Iterative Constrained Back-Translation for Unsupervised Domain Adaptation of Machine Translation

Hongxiao Zhang[1], Hui Huang[2], Jiale Gao[1], Yufeng Chen[1]*, Jinan Xu[1], Jian Liu[1]
[1]Beijing Jiaotong University, Beijing, China
[2]Harbin Institute of Technology, Harbin, China
{hongxiaozhang,jialegao,chenyf,jaxu,jianliu}@bjtu.edu.cn,
huanghui_hit@126.com

Abstract

Back-translation (BT) has been proven to be effective in unsupervised domain adaptation of neural machine translation (NMT). However, the existing back-translation methods mainly improve domain adaptability by generating in-domain pseudo-parallel data that contains sentence-structural knowledge, paying less attention to the in-domain lexical knowledge, which may lead to poor translation of unseen in-domain words. In this paper, we propose an **I**terative **C**onstrained **B**ack-**T**ranslation (**ICBT**) method to incorporate in-domain lexical knowledge on the basis of BT for unsupervised domain adaptation of NMT. Specifically, we apply lexical constraints into back-translation to generate pseudo-parallel data with in-domain lexical knowledge, and then perform round-trip iterations to incorporate more lexical knowledge. Based on this, we further explore sampling strategies of constrained words in ICBT to introduce more targeted lexical knowledge, via domain specificity and confidence estimation. Experimental results on four domains show that our approach achieves state-of-the-art results, improving the BLEU score by up to 3.08 compared to the strongest baseline, which demonstrates the effectiveness of our approach. The codes and models are publicly available at https://github.com/zzzxiaohong/ICBT.

1 Introduction

Neural machine translation (NMT) has made breakthroughs in resource-rich domains (Bahdanau et al., 2015; Vaswani et al., 2017), which requires abundant in-domain parallel data (Koehn and Knowles, 2017). Unfortunately, there is no enough parallel data for many domains, while the monolingual corpus is much easier to obtain. Therefore, unsupervised domain adaptation of NMT, which aims to improve in-domain translation through out-of-domain parallel corpus and in-domain monolingual corpus (Chu and Wang, 2018), has been extensively researched in recent literatures (Gulcehre et al., 2015; Sennrich et al., 2016a; Dou et al., 2019a).

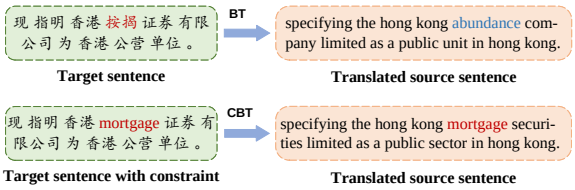

Figure 1: An example of the English-Chinese translation task to illustrate the effect of constrained back-translation (CBT) compared to back-translation (BT). The red fonts indicate the ground truth of the term, and the blue fonts show incorrect translation.

Among the existing techniques for unsupervised domain adaptation of NMT, the data-based ones are a significant part (Chu and Wang, 2018), which usually use the in-domain monolingual corpus to build pseudo-parallel data, and then use the synthetic data to fine-tune the pre-trained NMT model (Currey et al., 2017; Hu et al., 2019). Back-translation (BT) (Sennrich et al., 2016a) is one of the most basic data-based approaches. A series of studies on improving BT have emerged in recent years. For example, Wang et al. (2019) make better use of the back-translated synthetic data by introducing confidence estimation. Hoang et al. (2018) design iterative BT through iterations of forward and backward translation models to improve the translation quality of single-round BT. And some work further optimizes the iterative BT by filtering or selecting the data (Dou et al., 2020; Kumari et al., 2021). Although the above methods are proven to be effective, they only focus on reinforcing the sentence-structural knowledge provided by BT when building pseudo-parallel data, which pay less attention to in-domain lexical knowledge.

Daumé III and Jagarlamudi (2011) point out that the mistranslation of unseen (out-of-vocabulary) words accounts for a large proportion when trans-

*Yufeng Chen is the corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5054–5065
October 12–17, 2022.

ferring to a new domain. As we all know that a dictionary is a necessary aid to translate new words properly, so it is important to introduce in-domain lexical knowledge when generating pseudo-parallel data. However, the existing BT-based methods also suffer from the problem of domain shift, which leads to inaccurate translation of unseen in-domain words. Take Figure 1 as an example, BT incorrectly translates "按揭 (mortgage)" to "abundance", which fails to introduce the lexical knowledge "按揭-mortgage" to the pseudo-parallel data.

Aiming at the above issue, this paper focuses on introducing lexical knowledge into pseudo-parallel data from BT, and proposes a novel method named **I**terative **C**onstrained **B**ack-**T**ranslation (**ICBT**). Specifically, assuming that in-domain monolingual data on both source and target sides can be obtained, we firstly impose lexical constraints (by word replacement, as shown in Figure 1) on target data for the inference of back-translation. The constrained words will be forced to translate, so that the lexical knowledge is introduced. Then, we utilize round-trip iterations to incorporate more lexical knowledge into the pseudo-parallel data.

To exert more targeted constraints on BT, we further propose two sampling strategies, one is to select domain-specific words by domain difference, and the other is to select poorly translated words in BT by confidence estimation. During our BT process, we preferentially constrain these two types of words, so that more significant lexical knowledge can be incorporated.

