.

References

Mihael Arcan and Paul Buitelaar. 2017. Translating domain-specific expressions in knowledge bases with neural machine translation. *CoRR*, abs/1709.02184.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. *International Conference on Learning Representations*.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Mara Chinea-Rios, Álvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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, Copenhagen, Denmark. Association for Computational Linguistics.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 407–412. Association for Computational Linguistics.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. 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.

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

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. 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 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, Melbourne, Australia. Association for Computational Linguistics.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

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*.

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

Prajit Ramachandran, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Copenhagen, Denmark. Association for Computational Linguistics.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *International Workshop on Spoken Language Translation (IWSLT) 2008*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

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

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised neural machine translation with weight sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–55, Melbourne, Australia. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. 2018. Addressing troublesome words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 391–400, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

A.1 Hyper-parameters

For the RNN-based model, we use two stacked LSTM layers for both the encoder and the decoder with a hidden size and a embedding size of 512, and use feed-forward attention (Bahdanau et al., 2015). We use a Transformer model building on top of the OpenNMT toolkit (Klein et al., 2017) with six stacked self-attention layers, and a hidden size and a embedding size of 512. The learning rate is varied over the course of training (Vaswani et al., 2017).

	LSTM	XFMR
Embedding size	512	512
Hidden size	512	512
# encoder layers	2	6
# decoder layers	2	6
Batch	64 sentences	8096 tokens
Learning rate	0.001	-
Optimizer	Adam	Adam
Beam size	5	5
Max decode length	100	100

Table 8: Configurations of LSTM-based NMT and Transformer (XFMR) NMT, and tuning parameters during training and decoding

A.2 Domain Shift

To measure the extend of domain shift, we train a 5-gram language model on the target sentences of the training set on one domain, and compute the average perplexity of the target sentences of the training set on the other domain. In Table 9, we can find significant differences of the average perplexity across domains.

Domain	Medical	IT	Subtitles	Law	Koran
Medical	1.10	2.13	2.34	1.70	2.15
IT	1.95	1.21	2.06	1.83	2.05
Subtitles	1.98	2.13	1.31	1.84	1.82
Law	1.88	2.15	2.50	1.12	2.16
Koran	2.09	2.23	2.08	1.94	1.11

Table 9: Perplexity of 5-gram language model trained on one domain (columns) and tested on another domain (rows)

A.3 Lexicon Overlap

Table 10 shows the overlap of the induced lexicons from supervised, unsupervised induction and GIZA++ extraction across five domains. The second and third column show the percentage of unique lexicons induced only by unsupervised induction and supervised induction respectively, while the last column shows the percentage of the lexicons induced by both methods.

Corpus	Unsupervised	Supervised	Intersection
Medical	5.3%	5.4%	44.7%
IT	4.1%	4.1%	45.2%
Subtitles	1.0%	1.0%	37.1%
Law	4.4%	4.5%	45.7%
Koran	2.1%	2.0%	40.6%

Table 10: Lexicon overlap between supervised, unsupervised and GIZA++ lexicon.

Domain	|In|	Medical	IT	Subtitles	Law	Koran
Medical	125724	0 (0.00)	123670 (0.98)	816762 (6.50)	159930 (1.27)	12697 (0.10)
IT	140515	108879 (0.77)	0 (0.00)	818303 (5.82)	167630 (1.19)	12512 (0.09)
Subtitles	857527	84959 (0.10)	101291 (0.12)	0 (0.00)	129323 (0.15)	3345 (0.00)
Law	189575	96079 (0.51)	118570 (0.63)	797275 (4.21)	0 (0.00)	10899 (0.06)
Koran	18292	120129 (6.57)	134735 (7.37)	842580 (46.06)	182182 (9.96)	0 (0.00)

Table 11: Out-of-Vocabulary statistics of German Words across five domains. Each row indicates the OOV statistics of the out-of-domain (row) corpus against the in-domain (columns) corpus. The second column shows the vocabulary size of the out-of-domain corpus in each row. The remaining columns (3rd-7th) show the number of domain-specific words in each in-domain corpus with respect to the out-of-domain corpus, and the ratio between the number of out-of-domain corpus and the domain specific words.

Domain	|In|	Medical	IT	Subtitles	Law	Koran
Medical	68965	0 (0.00)	57206 (0.83)	452166 (6.56)	72867 (1.06)	15669 (0.23)
IT	70652	55519 (0.79)	0 (0.00)	448072 (6.34)	75318 (1.07)	14771 (0.21)
Subtitles	480092	41039 (0.09)	38632 (0.08)	0 (0.00)	53984 (0.11)	4953 (0.01)
Law	92501	49331 (0.53)	53469 (0.58)	441575 (4.77)	0 (0.00)	13399 (0.14)
Koran	22450	62184 (2.77)	62973 (2.81)	462595 (20.61)	83450 (3.72)	0 (0.00)

Table 12: Out-of-Vocabulary statistics of English Words across five domains. Each row indicates the OOV statistics of the out-of-domain (row) corpus against the in-domain (columns) corpus. The second column shows the vocabulary size of the out-of-domain corpus in each row. The remaining columns (3rd-7th) show the number of domain-specific words in each in-domain corpus with respect to the out-of-domain corpus, and the ratio between the number of out-of-domain corpus and the domain specific words.