FGraDA: A Dataset and Benchmark for Fine-Grained Domain Adaptation in Machine Translation

Wenhao Zhu[1,2], **Shujian Huang**[1,2], **Tong Pu**[1,2], **Pingxuan Huang**[3], **Xu Zhang**[4], **Jian Yu**[4], **Wei Chen**[4], **Yanfeng Wang**[4], **Jiajun Chen**[1,2]
[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization
[3] University of Michigan, USA [4] Sogou Inc. Beijing, China
{zhuwh, putong}@smail.nju.edu.cn, {huangsj, chenjj}@nju.edu.cn, pxuanh@umich.edu
{zhangxu216526, yujian216093, chenweibj8871, wangyanfeng}@sogou-inc.com

Abstract

Previous research for adapting a general neural machine translation (NMT) model into a specific domain usually neglects the diversity in translation within the same domain, which is a core problem for domain adaptation in real-world scenarios. One representative of such challenging scenarios is to deploy a translation system for a conference with a specific topic, e.g., global warming or coronavirus, where there are usually extremely less resources due to the limited schedule. To motivate wider investigation in such a scenario, we present a real-world fine-grained domain adaptation task in machine translation (FGraDA). The FGraDA dataset consists of Chinese-English translation task for four sub-domains of information technology: autonomous vehicles, AI education, real-time networks, and smart phone. Each sub-domain is equipped with a development set and test set for evaluation purposes. To be closer to reality, FGraDA does not employ any in-domain bilingual training data but provides bilingual dictionaries and wiki knowledge base, which can be easier obtained within a short time. We benchmark the fine-grained domain adaptation task and present in-depth analyses showing that there are still challenging problems to further improve the performance with heterogeneous resources.

Keywords: Domain Adaptation, Fine-Grained Domains, Machine Translation

1. Introduction

Recent years have witnessed the great thrive in neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). These neural-network-based models are wildly successful when there is abundant parallel data for training. However, in most real-world scenarios, the amount of data in a specific domain is limited. Therefore, domain adaptation becomes a popular topic that aims at adapting translation models in the general domain (or a source domain) to a target domain (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Chu et al., 2017; Barone et al., 2017; Michel and Neubig, 2018; Vilar, 2018; Hu et al., 2019; Zhao et al., 2020).

We notice that current research of domain adaptation usually considers very broad target domains. E.g., the popular dataset OPUS[1] (Tiedemann, 2012) is tested for the following domains: law, medical, information technology, Koran, and subtitles (Koehn and Knowles, 2017). There are still strong diversities within each domain. For example, the subtitles domain contains subtitles from action movies, political movies, Sci-Fi movies, etc.

We suggest that there are fine-grained sub-domains within these coarse domains. The sentences or words in different sub-domains may have different language phenomena, which requires a fine-grained treatment. As shown in Table 1, at the word level, the same Chinese word "卡" may correspond to different English translations in different fine-grained information technology (IT) domains. Capturing semantic diversity may be hard in traditional coarse domain adaptation but is often needed in real-world scenarios, such as translation services for a specific conference or translating a technical monograph.

Domain	translations around the word "卡"
Autonomous Vehicles	... the wheel is *stuck* and you can't ...
AI Education	... some of these math *card* games ...
Real-Time Networks	... how to fix video *stuttering* ...
Smart Phone	... find your *SIM card* slot and ...

Table 1: An example where the Chinese word "卡" have different translations (shown in red italics fonts) in different sub-domains of information technology.

To make the situation even worse, adaptation to these fine-grained domains often face challenges as a low-resource scenario, because there are limited time and budget, e.g. for the translation service provider, to collect data (especially parallel data) in the fine-grained domain. Specific research may be needed to explore other heterogeneous resources which are more available. The hierarchy between coarse and fine-grained domains may raise other research challenges as well.

In this paper, we introduce a novel and challenging dataset for fine-grained domain adaptation in machine translation, namely FGraDA (Section 3). FGraDA includes Chinese-English translation tasks on four fine-

[1] http://opus.nlpl.eu

Domain	Dictionary (items)	Wiki knowledge base (wiki pages)	Development set (sent. pairs)	Test set (sent. pairs)
Autonomous Vehicles (AV)	275	116,381	200	605
AI Education (AIE)	270	195,339	200	1,309
Real-Time Networks (RTN)	360	111,101	200	1,303
Smart Phone (SP)	284	90,337	200	750

Table 2: Main statistics of our dataset. We report the number of items for the dictionary, the number of wiki pages in the extracted wiki knowledge base, and the number of sentence pairs of development set and test set, respectively.

grained domains: autonomous vehicles (AV), AI education (AIE), real-time networks (RTN), and smart phone (SP), which are all sub-domains of IT. The development and test sets used for evaluation are collected and anonymized from real-world conferences.