The main contributions of this paper can be summarized as follows:

- We are the first to apply lexical constraints to BT, and propose an Iterative Constrained Back-Translation (ICBT) method to improve the unsupervised domain adaptation of NMT.
- To create more targeted lexical constraints, we propose two strategies for sampling constrained words, via domain specificity and confidence estimation. Experiments show that the two strategies are complementary.
- We conduct experiments on four domains of the English-Chinese public datasets. The experimental results show that our method claimed improvement in all domains, with a maximum of +8.45 BLEU scores over the strongest baseline and a maximum of +32.33 BLEU scores over the unadapted model. Besides, the translation accuracy of in-domain lexicons is improved by up to 7.99%.

2 Related Work

2.1 Supervised Domain Adaptation of NMT

If a small number of in-domain parallel data can be obtained, the domain adaptation of NMT can be performed in a supervised manner. The easiest way is to directly fine-tune a model pre-trained on an out-of-domain corpus with a small amount of in-domain parallel data (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016). Based on this, Khayrallah et al. (2018) add additional items to minimize the cross-entropy between the output word distribution of the model and the out-of-domain model. Gu et al. (2019) model domain-specific information and enhance the performance of translation through adversarial training. Gu and Feng (2020) address the catastrophic forgetting problem of domain adaptation by freezing some module or neuron.

2.2 Unsupervised Domain Adaptation of NMT

Unlike in-domain parallel data, the in-domain monolingual data is much easier to obtain, which makes the research on unsupervised domain adaptation of NMT increasingly popular.

The unsupervised domain adaptation of NMT is mainly divided into model-based and data-based methods (Chu and Wang, 2018). Some model-based approaches introduce language models (Gulcehre et al. (2015); Dou et al. (2019b)) or auto-encoders (Cheng et al., 2016) for NMT models. Other studies introduce domain and task embedding learners during training (Dou et al., 2019a), or extend back-translation with additional Domain-Repaired models (Wei et al., 2020).

Among the data-based approaches, some studies generate pseudo-parallel data by back-translation (Sennrich et al., 2016a) and copy-based methods (Currey et al., 2017). Alternatively, Aharoni and Goldberg (2020) select domain-appropriate data from a common corpus through a self-supervised language model. In addition, some other studies use in-domain lexical knowledge to help domain adaptation. Hu et al. (2019) use lexical induction to generate dictionary, and perform word-by-word translation to generate pseudo-parallel data. Pourdamghani et al. (2019) utilize word-by-word trans-

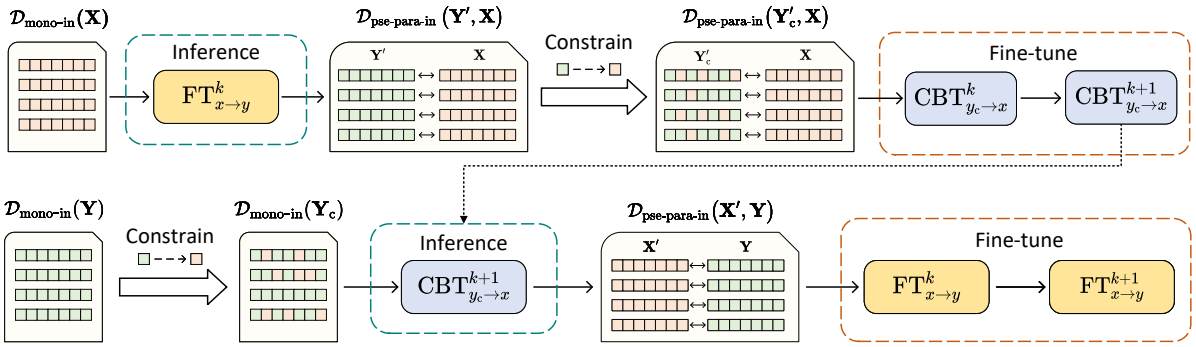

Figure 2: The schematic of our proposed ICBT method at iteration k; $\mathcal{D}_{\text{mono-in}}(\cdot)$ represents the in-domain monolingual corpus, $D_{\text{pse-para-in}}(\cdot)$ represents the in-domain pseudo-parallel sentence pairs, and $\mathbf{X}$ and $\mathbf{Y}$ represent the source and target language, respectively.

lation to generate *translationese*, and then translate *translationese* into fluent sentences.

In this work, we focus on the data-based unsupervised domain adaptation of NMT and propose a method to improve the ability of back-translation (BT) (Sennrich et al., 2016a) in domain adaptation. There are several similar studies, Wang et al. (2019) use confidence estimation to better handle the noise of synthetic corpus, Hoang et al. (2018) generate better synthetic parallel data by iterating forward and backward translation models, and Kumari et al. (2021) improve iterative back-translation by introducing classifiers to filter the synthetic data. But they only focus on utilizing the sentence-structural knowledge of BT and pay less attention to the use of lexical knowledge from in-domain monolingual data, which may lead to the mistranslation of unseen in-domain words.

3 Our Approach

The Iterative Constrained Back-Translation (ICBT) method proposed in this paper aims to introduce the lexical knowledge from monolingual data into back-translation. In this section, we first describe how to constrain back-translation and perform round-trip iterations (§ 3.1), and then we introduce two strategies for sampling more targeted constrained words (§ 3.2).

