FDMT: A Benchmark Dataset for Fine-grained Domain Adaptation in Machine Translation

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

Abstract

Previous domain adaptation research usually neglect the diversity in translation within a same domain, which is a core problem for adapting a general neural machine translation (NMT) model into a specific domain 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. computer networks or natural language processing, where there is usually extremely less resources due to the limited time schedule. To motivate a wide investigation in such settings, we present a real-world fine-grained domain adaptation task in machine translation (FDMT). The FDMT dataset (Zh-En) consists of four sub-domains of information technology: autonomous vehicles, AI education, real-time networks and smart phone. To be closer to reality, FDMT does not employ any in-domain bilingual training data. Instead, each sub-domain is equipped with monolingual data, bilingual dictionary and knowledge base, to encourage in-depth exploration of these available resources. Corresponding development set and test set are provided for evaluation purpose. We make quantitative experiments and deep analyses in this new setting, which benchmarks the fine-grained domain adaptation task and reveals several challenging problems that need to be addressed.

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). Neural-network-based translation models are wildly successful when there is abundant parallel data for training. However, in most real-world scenarios, there is limited data in specific domains. Therefore, domain adaptation becomes a popular topic, which aims at adapting translation models in a source domain (or general 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).

Domain	Correct 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 in different sub-domains (shown in red italics fonts). Autonomous Vehicles, AI Education, Real-Time Networks and Smart Phone are four fine-grained information technology (IT) domains.

Currently, the research of domain adaptation usually consider target domains that are very broad. E.g., the popular dataset OPUS[1] (Tiedemann, 2012) has a wide collection of corpora and is tested for the following domains: law, medical, information technology (IT), Koran and subtitles (Koehn and Knowles, 2017). We notice that there are still strong diversities within each domain, which are different but related. For example, the subtitles domain data contains subtitles from action movies, political movies, Sci-Fi movies, etc.

Different from existing dataset, 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. For example in Table 1, at the word level, the same Chinese word "卡" may correspond to different English translations in different fine-grained IT domains. Capturing semantic diversity may be hard in traditional coarse domain adaptation, but is often

* Corresponding author.

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

Domain	Test set	Development set	Monolingual Data	Bilingual Dictionary
Autonomous Vehicles (AV)	605	200	733,954	118
AI Education (AIE)	1,309	200	726,356	126
Real-Time Networks (RTN)	1,303	200	939,332	120
Smart Phone (SP)	750	200	1,048,250	123

Table 2: Statistics of our dataset. We report number of sentences for test set, development set and monolingual set, and number of phrases for the dictionary.

needed in real-world scenarios, such as translation services for a specific conference or translating a technical monograph.

To make the situation even worse, adaptation to these fine-grained domain often face the challenges as a low-resource scenario, because there is limited time for translation service provider to collect data, especially parallel data, in the fine-grained domain. Specific research may be needed to explore other types of more available resources.

In this paper, we introduce a novel and challenging dataset for fine-grained domain adaptation in machine translation, namely FDMT[2], to simulate the above real-world settings. FDMT includes four fine-grained domains: autonomous vehicles (AV), AI education (AIE), real-time networks (RTN) and smart phone (SP) , which are all sub-domains of information technology (IT). For each fine-grained domain, no training parallel data is provided for better simulation of the real-world scenarios. For the purpose of adaptation, we provide other resources: a monolingual corpus, a bilingual dictionary and a knowledge base. We also manually annotated the development and test set, which is used to evaluate the performance of translation systems. In the end, we benchmark our dataset to facilitate further comparison and provide in-depth analyses.

Please note that, we are not here providing the solution to this fine-grained domain adaptation problem, but presenting this challenging task and calling for attention and solutions.

2 Related Work

Apart from OPUS, there are other popular domain adaptation datasets, including IWSLT[3] (Cettolo et al., 2012) and ASPEC[4] (Nakazawa et al., 2016). The IWSLT corpus is made up of a collection of Ted talks, and usually used as a single target domain. These talks also have various different topics, such as biology, chemistry, education, etc. ASPEC includes papers from several different scientific disciplines, such as environmentology, materials, IT, etc. These domains are not as specific as those for fine-grained scenarios. Due to the strong 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 in sentiment analysis (Pontiki et al., 2014). We hope our dataset also offers a step stone for further research of fine-grained domain adaptation.

3 FDMT Dataset

In practice, many scenarios requires a fine-grained domain adaptation. One representative of these scenarios 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, corona virus, etc., which requires the neural machine translation (NMT) models to be adapted in a fine-grained way. However, it's difficult and costly for translation service provider 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 NMT system with other more available resources in such a predicament.

Following the above setting, we select four conferences[5] as examples to construct the dataset. Each conference is organized for a particular topic of information technology, namely autonomous vehicles, AI education, real-time networks and smart phone, which could be seen as four sub-domains

[2] All the resources of the FDMT dataset will be released soon

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

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

[5] The four conferences are the Global AI and Robotics Conference (CCF-GAIR2019), the Real-Time Internet Conference (RTC2019), the GIIS China Education Industry Innovation Summit (GIIS2019) and Apple Events (held in 2018 and 2019).