For each fine-grained domain, no parallel training data is provided, as in the real-world scenarios parallel training data is expensive to obtain. For the purpose of adaptation, we provide heterogeneous but more available resources: bilingual dictionaries and wiki knowledge base.

We conduct benchmark experiments to facilitate further comparison (Section 4), as well as in-depth analyses to show translation errors (Section 5) and interesting research challenges (Section 6). Please note that this fine-grained domain adaptation problem is so challenging that we are here only presenting and benchmarking this task and calling for attention and solutions.

2. Related Work

Besides OPUS, there are other popular datasets for domain adaptation, including IWSLT[2] (Cettolo et al., 2012) and ASPEC[3] (Nakazawa et al., 2016). The IWSLT corpus is usually used as a single target domain of technical talks. But it comprises a collection of Ted talks coming from very diverse areas, such as biology, chemistry, psychology, etc. The ASPEC corpus includes scientific papers from several different disciplines, such as physics, earth science, agriculture, etc. And these domains are not as specific as those for fine-grained scenarios. Due to the substantial diversity within each domain, these current datasets cannot be directly used to simulate the fine-grained setting.

We notice that novel fine-grained datasets always set advance a research field in natural language processing, such as the fine-grained task in entity recognition (Hovy et al., 2006) and sentiment analysis (Pontiki et al., 2014). We hope our dataset also offers a stepping stone for further research of fine-grained domain adaptation.

[2] https://wit3.fbk.eu

[3] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

3. FGraDA Dataset

The task of FGraDA[4] is to improve the translation quality for each fine-grained domain with the provided heterogeneous resources, such as bilingual dictionaries and wiki knowledge bases, so that the methodology could be used to improve real-world applications in a quick and economical way.

The dataset is built from one representative of real-world fine-grained adaptation scenarios, which is to provide translation services (i.e., simultaneous interpretation) for specific international conferences (Gu et al., 2017). These conferences mainly focus on very specific topics, such as global warming, coronavirus, etc. However, it is difficult and costly for translation service providers to obtain massive in-domain parallel data for each specific conference in a short time. Therefore it is potentially useful to explore methods of adapting the NMT system with other more available resources in such a predicament.

We present the fine-grained domains and the resources in the following subsections (main statistics are shown in Table 2).

3.1. Fine-Grained Domains

We select four real-world conferences as representatives to construct the dataset. Each conference[5] is organized for a particular topic of IT, namely autonomous vehicles, AI education, real-time networks, and smart phone, which could be seen as four fine-grained domains.

These fine-grained domains have obviously different topics. But they may still share some common words or language phenomenons of the IT domain, showing an interesting characteristic inside the domain hierarchy, which are less studied in previous researches.

3.2. Bilingual Dictionary

Compared to parallel data, a set of domain-specific keywords or phrases may be much easier or cheaper to obtain. Translating these keywords or phrases can

[4] All the FGraDA dataset resources is released at https://github.com/OwenNJU/FGraDA

[5] The four conferences are the Global AI and Robotics Conference (CCF-GAIR2019), the GIIS China Education Industry Innovation Summit (GIIS2019), the Real-Time Internet Conference (RTC2019), and Apple Events (held in 2018

Autonomous Vehicles	AI Education	Real-Time Networks	Smart Phone
自动驾驶 - self-driving	知识检索 - knowledge retrieval	直播 - live streaming	蓝牙 - bluetooth
超声波雷达 - ultrasonic radar	虚拟教学 - virtual teaching	丢包 - packet loss	高动态范围成像 - HDR
车道协同 - lane coordination	脑电图 - EEG	网络地址转换 - NAT	焦外 - bokeh
激光雷达 - LiDAR	聊天机器人 - chatbot	传输层 - transport layer	帧率 - fps
行人检测 - pedestrian detection	机器学习 - machine learning	延迟 - latency	蜂窝网络 - cellular network

Table 3: Examples of the annotated bilingual dictionary

Domain	seed pages	one-hop-link pages
Autonomous Vehicle	19,277 / 490	97,104 / 1,522
AI Education	35,615 / 636	159,724 / 1,536
Real-Time Networks	17,930 / 565	93,171 / 1,386
Smart Phone	15,944 / 452	74,393 / 1,736

Table 4: Detailed statistics of our wiki knowledge base. In each cell, the left numbers correspond to the number of extracted pages and the right numbers correspond to the average number of words contained in one page.