3.1 Iterative Constrained Back-Translation

3.1.1 Lexically Constrained Back-Translation

We perform lexical constraints by replacing target constrained words with their source corresponding words in the bilingual dictionary. When training the constrained back-translation, the model learns to directly copy the constraints into the translated sentences, so that during the inference, constrained words are forced to translate to their corresponding words (Song et al., 2019).

Supposing that two in-domain (IND) monolingual corpora ($\mathcal{D}_{\text{mono-in}}(\mathbf{X})$ and $\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$) and an out-of-domain (OOD) parallel corpus ($\mathcal{D}_{\text{para-out}}(\mathbf{X}, \mathbf{Y})$) can be obtained in the unsupervised scenario, we firstly constrain the target of OOD parallel data to generate $\mathcal{D}_{\text{para-out}}(\mathbf{X}, \mathbf{Y}_{\text{c}})$. We use $\mathcal{D}_{\text{para-out}}(\mathbf{X}, \mathbf{Y})$ to train a source-to-target forward translation model $\text{FT}^{0}_{x\to y}$, and use $\mathcal{D}_{\text{para-out}}(\mathbf{X}, \mathbf{Y}_{\text{c}})$ to train a target-to-source constrained back-translation model $\text{CBT}^{0}_{y_{\text{c}}\to x}$.

Then, we match each target sentence y ($y \in \mathcal{D}_{\text{mono-in}}(\mathbf{Y})$) with the in-domain bilingual dictionary V and select words to be constrained, thereby generating $\mathcal{D}_{\text{mono-in}}(\mathbf{Y}_{\text{c}})$. Since the dictionary matches a large proportion of words, and too many replacement words may destroy the syntactic structure of the sentence (see detailed analysis in § 5.3), we limit the number of replacements per sentence to n. If there are more than n matched words, they will be randomly sampled n for replacement.

Finally, $\mathcal{D}_{\text{mono-in}}(\mathbf{Y}_{\text{c}})$ is fed into the pre-trained constrained back-translation (CBT) model for inference. In this way, we can get the pseudo-parallel data $\mathcal{D}_{\text{pse-para-in}}(\mathbf{X}', \mathbf{Y})$, which will be used to fine-tune the forward translation (FT) model.

Since the constrained lexicon pairs are integrated into the inference results, the FT model can learn the in-domain lexical knowledge. On the contrary, BT without constraints can hardly provide in-domain lexical knowledge.

3.1.2 Round-Trip Iteration

To further encourage the integration of lexical knowledge, we utilize an iterative translation process, as illustrated in Algorithm 1. For intuitive-

ness, we also show a schematic diagram of the round-trip iteration in Figure 2.

The iteration starts with models ($\mathrm{FT}^0_{x\to y}$ and $\mathrm{CBT}^0_{y_c\to x}$) pre-trained with out-of-domain parallel data, and takes the in-domain monolingual data ($\mathcal{D}_{\text{mono-in}}(\mathbf{X})$ and $\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$) as well as the in-domain bilingual dictionary V as input. The CBT model is first fine-tuned with pseudo-parallel data $\mathcal{D}_{\text{pse-para-in}}(\mathbf{X}, \mathbf{Y}'_c)$, and then the FT model is fine-tuned with pseudo-parallel data $\mathcal{D}_{\text{pse-para-in}}(\mathbf{X}', \mathbf{Y})$.

During iteration, CBT produces pseudo-parallel data with in-domain lexical knowledge, which will be integrated with sentence-structural knowledge when fine-tuning FT. So that FT produces target sentences more in line with the domain. And correspondingly, there will be more domain-related lexical constraints on these target sentences, which is also beneficial for fine-tuning CBT.

Algorithm 1 Round-trip iterative training process for ICBT

Input: pre-trained NMT models $\mathrm{FT}^0_{x\to y}$ and $\mathrm{CBT}^0_{y_c\to x}$), in-domain monolingual data $\mathcal{D}_{\text{mono-in}}(\mathbf{X})$ and $\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$, bilingual dictionary V, maximum number of iterations K

Output: forward translation model $\mathrm{FT}^K_{x\to y}$

$k = 0$;
for $k < K$ **do**
  Fine-tune the CBT model:
  Use $\mathrm{FT}^k_{x\to y}$ to infer $\mathcal{D}_{\text{mono-in}}(\mathbf{X})$ and create the pseudo data $\mathcal{D}_{\text{pse-para-in}}(\mathbf{Y}', \mathbf{X})$;
  Constrain on the target of the pseudo data with V to generate $\mathcal{D}_{\text{pse-para-in}}(\mathbf{Y}'_c, \mathbf{X})$;
  Use $\mathcal{D}_{\text{pse-para-in}}(\mathbf{Y}'_c, \mathbf{X})$ to fine-tune $\mathrm{CBT}^k_{y_c\to x} \Rightarrow \mathrm{CBT}^{k+1}_{y_c\to x}$;
  Fine-tune the FT model:
  Constrain on the target monolingual data $\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$ with V to get $\mathcal{D}_{\text{mono-in}}(\mathbf{Y}_c)$;
  Use $\mathrm{CBT}^{k+1}_{y_c\to x}$ to infer $\mathcal{D}_{\text{mono-in}}(\mathbf{Y}_c)$, and get the pseudo data $\mathcal{D}_{\text{pse-para-in}}(\mathbf{X}', \mathbf{Y})$;
  Use $\mathcal{D}_{\text{pse-para-in}}(\mathbf{X}', \mathbf{Y})$ to fine-tune $\mathrm{FT}^k_{x\to y} \Rightarrow \mathrm{FT}^{k+1}_{x\to y}$;
  $k = k + 1$;
end for
return $\mathrm{FT}^K_{x\to y}$

3.1.3 Bilingual Dictionary Induction

Since the bilingual dictionary is required for lexical constraints, we obtain them through automatic induction. Considering there are no IND parallel corpora, we use unsupervised lexical induction to create a bilingual dictionary.