Autonomous Vehicles	AI Education	Real-Time Networks	Smart Phone
自动驾驶 - self-driving	知识点 - knowledge point	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	托福 - TOEFL	网络地址转换 - NAT	焦外 - bokeh
激光雷达 - LiDAR	慕课 - MOOC	传输层 - transport layer	帧率 - fps
行人检测 - pedestrian detection	智能测评 - intelligent evaluation	延迟 - latency	蜂窝网络 - cellular network

Table 3: Examples of the annotated bilingual dictionary

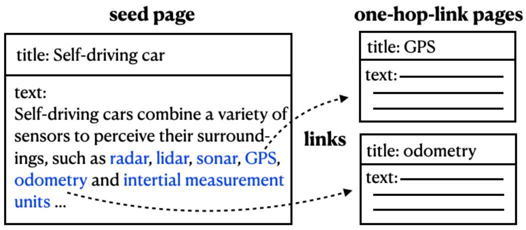

Figure 1: Illustration of the wiki knowledge graph provided in our dataset. The seed page contains words (in blue font) that have links pointing to other pages.

Domain	page nums	words per page
Autonomous Vehicle	19,277 / 97,104	487 / 1,519
AI Education	35,615 / 159,724	629 / 1,532
Real-Time Network	17,930 / 93,171	559 / 1,733
Smart Phone	15,944 / 74,393	445 / 1,733

Table 4: Statistics of our knowledge graph. We report the number of extracted pages (in column "page nums"), and average number of words contained in one page (in column "words per page"). In each cell, the left and right number correspond to the statistics of seed pages and one-hop-link pages, respectively.

of IT.

The task of FDMT is to improve the translation quality for the automatic translation services at international conferences. It usually starts with a strong translation system in the general domain. And we try to build the dataset to simulate the real-world adaptation to the fine-grained domain.

3.1 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 might provide important, domain-specific word level correspondences between the two language, which is acceptable in time and cost, and acts as the starting point of the adaptation process.

We manually build a small set of domain-specific keywords/phrases for each domain as bilingual dictionaries (Table 3) and provide them in our dataset as one kind of translation resource. To ensure the quality of the dictionary, we have all the dictionary items checked by linguistic experts. So the selection and translation of domain-specific words is reliable. However, in this case, the size of the dictionary could not be as large as traditional dictionaries.

3.2 Monolingual Data

According to the research of low-resource neural machine translation (Sennrich et al., 2016b), monolingual data in the target language will be a great alternative to parallel data, as it describes the data distribution in the target language in the specific domain. However, obtaining such monolingual data is not trivial, because the data distribution is unknown.

As we have obtained a manually checked in-domain dictionary, a quick way of obtaining monolingual data is to crawl the web, or query a search engine, with the given dictionary. We extract monolingual English sentences containing at least one dictionary item from a large pre-crawled English web text. The extracted monolingual resources is cleaned to remove duplication and the sentences longer than 80 tokens (Table 2).

3.3 Knowledge Graph

We notice that there are existing structured resources, such as knowledge graphs (KG), which may also contain domain specific information. It might have potential benefit as a complement to the monolingual and dictionary resources. Although currently there is very little research effectively employing KG for machine translation, we call for more explorations in our setting.

We build a knowledge graph from wiki in our dataset. More specifically, we collect Wikipedia pages[6] containing annotated dictionary keywords in their titles, and use them as seed pages. Since seed pages naturally contain words and links pointing to other wikipages (illustrated in Figure 1),

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

we leverage one-hop-link relations to collect more wikipages. One-hop-link pages are the wikipages directly pointed by at least one seed pages. This one-hop constraint makes sure that these pages are relevant to our domain.

Our final knowledge graph for each domain is made up of seed pages and one-hop-link pages. Statistical information can be found in Table 4. Compared with existing wiki knowledge base, such as DBpedia[7], which is in the form of *(instance, properties)*, our knowledge graph not only contains more information (in the text of the pages), but also is more closely related to the target fine-grained domains.

3.4 Development and Test Set

To evaluate the performance on the FDMT task, we also collect and label parallel data. We collected 70 hours of real-world audio recordings from the above four mentioned conferences. We transcript 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 step is consulted with linguistic experts, so the labeling process is expensive, which also shows why 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 development set is relative small for a typical large scale machine translation system, but improving the translation under this condition may raise more interesting challenges.

4 Benchmarks

4.1 Backgrounds

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 can be factorized as:

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

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

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

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 cross entropy loss on the training set D:

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

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 Existing Domain Adaptation Approaches

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

Using parallel data By given in-domain parallel data, the most effective 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} with a fixed small learning rate by minimizing the loss:

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

where ϕ is the parameters of the adapted model and initialized by θ. The model usually converges quickly within a few epochs (Barone et al., 2017).