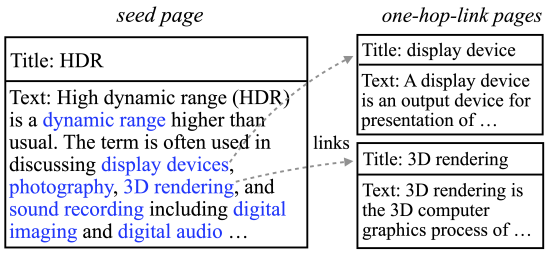

Figure 1: Illustration of the wiki knowledge base provided in our dataset. The seed page contains words (in blue font) that have links pointing to other pages.

provide important, domain-specific word-level correspondences between the two languages and act as the starting point of the adaptation process. As the dictionary will later be used for retrieving wiki resources, the quality and domain relevance of the dictionary is quite important.

Therefore, we manually build a small set of domain-specific keywords/phrases for each domain as bilingual dictionaries. Table 3 shows some examples. To make sure that the selection and translation of domain-specific words are reliable, we have all the dictionary items checked by linguistic experts.

3.3. Wiki Knowledge Base

Wikipedia is a useful resource for machine translation (Hálek et al., 2011; Wu et al., 2019). As we have obtained a manually checked in-domain dictionary, it is convenient to retrieve wikipages with the given dictionary. Because aligned wikipedia pages in different languages are not always available, we only use the wikipages in the target language as our resources.

More specifically, we first collect English wikipages[6] containing annotated dictionary keywords in their titles. These pages are closely related to the fine-grained domain (later mentioned as seed pages). Since each wikipage naturally contains words and links pointing to other related wikipages (illustrated in Figure 1), we also leverage this structural knowledge to collect more information. We collect the wikipages directly linked by links in the seed pages (later mentioned as one-hop-link pages). This one-hop constraint makes sure that these pages are relevant to our domain.

Our final knowledge base for each domain consists of both seed pages and one-hop-link pages. Statistics are shown in Table 4. Our knowledge base not only contains rich monolingual resources, but also have additional structural knowledge, e.g. link relations, which may be of potential usage.

Compared with the existing wiki knowledge base, such as DBpedia[7], which is in the form of instance-properties pairs, our wiki knowledge base costs less time to build, contains more information (in the text of the pages), and is more closely related to the fine-grained target domains.

3.4. Development and Test Set

To evaluate performances on the FGraDA task, we also collect and label parallel data. We collect 70 hours of real-world audio recordings from the four conferences mentioned above, transcribe the audio recordings with in-house tools, filter out domain-irrelevant sentences[8], and annotate them into 4,767 parallel sentence pairs (Table 2). Data desensitization is then conducted as post-editing to hide human names and company names in the annotation data to protect privacy. Each of the above steps is consulted with linguistic experts, so the labeling process is expensive, which is why a large amount of parallel data is no easy to obtain.

We split annotated data in each domain into two parts: 200 sentence pairs as the development set, and the rest as the test set. We do understand the size of the development set is relatively small for a typical large scale machine translation system, but improving the translation under this condition may be a practical problem. In contrast, the test sets are larger for better and consistent evaluation.

and 2019).

[6] https://dumps.wikimedia.your.org/enwiki/20200701/

[7] https://wiki.dbpedia.org

[8] Note that filtering is conducted as we only concern translation performance on the domain related part.

4. Benchmarks

4.1. Notations

NMT systems typically generate a target language sentence $\mathbf{y}=\{y_1, y_2, \cdots, y_{|\mathbf{y}|}\}$ given a source language sentence $\mathbf{x}=\{x_1, x_2, \cdots, x_{|\mathbf{x}|}\}$ in an end-to-end fashion. The translation probability distribution is factorized as:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} p(y_i|\mathbf{x}, \mathbf{y}_{<i}; \theta), \tag{1}$$

where y_i is the current predicted token; $\mathbf{y}_{<i}$ is the previous predicted tokens and θ is the parameters of the NMT model. The model can be trained by minimizing the loss on the training set D:

$$\mathcal{L}(D;\theta) = \sum_{(\mathbf{x},\mathbf{y})\in D} -\log p_\theta(\mathbf{y}|\mathbf{x}). \tag{2}$$

For domain adaptation, we denote the general domain parallel data as $D_{\text{g}} = \{(\mathbf{x}_{\text{g}}, \mathbf{y}_{\text{g}})\}$, the in-domain parallel data as $D_{\text{in}} = \{(\mathbf{x}_{\text{in}}, \mathbf{y}_{\text{in}})\}$ and the in-domain monolingual data in target language as $M_{\text{in}} = \{\mathbf{y}_{\text{in}}\}$, where the subscript is used to identify the corresponding domain.