Unsupervised lexical induction aims to extract dictionaries from non-parallel data automatically. The mainstream approach is to map the source and target word embeddings to the same representation space and find words with close distances in the cross-lingual space as translation candidates.

In this paper, we first use $\mathcal{D}_{\text{para-out}}(\mathbf{X}, \mathbf{Y})$, $\mathcal{D}_{\text{mono-in}}(\mathbf{X})$, and $\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$ to train word embeddings in the source and target languages by *FastText* (Bojanowski et al., 2017). Then we follow Artetxe et al. (2018) to build cross-lingual embedding representations by self-learning, and find the nearest neighbors of the source and target word.

3.2 Constrained Lexicon Sample Strategy

In the lexical constraints process in § 3.1, matched words are randomly sampled, which leads to some common words (such as "the", "she", "this") being constrained. So we explore delicate strategies for sampling constrained words, based on two standards: domain specificity (§ 3.2.1) and translation confidence (§ 3.2.2).

3.2.1 Domain Specificity

As shown in the ICBT-DomainSpec method of Figure 3, we use masked language models (MLMs) to calculate domain difference to help us judge the domain specificity of words. Specifically, we follow Devlin et al. (2019) to train an out-of-domain MLM (MLM_{out}) and an in-domain MLM (MLM_{in}) using the target of the out-of-domain parallel data and the in-domain target monolingual data, respectively. Assuming that the dictionary matches the set of words $\boldsymbol{w} = \{w_1, w_2, \ldots, w_m\}$ in the sentence $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\} (\boldsymbol{y} \in \mathcal{D}_{\text{mono-in}}(\mathbf{Y}))$, for each word $w \in \boldsymbol{w}$, we perform the following operations:

- Mask y_i ($y_i = w$) in $\boldsymbol{y}$ to creat $\boldsymbol{y}^{mask} = \{y_1, y_2, \ldots, y_{i-1}, [\text{mask}], y_{i+1}, \ldots, y_n\}$.
- Feed $\boldsymbol{y}^{mask}$ into MLM_{out} and MLM_{in} respectively, and obtain the outputs $\boldsymbol{y}_t^{d_{out}}$ and $\boldsymbol{y}_t^{d_{in}}$ of the two models. The probability of predicting the [mask] position as word w is defined as:

$$p_d = \log p\left(\boldsymbol{y}_t^d[i] = w | \boldsymbol{y}^{mask}\right) \quad (1)$$

  where d represents domain, $d \in \{d_{out}, d_{in}\}$.
- Calculate the probability difference:

$$\Delta p = p_{d_{in}} - p_{d_{out}} \quad (2)$$

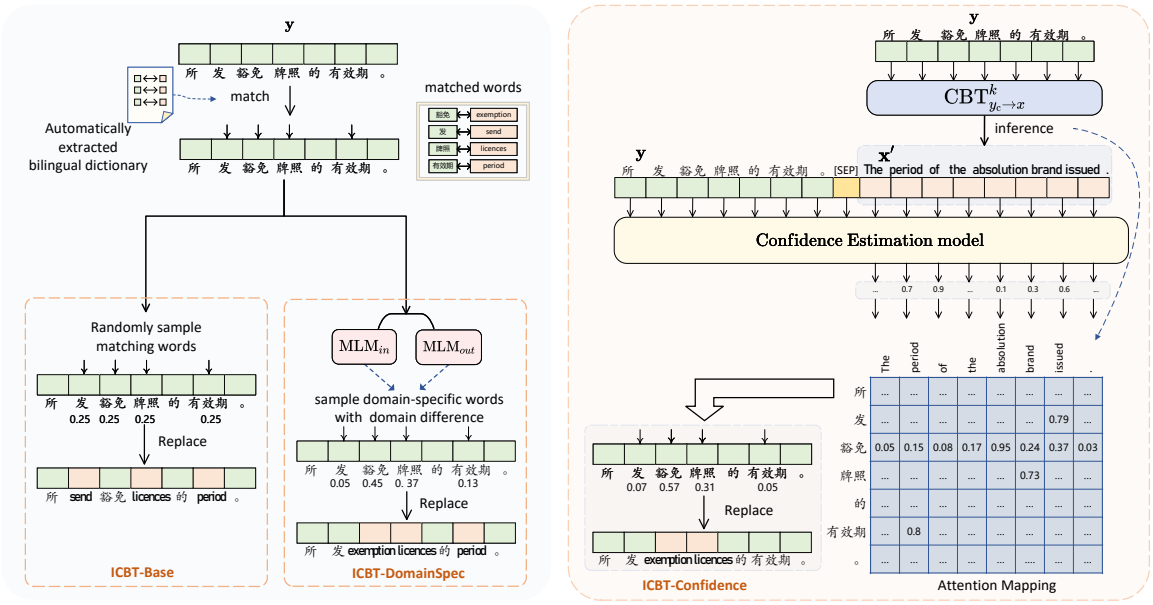

Figure 3: Different word constraints sampling methods proposed in this paper. We introduced three forms, namely ICBT-Base for baseline (§ 3.1), ICBT-DomainSpec for domain-specific words (§ 3.2.1), and ICBT-Confidence for poorly translated words (§ 3.2.2). In the attention matrix, value $w_{i,j}$ represents the correlation between the i-th word in the input sentence and the j-th word in the predicted sentence, for example, $w_{3,5}$ represents the correlation between the word "豁免 (which means exemption)" and the word "absolution" is 0.95.

We use Δp as the criterion for judging the domain specificity of the word w. A large Δp means that the word has a strong domain specificity. When constraining, we replace the top-n words with the largest Δp.

3.2.2 Confidence Estimation

Confidence estimation is utilized to select words that are poorly translated in the target monolingual data ($\mathcal{D}_{\text{mono-in}}(\mathbf{Y})$) during back-translation inference. Generally, these words also need to be supplemented by lexical knowledge.

As the ICBT-Confidence method shown in Figure 3, we first use a masked language model to perform confidence estimation on words in source translation data ($\mathcal{D}_{\text{pse-in}}(\mathbf{X}')$), following Zheng et al. (2021). Specifically, OOD parallel data is used to train an estimation model. We splice each source and target sentence pair of the parallel data, mask some words on the target, then feed them into the estimation model, and minimize the following loss function:

$$\mathcal{L} = -\sum_{n=1}^{N} \log p\left(\boldsymbol{y}_n^m | \boldsymbol{x}_n, \boldsymbol{y}_n^o; \boldsymbol{\theta}\right) \quad (3)$$

where $\boldsymbol{y}_n^m$ is the masked part of the target sentence, and $\boldsymbol{y}_n^o$ is the unmasked part, $\boldsymbol{x}_n$ is the source sentence, N is the number of OOD parallel sentence pairs, and $\boldsymbol{\theta}$ is the model parameter.

During inference, the target monolingual sentence $\boldsymbol{y}$ ($\boldsymbol{y} \in \mathcal{D}_{\text{mono-in}}(\mathbf{Y})$) and the source predicted sentence $\boldsymbol{x}$ ($\boldsymbol{x} \in \mathcal{D}_{\text{pse-in}}(\mathbf{X}')$) are spliced and fed into the pre-trained estimation model. The estimation model calculates the probability that words in $\boldsymbol{x}$ can be recovered after being masked. The higher the recovery probability is, the higher the confidence score of the word is, and vice versa. In this way, the confidence score of each word in the source predicted data is obtained.

Then we obtain the confidence score of each word in the target monolingual data through the attention mapping, which is obtained from the penultimate layer (proved by Garg et al. (2019) to be more inclined to learn alignment) of the back-translation model. More specifically, for a sentence $\boldsymbol{x}$ of the predicted data, the confidence score of each word $x_j \in \boldsymbol{x}$ is denoted as s_{x_j}. The confidence score s_{y_i} of the word y_i in the input sentence $\boldsymbol{y}$ is calculated by:

$$s_{y_i} = \sum_{j=1}^{N} w_{i,j} \times s_{x_j} \quad (4)$$

where N is the length of the predicted sentence $\boldsymbol{x}$, $w_{i,j}$ is a value in attention mapping (see the example at Figure 3), which is taken as the weight, in other words, the contribution of each s_{x_j} to s_{y_i}.

	Education	Laws	Science	Thesis	Average
without iteration					
Unadapted	10.52	23.88	6.00	6.67	11.77
Back-Translation (Sennrich et al., 2016a)	12.71	32.70	6.62	11.22	15.81
Translationese (Pourdamghani et al., 2019)	12.79	32.61	6.27	11.18	15.71
CBT-Base (*ours*)	14.25†	37.05†	7.17†	12.12†	17.65
CBT-DomainSpec (*ours*)	**14.65**†	**38.41**†	7.19†	**12.48**†	**18.18**
CBT-Confidence (*ours*)	14.30†	38.11†	**7.23**†	12.29†	17.98
with iteration					
Iterative Back-Translation (Hoang et al., 2018)	14.63	47.56	8.25	13.31	20.94
CFIBT (Kumari et al., 2021)	14.84	47.30	8.80	12.79	20.93
ICBT-Base (*ours*)	14.91†	49.51†	**8.91**†	14.05†	21.85
ICBT-DomainSpec (*ours*)	16.11†	52.12†	8.62†	14.76†	22.90
ICBT-Confidence (*ours*)	**16.36**†	52.79†	7.81†	14.70†	22.92
ICBT-ALL (*ours*)	16.10†	**56.21**†	8.50†	**15.25**†	**24.02**

Table 1: Comparative results of unsupervised domain adaptation on the English-Chinese translation task. *News* is used as the out-of-domain data. "CBT-*" are our methods without iteration. † denotes the improvement over other methods is statistically significant with $p < 0.01$.

Domain	Train	Dev	Test
News	1,252,977	1,664	1,357
Education	447,000	3,000	790
Laws	217,000	3,000	456
Science	267,000	3,000	503
Thesis	297,000	3,000	625

Table 2: Corpus statistics for our experiments.

After getting the confidence score of each word in the target, we choose the top-n words that have the lowest score and are in the domain dictionary to be constrained.

4 Experiments

4.1 Setup

Datasets. We conduct our experiments on English-Chinese datasets. The parallel corpus LDC[1] in *News* domain is used as the out-of-domain dataset. For in-domain dataset, we use the UM-corpus (Tian et al., 2014), which provides parallel sentence pairs in eight domains, and we choose four domains of *Laws*, *Education*, *Science*, and *Thesis* to conduct experiments. To obtain the in-domain development set, we randomly sample 3K sentence pairs in each domain. Data statistics are shown in Table 2. We follow Hu et al. (2019) to construct a non-parallel monolingual corpus for each domain. Specifically, we randomly divide the parallel corpus into two equal parts, and take the source sentences of the former part and the target sentences of the latter part as our monolingual data.

Jieba[2] is used to segment Chinese data and Moses[3] is used to segment English. And all English words are converted to lowercase. After that, we segment words into subwords through Byte Pair Encoding (Sennrich et al., 2016b) and construct joint vocabulary for both languages.

Models and Parameters. We implement the Transformer$_{\text{base}}$ (Vaswani et al., 2017) based on the Fairseq (Ott et al., 2019) as our translation model. BERT-base-Chinese[4] model is used as the masked language model in the ICBT-DomainSpec method. To get the in-domain model MLM_{in} and the out-of-domain model MLM_{out}, we fine-tune the BERT on two kinds of data for 5 epochs respectively. Moreover, we use multilingual BERT[4] (mBERT) as the confidence estimation model for the ICBT-Confidence method. For each method, we conduct 3 iterations. SacreBLEU[5] python package is used to calculate the BLEU score.

Baselines. We compare our method with the following methods:

- **Unadapted**. The translation model is trained on the out-of-domain training set and directly evaluated on the in-domain test set.

[1] https://www.ldc.upenn.edu/

[2] https://github.com/fxsjy/jieba

[3] http://www.statmt.org/moses/

[4] https://huggingface.co

[5] https://github.com/mjpost/sacrebleu

- **Back-Translation** (Sennrich et al., 2016a). A method for generating in-domain pseudo-parallel data from a target-to-source NMT model and target monolingual data.

- **Iterative Back-Translation** (Hoang et al., 2018). Both forward and backward translation models are used for round-trip iteration to optimize and generate better in-domain pseudo-parallel data.

- **Translationese** (Pourdamghani et al., 2019). This method first uses word-by-word translation to generate *translationese*, and then generates fluent translation sentences using *translationese*. It improves translation performance by introducing in-domain lexical knowledge.

- **CFIBT** (Kumari et al., 2021). Classifier models are introduced to filter pseudo-parallel data generated by the back-translation, thereby optimizing the iterative back-translation.

4.2 Main Results

This paper introduces three forms of ICBT, namely ICBT-Base (§ 3.1), ICBT-DomainSpec (§ 3.2.1), and ICBT-Confidence (§ 3.2.2). To prove the complementarity of our methods, we combine the data obtained from both ICBT-DomainSpec and ICBT-Confidence at each iteration to conduct experiments (ICBT-ALL). Furthermore, we also compare the performance of our method without iteration. The experimental results are shown in Table 1.

Firstly, our proposed methods achieve optimal results in all domains. The average of four domains improves by up to 12.25 BLEU over the unadapted model and 3.08 BLEU over the strongest baseline. The most considerable improvement is in *Laws*, which improves by up to 8.45 BLEU over the strongest baseline. We believe that it is because *Laws* includes more terms needed to be supplemented, so the introduction of lexical knowledge can bring a significant improvement.

Secondly, *Translationese* achieves comparable performance to *Back-Translation* and has improved in all domains compared with the unadapted baseline, illustrating the benefits of introducing in-domain lexical knowledge. Our method is better than *Translationese*, which we believe is because we fuse lexical knowledge and sentence-structural knowledge through lexically constrained BT, while *Translationese* uses knowledge separately.

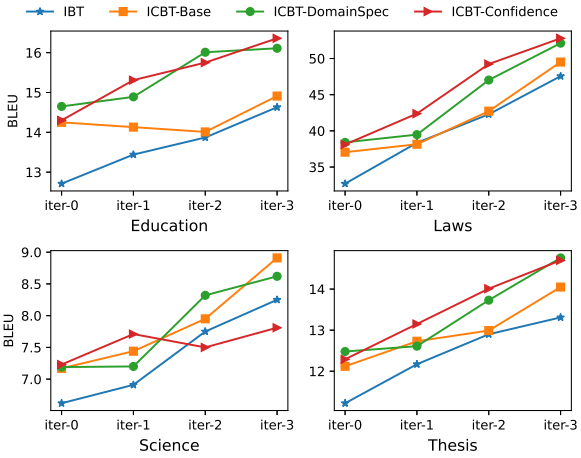

Figure 4: BLEU score of different methods according to the number of iterations on the test sets of four domains. IBT represents *Iterative Back-Translation*.