4.2. Domain Adaptation Approaches

We briefly categorize and discuss existing domain adaptation approaches according to the resources they employ.

Using (pseudo) parallel data: Given in-domain parallel data, the most popular domain adaptation approach is fine-tuning (Luong and Manning, 2015), where a general domain model trained on D_{g} is continuously trained on D_{in} by minimizing the loss:

$$\mathcal{L}_{\mathcal{FT}}(D_{\text{in}};\theta_{\text{in}},\theta_{\text{g}}) = \sum_{(\mathbf{x}_{\text{in}},\mathbf{y}_{\text{in}})\in D_{\text{in}}} -\log p_{\theta_{\text{in}}}(\mathbf{y}_{\text{in}}|\mathbf{x}_{\text{in}}), \tag{3}$$

where θ_{g} and θ_{in} is the parameters of the general and adapted model, respectively. θ_{in} is initialized by θ_{g}.

When parallel data is not available, leveraging monolingual data with back-translation (BT) is an alternative. BT translates monolingual data in the target language M_{in} back to the source language to construct pseudo parallel data $\hat{D}_{\text{in}} = \{(\hat{\mathbf{x}}_{\text{in}}, \mathbf{y}_{\text{in}})\}$, and uses $\hat{D}_{\text{in}}$ as an augmentation for fine-tuning (Sennrich et al., 2016a; Hoang et al., 2018).

Parallel data is the exact type of resource that NMT systems are trained from, thus is useful for adaptation. But high quality parallel data is expensive to obtain in our scenario.

Using dictionaries: Grid beam search (GBS) (Hokamp and Liu, 2017) is proposed to incorporate dictionaries into NMT at decoding time, which extends beam search to allow the inclusion of pre-specified lexical constraints. Large-scale bilingual dictionaries can also be treated as pseudo bitext for fine-tuning (Kothur et al., 2018; Thompson et al., 2019) at training time, or for building word-by-word translation to generate pseudo bitext (Hu et al., 2019).

FGraDA provides a small but high quality set of domain specific dictionary. Using it as decoding constraints might be a reasonable choice. However, compared to general lexical constraints, domain specific constraints are harder to generate because these constraints are often rare in the general domain. To balance the constraints and the generation process, we reweight the log-likelihood during constrained generation steps:

$$\begin{aligned}\text{score}(\hat{\mathbf{y}}, \mathbf{x}) = -\sum_{i=1}^{|\hat{\mathbf{y}}|} [&\mathcal{I}(\hat{y}_i \notin \mathcal{C}) \log p(\hat{y}_i|\mathbf{x}, \hat{\mathbf{y}}_{<i}; \theta) \\ &+(1-w)\mathcal{I}(\hat{y}_i \in \mathcal{C}) \log p(\hat{y}_i|\mathbf{x}, \hat{\mathbf{y}}_{<i}; \theta)],\end{aligned} \tag{4}$$

where $\hat{\mathbf{y}}$ is the generated hypothesis sentence and $\mathcal{C}$ is the constraint set. Moreover, rather than selecting hypotheses only from beams where all the constraints are satisfied, we select highest scored hypotheses from all beams, which enables the model to ignore constraints that are harmful to the generation process.

Using Knowledge Base: Bilingual knowledge bases could be used for extracting bilingual lexicons (Zhao et al., 2020). To our best knowledge, there is not previous attempt in exploring domain related information in monolingual wikipages, as provided in FGraDA. As one of the first attempts, we simply take all sentences from seed pages as M_{in} and apply back-translation.

4.3. Benchmark Systems

We implement the following systems as benchmark baselines.

Base: Directly using a Transformer (Vaswani et al., 2017) trained on D_{g} on the target domains without any adaptation.

Dict$_{\text{GBS}}$: Performing constrained decoding (Hokamp and Liu, 2017) for Base with in-domain dictionary. We implement a weighted version described in the previous sub-section. The weight is selected on the development set (see Section 6.1 for discussions about the weight).