Thirdly, CFIBT achieves better performance than *iterative back-translation* (IBT) in *Education* and *Science*, and these two methods are comparable on the whole. This is in keeping with the elaboration of Kumari et al. (2021) on the performance of CFIBT in high-resource scenarios. Our methods have the improvement over both approaches. Compared to the case without iteration, our methods achieve more significant improvements over others, implying that more lexical knowledge can be incorporated through iterations.

Finally, both ICBT-DomainSpec and ICBT-Confidence achieve higher results than ICBT-Base. We conclude that it is because more targeted constraints can introduce more lexical knowledge lacking by models. ICBT-ALL achieves the best performance and obtains more significant improvement than ICBT-DomainSpec and ICBT-Confidence, especially in *Laws*. It shows that domain specificity and confidence estimation are complementary.

5 Analysis

5.1 Impact of Iterations

To further investigate the impact of iterations, we separately validate the models of each method after each iteration. The variation of BLEU scores with the number of iterations is shown in Figure 4.

Our methods show continuous improvement over IBT and the improvement is larger in the later iterations in general. We conjecture the main reasons are as follows: 1) Lexical constraints as semantic spatial anchors bring the lexical-level representations closer (Lin et al., 2020). The alignment of lexical-level makes sentence-level representa-

	Laws		Thesis	
	BLEU	$\text{Trans}_{\text{acc}}$	BLEU	$\text{Trans}_{\text{acc}}$
BT	32.70	79.85%	11.22	44.39%
CBT	**37.05**	**82.58%**	**12.12**	**49.71%**
IBT	47.56	84.99%	13.31	51.58%
ICBT	**49.51**	**88.46%**	**14.05**	**59.57%**

Table 3: Comparison of BLEU score and lexicon translation accuracy of BT and constrained BT. CBT and ICBT are the base type of our methods (§ 3.1).

tions in BT also closer in semantic space, which allows the model to transfer at both levels. 2) When the sentence-level representations are aligned, more in-domain lexicons can be generated, so that the knowledge of the in-domain dictionary is more fully utilized. 3) Lexical constraints increase the diversity of constraints in the iteration. Thus in the obtained pseudo data, the same target sentence may correspond to different source sentences, which further improves the robustness of the model.

5.2 Lexicon Translation Accuracy

To verify whether constrained back-translation really helps the translation of the in-domain lexicon, we analyze the lexical translation accuracy. Concretely, we extract the high-quality in-domain bilingual dictionary from in-domain parallel data to act as the testbed for lexical translation accuracy. For each source word in the high-quality dictionary that appears at the source of the test set, we consider it to be successfully translated if its corresponding target word occurs in the predicted data.

We conduct experiments in *Laws* and *Thesis*,[6] comparing BT and CBT with and without iteration. The comparison results are shown in Table 3. With or without iteration, the lexicon translation accuracy of CBT is always higher than that of BT, with a maximum of 3.47% in *Laws* and 7.99% in *Thesis*, verifying the benefit of constraints. Besides, the improvement with iterations is more significant than without iterations, which indicates that more lexical knowledge can be introduced through iterations, so that more in-domain lexicons can be translated correctly. We also present some cases of BT and CBT in each domain in Appendix A.

[6]We focus on domains where there is room for improvement. For *Education* and *Science*, BT can already achieve high translation accuracy of lexicon, so we do not discuss.

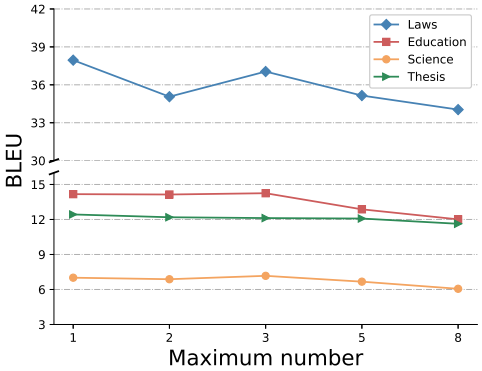

Figure 5: The effect of the maximum number of constraints on the BLEU score.

5.3 Effect of Constraints Amount

Few constraints may not achieve the desired effect, while too many constraints may destroy the syntactic structure of sentences. To validate this hypothesis, we investigate the impact of the number of constraints. Specifically, we vary the value of the maximum number n of constraints per sentence in the range of $[1, 2, 3, 5, 8]$. For each value of n, we perform CBT as described in § 3.1 and test on each resulting FT model with the test set.

As shown in Figure 5, models can achieve a high performance generally when $n \leq 3$. However, when $n = 1$ or $n = 2$, the model is unstable. For example, when $n = 1$, the model performs slightly worse on *Education* and *Science*, and when $n = 2$, it performs poorly on *Laws*. It is consistent with our conjecture that the lexical knowledge can not be introduced into the model enough when there are few constrained words. When $n > 3$, the performance deteriorates as the increase of n, indicating that too many constraints may damage the syntactic structure of sentences, thus making the performance worse. With the above analysis, we set $n = 3$ in other experiments in this paper.