Dict$_{\text{FT}}$: Fine-tuning the Base model on the in-domain dictionary (Kothur et al., 2018).

Wiki$_{\text{BT}}$: Using sentences of wiki seed pages for back-translation with the Base model (Sennrich et al., 2016a).

Wiki$_{\text{BT}}$+Dict$_{\text{GBS}}$: Applying constrained decoding on Wiki$_{\text{BT}}$ model.

4.4. Experiment Settings

General Domain: We use WMT-CWMT-17 Chinese-English dataset (9 million sentence pairs) as the general domain data and train Base model with newsdev2017[9] as the development set. For back

[9]We report validation performance in Appendix A.

Model	AV	AIE	RTN	SP	Avg.
Base	34.0	31.1	16.6	22.9	26.2
Dict$_{\text{GBS}}$	34.5	31.1	17.0	23.0	26.4
Dict$_{\text{FT}}$	34.0	31.1	16.7	22.9	26.2
Wiki$_{\text{BT}}$	34.8	31.8	16.8	23.4	26.7
Wiki$_{\text{BT}}$+Dict$_{\text{GBS}}$	**35.1**	**31.9**	**17.2**	**23.6**	**27.0**

Table 5: Translation results (BLEU scores) on four fine-grained domains. Bold text identifies the best result among the benckmark result. "Avg" denotes the average results across four domains.

translation, the backward model is also trained on WMT-CWMT-17 zh-en dataset.

Data Processing: We use the open-source toolkit sentence_splitter[10] to split paragraphs in wikipages into sentences. We use the script in *moses*[11] and *jieba*[12] to tokenize the English and Chinese corpus, respectively. Byte-pair encoding (Sennrich et al., 2016b) is applied with 32k merge operations.

Implementation Details: All the models are implemented with an open source tool NJUNMT[13] and follow the architecture of transformer-base (Vaswani et al., 2017). Adam is used as the optimizer and Noam as the learning rate scheduler. We set 8k warm-up steps and a maximum learning rate as 9e-4. We train the Base model on 4 Tesla V100, which takes three days. The batch size is 3000 tokens, and the update circle is 10. Beam size is set as 5.
In all fine-tuning experiments, we choose learning rate from {1e-6, 5e-7, 1e-7, 5e-8} according to model's BLEU score on the development set.
We report detokenized case-insensitive BLEU scores calculated with *mteval-v13.pl*[14].

4.5. Benchmark Results

The benchmark results are presented in Table 5. For each domain, Dict$_{\text{GBS}}$ and Wiki$_{\text{BT}}$ improve the baseline model to some extent, while Dict$_{\text{FT}}$ barely brings any improvements. With both resources, Wiki$_{\text{BT}}$+Dict$_{\text{FT}}$ achieves the best performance among all systems. These results demonstrate the effectiveness of the heterogeneous resources. We will use this best model as the adapted model for further analyses.
However, the translation quality does not improve as greatly as reported in other research (Freitag and Al-Onaizan, 2016; Chu et al., 2017); the performance on

[10] https://github.com/mediacloud/sentence-splitter
[11] https://github.com/moses-smt/mosesdecoder
[12] https://github.com/fxsjy/jieba
[13] https://github.com/whr94621/NJUNMT-pytorch
[14] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/mteval-v13a.pl

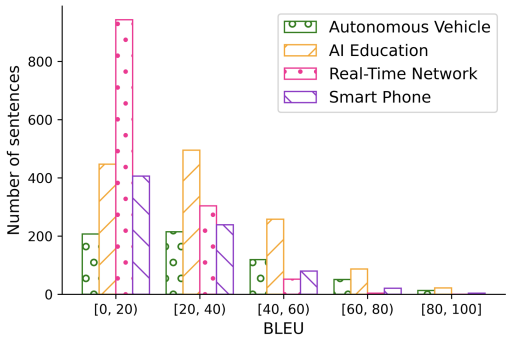

Figure 2: Distribution of sentence BLEU scores on four domains.

real-time networks and smart phone are much lower than on the other two domains, showing the diversity and difficulty of these fine-grained domains. In the following sections, we conduct further analyses to better understand the difficulty and challenges in this scenario.

5. Translation Analyses

For analysis, we translate the four test sets with the best adapted system, i.e. Wiki$_{\text{BT}}$+Dict$_{\text{GBS}}$. We compute the BLEU score for each sentence and plot their distributions in Figure 2. It is obvious that the translation performance on a large portion of test sentences is not satisfactory, e.g. under 20, leaving a large room for improvement.
We carefully go through generated translations of test sentences in each domain to analyze translation errors. We find that there are three typical types of error, which are challenging and require more attention. The error types and examples are presented in Table 6.

Mistranslating domain-specific words: Some words in the specific domain rarely appear in general domain training data, which increases the difficulty for the model to translate them. For example, the Chinese phrase "网页即时通信技术" is expected to be translated as "WebRTC", which is a domain-specific word in real-time networks. Sadly, the system generates an wordy and incorrect translation (the first case in Table 6).

Misunderstanding common words with domain specific meaning: Some frequent words are endowed with a new meaning under the context of a specific domain. Taking the Chinese word "卡" as an example, it means "card" or "carton" in most cases in the general domain; but in the networking domain, it means "stutter" (as shown in the second case in Table 6).

Under-translating the source sentence: Part of the domain-related source sentence information is missing after translation. For example, in the third case in Table 6, the translation model can not completely capture the semantic meaning of "送达模式" and only translates it as "service pattern" rather than "delivery mode".

Type I: mistranslating domain-specific words	
Source	如果你想直接从一个浏览器发送信息到另一个浏览器，唯一的办法就是使用网页即时通信技术 。
Hypothesis	If you want to send messages directly from one browser to another, the only way to do so is to use *web instant communication technology*.
Reference	The only way in which you can send a message directly from one browser to the other is using *WebRTC*.
Type II: misunderstanding common words with domain specific meaning	
Source	左边是相对卡 很多，右边是相对流畅，也有卡顿 ，但是总体上流畅度有巨大的提升。
Hypothesis	On the left is a lot of relative *cards*, on the right is relatively fluid, also there is *Carton*, but overall fluency has a great increase.
Reference	The left is relatively *stutter*. The right is relatively smooth, and there are *stutters*, but the overall fluency is greatly improved.
Type III: under-translating the source sentence	
Source	但是我们也注意到，这种送达模式在以前非常重要。
Hypothesis	But we also note that this *service pattern* was important in the past.
Reference	However, we also notice that although this *delivery mode* used to be very important.

Table 6: Typical types of errors and examples. Source fragments that are wrongly translated are shown in blue font. The error in the hypothesis and the correct translation in the reference are shown in red and italic font.

Model	AV	AIE	RTN	SP
Base	63.04	57.81	65.86	59.42
Dict$_{GBS}$	**65.84**	59.69	76.94	61.85
Wiki$_{BT}$	63.93	59.38	67.30	58.97
Wiki$_{BT}$+Dict$_{GBS}$	**65.84**	**64.22**	**87.84**	**63.07**

Table 7: The translation accuracy (%) of items in the domain dictionary. Bold font marks the model with the highest accuracy.

All the above errors are closely related to domain specific problems, which brings interesting challenges: how to solve these problems with limited resources.

6. Remaining challenges

6.1. Mining from the Dictionary

The domain dictionary contains accurate translation knowledge about the domain specific words. We conduct analyses to see whether this knowledge are properly used. We count the occurrences of the dictionary item in the source sentence and the translation results, and compute the accuracy on each test set (Table 7).

Compared with Base model, using these dictionary items as constraints Dict$_{GBS}$ improves the translation accuracy. It is interesting to see that the constrained decoding achieves even better performance when using together with the domain related monolingual data Wiki$_{BT}$. Considering that finetuning the dictionary does not lead to any improvement (Table 5), it is likely that the wiki data provides sentence-level context for the domain specific words, which encourages the generation of these words. However, even with the adapted models, a large portion of items are still mis-translated.

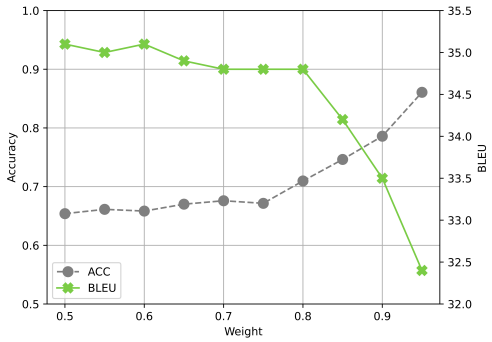

Figure 3: Dictionary words translation accuracy and BLEU scores w.r.t different weights in grid beam search on AV test set.

We set the weight of constraints at different values, and plot the translation accuracy of dictionary items and overall BLEU scores for the adapted system (Wiki$_{BT}$+Dict$_{GBS}$). As shown in Figure 3, although higher weight ensures more domain specific words are translated, BLEU score drops significantly.