6 Conclusion and Future Work

This paper proposes a method for unsupervised domain adaptation of NMT named Iterative Constrained Back-Translation (ICBT), in which lexical constraints are applied to back-translation, aiming to incorporate in-domain lexical knowledge into synthetic parallel data from BT. Besides, we propose two strategies for sampling constraints to exert more targeted constraints. We conduct experiments on English-Chinese translation tasks in four domains. The experiments show that our method can introduce beneficial lexical knowledge to BT, thus

achieving state-of-the-art results.

We believe that the lexical constraint is not only suitable for unsupervised domain adaptation, but also promising in the semi-supervised scenario with a small amount of in-domain parallel corpus. In the future, we will explore the application of lexical constraints in supervised or semi-supervised domain adaptation of NMT.

Acknowledgements

The research work descried in this paper has been supported by the National Key R&D Program of China (2020AAA0108001) and the National Nature Science Foundation of China (No. 61976016, 61976015, and 61876198). The authors also would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

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.

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.

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*, pages 407–412, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5894–5904, Online. Association for Computational Linguistics.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. 2019a. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1417–1422, Hong Kong, China. Association for Computational Linguistics.

Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. 2019b. Domain differential adaptation for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 59–69, Hong Kong. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *ArXiv preprint*, abs/1612.06897.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

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

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3081–3091, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, abs/1503.03535.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

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

Huda Khayrallah, Brian Thompson, Kevin Duh, and Philipp Koehn. 2018. Regularized training objective for continued training for domain adaptation in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 36–44, Melbourne, Australia. Association for Computational Linguistics.

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.

Surabhi Kumari, Nikhil Jaiswal, Mayur Patidar, Manasi Patwardhan, Shirish Karande, Puneet Agarwal, and Lovekesh Vig. 2021. Domain adaptation for NMT via filtered iterative back-translation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 263–271, Kyiv, Ukraine. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

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

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.

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.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 791–802, Hong Kong, China. Association for Computational Linguistics.

Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. 2020. Iterative domain-repaired back-translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5884–5893, Online. Association for Computational Linguistics.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Case Study

We feed one case from each domain into the final fine-tuned forward translation models of BT, CBT, IBT, and ICBT for inference, and the obtained results are shown in Table 4. It can be seen that the constrained BT outperforms the BT in the translation of in-domain words, no matter with or without iteration. In the case of *Education*, the ordinary BT translates "spring" to "春天", which does not fit the current context. On the contrary, constrained BT can correctly translate "spring" to "弹簧". It shows that lexical constraints can introduce in-domain lexical knowledge into the synthetic pseudo-parallel data.

From the cases of *Laws* and *Thesis*, the performance of ICBT is stronger than that of CBT, and it can translate more in-domain words. For example, in the case of *Laws*, CBT only successfully translates "public office" into "公职", but does not successfully translate "appoint or remove", which is a strongly domain-specific expression. But through iteration, the in-domain lexical knowledge and specific expressions are further enriched, so ICBT successfully translates "appoint or remove" into "任免". In addition, ICBT can also learn the domain-specific expression of translating "holders of public office" into "公职 人员" under the promotion of lexical knowledge "公职", indicating that lexical knowledge can not only bring alignment knowledge, but also enrich the sentence-structural knowledge of BT, so that the generated sentences are more in line with the domain expression.

	Education
Source	deformation of a spring is linear with the force applied on it.
Reference	弹簧 的 变形 量 与 施加 在 其 上 的 作用力 成 线性关系 。
BT	春天 变形 是 线性 的 , 应用 于 它 的 力量 。
CBT	弹簧 变形 与 应用 的 力量 是 线性 的 。
IBT	一个 春天 的 变形 与 应用 在 它 上 的 力量 是 线性 的 。
ICBT	弹簧 的 变形 与 应用 在 它 上 的 力量 是 线性 的 。
	Laws
Source	to appoint or remove holders of public office in accordance with legal procedures;
Reference	依照 法定程序 任免 公职 人员 ;
BT	按照 法定程序 委任 或 免任 公用 处所 的 持有人 ;
CBT	依照 法定程序 委任 或 免除 公职 人员 ;
IBT	依照 法定程序 委任 或 罢免 公职 的 持有人 ;
ICBT	依照 法定程序 任免 公职 人员 ;
	Science
Source	only when its melting-point temperature is reached does iron start to pass into a liquid.
Reference	只有 当 熔点 温度 达到 时 , 铁 才 开始 变成 液体 。
BT	只有 当 它 融化 的 温度 达到 时 , 才 会 有 铁的开始 , 进入 液态 。
CBT	当 熔点 温度 达到 时 , 铁 开始 进入 液体 。
IBT	只有 当 它 的 墨点 温度 达到 时 , 铁的开始 传入 液体 。
ICBT	只有 当 它 的 熔点 温度 达到 时 , 铁 才 会 开始 传入 液体 。
	Thesis
Source	effect of exciting frequency on the residual stress of the vibrating solidification casting
Reference	激振 频率 对 振动 凝固 铸件 残余 应力 的 影响
BT	振动 稳态 洞穴 的 振动 频度 对 振动 稳态 洞室 残余 压力 的 影响
CBT	振动 频率 对 振动 凝固 铸件 残余 应力 的 影响
IBT	激发 频率 对 振动 固化 铸件 残余 应力 的 影响
ICBT	激振 频率 对 振动 凝固 铸件 残余 应力 的 影响

Table 4: The translation examples show the effect of constrained back-translation. We identify aligned words in red and blue front.