The results indicate that simply forcing the models to generate infrequent in-domain words is not sufficient. Therefore, better methods for generating these domain words are still worth exploring.

6.2. Mining from Wiki Knowledge Base

Another possible reason of the poor generation of domain specific words is that they are rare in the general domain, so the model is not able to learn their proper representation. We notice that resources from Wiki knowledge base may contain rich structural knowledge that may help the system to learn the representations, i.e. to “understand” these words.

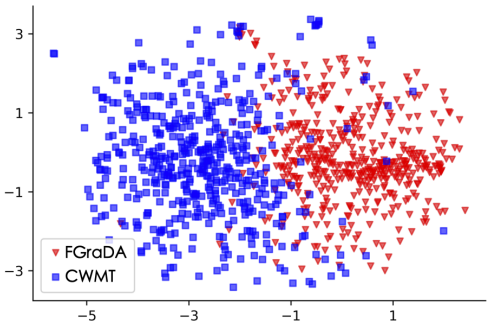

Figure 4: Visualization of sentences from FGraDA and CWMT.

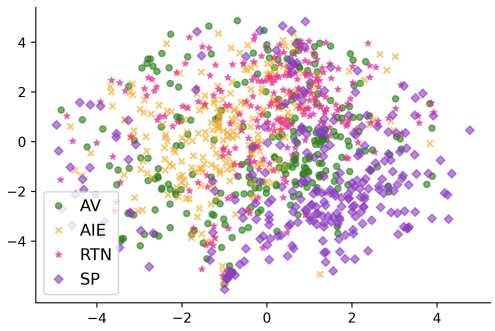

Figure 5: Visualization of sentences from the four fine-grained domains in FGraDA.

Adapt \ Test	AV	AIE	RTN	SP
AV	35.1	31.0	16.7	23.1
AIE	35.0	**31.9**	16.9	23.3
RTN	**35.2**	**31.9**	**17.2**	23.4
SP	**35.2**	**31.9**	16.9	**23.6**

Table 8: The performance on all four test sets. Each line represents a model adapted to a specific domain.

Resource	AV	AIE	RTN	SP
Dict-AV	-	5.45	5.82	3.27
Dict-AIE	5.45	-	3.70	1.45
Dict-RTN	5.82	3.70	-	5.83
Dict-SP	3.27	1.45	5.83	-
Wiki-AV	-	24.29	23.72	16.03
Wiki-AIE	24.29	-	15.20	9.98
Wiki-RTN	23.72	15.20	-	17.44
Wiki-SP	16.03	9.98	17.44	-

Table 9: Overlap ratio (%) of dictionary and wikipage resources between different domains.

On one hand, many wikipages contain definition for their title words, which is usually the first sentence in the page. For example, "High dynamic range (HDR) is a dynamic range higher than usual" gives the definition of "HDR" (illustrated in Figure 1). It might be possible to utilize such knowledge to better understand domain-specific words in sentences.

On the other hand, most wikipages have links pointing to other words or phrases that are closely related to the current title word (also illustrated in Figure 1). These words, e.g. dynamic range and 3D rendering, photography, may also help to learn the representation of the title word (e.g. HDR).

Moreover, compared to the limited amount of manually labeled dictionary items, wiki knowledge base also contains much more domain related words in the one-hop-link pages. It is interesting to explore effective methods to utilize this knowledge in the future.

6.3. Mining from the Domain Hierarchy

To get more insight into the relation between domains of FGraDA, we use BERT (Devlin et al., 2018) to encode source sentences in the test sets and visualize them with t-SNE (Maaten and Hinton, 2008). The visualizations of CWMT (general domain) and FGraDA (IT domain) data are shown in Figure 4. The two distributions are almost separated, which means that FGraDA is quite different from the general domain.

Then we plot sentences of the four fine-grained domains in Figure 5. Different from the situation in Figure 4, the four IT sub-domains have more overlaps, showing they are closely related. However, each domain still presents a unique distribution.

To quantify the diversity and relation among four sub-domains, We use the adapted model for each domain to translate the test sets of the other three domains (Table 8).

For each domain, the performance varies according to the adapted data, which demonstrates the diversity among different fine-grained domains. The best performance almost distributes on the diagonal, showing that exact resource for the fine-grained domain are most helpful in most cases. Also, the performances on some domains are different. We guess that is because some fine-grained domain, e.g. AIE and AV, may also include sentences related to other fine-grained domains.

To further understand the relation between different fine-grained domain, we also analyze the overlap ratio of other provided resources, i.e. dictionary and wiki knowledge base. Quantitative results in Table 9 show small but consistent overlap ratios of dictionary and wiki knowledge base resources (around 5% and 20%, respectively), which also indicates a close relation between these domains.

From the above results, it might be beneficial to leverage resources from other related sub-domains for adaptation. The hierarchy of coarse and fine-grained domains may also raise other interesting research topics.

7. Conclusion

This paper introduces the first fine-grained domain-adaptation dataset for machine translation, FGraDA, which presents a real-world problem. We benchmark the dataset and show the needs for fine-grained adap-

tation. We find that the NMT model usually mistranslates domain-specific words, misunderstands common words with domain-specific meaning, and undertranslates the source sentence in our task. Our analyses also show that the provided hetergeneous resources may contain useful information for the adaptation. However, current adaptation methods cannot effectively utilize these resources. The challenging problems of FGraDA encourages further exploration of dictionaries, wiki knowledge base, which might be more available than in-domain parallel data. It is also interesting to further explore the domain hierarchy as well.

8. Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. U1836221, 6217020152), National Key R&D Program of China (No. 2019QY1806).

9. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Barone, A. V. M., Haddow, B., Germann, U., and Sennrich, R. (2017). Regularization techniques for fine-tuning in neural machine translation. *arXiv preprint arXiv:1707.09920*.

Chu, C., Dabre, R., and Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Gu, J., Neubig, G., Cho, K., and Li, V. O. (2017). Learning to translate in real-time with neural machine translation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Hálek, O., Rosa, R., Tamchyna, A., and Bojar, O. (2011). Named entities from wikipedia for machine translation. In *ITAT*.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Workshop on Neural Machine Translation and Generation (WMT)*.

Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. G. (2019). Domain adaptation of neural machine translation by lexicon induction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Workshop on Neural Machine Translation (WMT)*.

Kothur, S. S. R., Knowles, R., and Koehn, P. (2018). Document-level adaptation for neural machine translation. In *Workshop on Neural Machine Translation and Generation (WMT)*.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *International Workshop on Spoken Language Translation (IWSLT)*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*.

Michel, P. and Neubig, G. (2018). Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Thompson, B., Knowles, R., Zhang, X., Khayrallah, H., Duh, K., and Koehn, P. (2019). Hablex: Human annotated bilingual lexicons for experiments in machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*.

Vilar, D. (2018). Learning hidden unit contribution for adapting neural machine translation models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Wang, Q., Li, B., Liu, J., Jiang, B., Zhang, Z., Li, Y., Lin, Y., Xiao, T., and Zhu, J. (2018). Towards building a strong transformer neural machine translation system. In *China Workshop on Machine Translation (CWMT)*.

Wu, L., Zhu, J., He, D., Gao, F., Qin, T., Lai, J., and Liu, T.-Y. (2019). Machine translation with weakly paired documents. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Zhao, Y., Zhang, J., Zhou, Y., and Zong, C. (2020). Knowledge graphs enhanced neural machine transla-

tion. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

10. Language Resource References

Mauro Cettolo and Christian Girardi and Marcello Federico. (2012). *WIT^3: Web Inventory of Transcribed and Translated Talks*.

Hovy, Eduard and Marcus, Mitchell and Palmer, Martha and Ramshaw, Lance and Weischedel, Ralph. (2006). *OntoNotes: The 90% Solution*.

Nakazawa, Toshiaki and Yaguchi, Manabu and Uchimoto, Kiyotaka and Utiyama, Masao and Sumita, Eiichiro and Kurohashi, Sadao and Isahara, Hitoshi. (2016). *Aspec: Asian scientific paper excerpt corpus*.

Pontiki, Maria and Galanis, Dimitris and Pavlopoulos, John and Papageorgiou, Harris and Androutsopoulos, Ion and Manandhar, Suresh. (2014). *SemEval-2014 Task 4: Aspect Based Sentiment Analysis*.

Tiedemann, Jörg. (2012). *Parallel Data, Tools and Interfaces in OPUS*.

A. Appendix

We report our Base model's performance on newsdev2017 and newsdev2018 in Table10. The comparison result shows that our transformer-base is comparable with other implementation (Wang et al., 2018).

	newsdev2017	newsdev2018
Ours	22.9	24.3
Wang et al.	-	24.4

Table 10: Evaluation results. All BLEU scores reported in this table is computed by *multi-bleu.pl*[16].