Various fine-tuning strategies are proposed to achieve better adaptation performance, including ensemble (Freitag and Al-Onaizan, 2016), freezing partial parameters during fine-tuning (Zoph et al., 2016; Vilar, 2018; Thompson et al., 2018), mixed fine-tuning (Chu et al., 2017), adding regularization (Barone et al., 2017; Khayrallah et al., 2018; Thompson et al., 2019) and meta-learning (Finn et al., 2017; Sharaf et al., 2020), etc.

Using dictionaries Recently, people start to explore the idea of utilizing bilingual dictionaries in domain adaptation. The proposed method try to build pseudo in-domain data with dictionaries (Hu et al., 2019) which relies on bilingual lexicon induction (BLI) (Mikolov et al., 2013; Lample et al., 2018). However, it has been found that the BLI performance on distant language pairs, such as Chinese-English, is rather poor (Nakashole and Flauger, 2018; Hoshen and Wolf, 2018). Because the primary assumption of BLI is not held in this condition. Such technical drawbacks limit the generalization of proposed data augmentation methods.

Using monolingual data Back-translation is the most successful way to leverage monolingual resources (Sennrich et al., 2016a; Hoang et al., 2018). It back-translates monolingual data in target language to construct pseudo parallel pairs as an augmentation for further training.

To achieve a better augmentation effect, monolingual data selection can be performed before back-translating (Dou et al., 2020), which aims at screening out representative samples that better describe the target domain at hand. Among studies of data selection, early works utilize the language model to rank and pick sentences (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013) and recent works are based on sentence embeddings (Wang et al., 2017).

Using knowledege graphs To the best of our knowledge, there is merely one work using KG for data augmentation to enhance NMT (Zhao et al., 2020) before. However, the paper focuses on the entity based KG, which is different from KG provided in our dataset.

4.3 Baselines

Only several existing approaches can be applied in our setting. Firstly, without in-domain parallel data, it is possible to obtain pseudo parallel data through back-translation (BT) and perform fine-tuning. Secondly, it is also possible to perform data selection (DS) on monolingual data before back-translating.

As a result, we choose the following five representative models as our baselines, which use public general domain training data, monolingual set provided in FDMT. Dictionaries and knowledge graphs are not used by our baseline models and we leave the effective usage of them to be explored in the future.

BASE: A Transformer (Vaswani et al., 2017) trained on D_{g} is used directly on the target domains without any adaptation.

BT-FT: Fine-tuning with back-translation. More specifically, each in-domain sentence $\mathbf{y}_{\text{in}} \in M_{\text{in}}$ is back-translated into $\hat{\mathbf{x}}_{\text{in}}$ (Sennrich et al., 2016a). The BASE model is fine-tuned on pseudo parallel pairs $\hat{D}_{\text{in}} = \{(\hat{\mathbf{x}}_{\text{in}}, \mathbf{y}_{\text{in}})\}$.

BT-MFT: Mixed fine-tuning with back-translation. Sentences are randomly sampled from D_{g} to form a new set D'_{g}, which has the same size of M_{in}. The pre-trained translation model is finetuned on the mixed data $D'_{\text{g}} \cup \hat{D}_{\text{in}}$ (Chu et al., 2017).

BT-FT(DS): BT-FT with data selection. Monolingual data M'_{in} is sampled from M_{in} based on relevance of sentence embedding (Wang et al., 2017). The BASE model is fine-tuned on the back-translated data from M'_{in}. During data selection, we use BERT (Devlin et al., 2019) to encode sentences and the detailed algorithm is described in the Appendix A.

BT-MFT(DS): BT-MFT with data selection. The above data selection on monolingual data is conducted and employed with the mixed fine-tuning strategy.

4.4 Benchmark Settings

We use CWMT-17 Chinese-English dataset (9 million sentence pairs) as the general domain data and train BASE model with newsdev2017 as the development set. We use *moses*[9] and *jieba*[10] to tokenize the English and Chinese corpus. Byte-pair encoding (Sennrich et al., 2016b) is applied with 32k merge operations. Our models are implemented with an open source tool NJUNMT[11] and follow the architecture of transformer-base (Vaswani et al., 2017). We use Adam 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 3 days. Batch size is 3000 tokens and update circle is 10. BT-FT, BT-MFT, BT-FT(DS) and BT-MFT(DS) are fine-tuned with the same fixed learning rate 1e-5. We report below detokenized case-sensitive BLEU (Papineni et al., 2002) scores calculated with *mteval-v13.pl*[12] (Post, 2018). For data selection, we download the pretrained BERT (*bert-base-uncased*) provided in *transformers*[13] and use hidden states of the last layer as semantic representations. We suggest that it's trivial to tune the size of selected monolingual set for each sub-domain, therefore we empirically set the size as 200,000.

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

Model	Autonomous Vehicle	AI Education	Real-Time Network	Smart Phone
BASE	33.2	31.1	16.6	22.7
BT+FT	34.3(+1.1)	32.4(+1.3)	**16.8(+0.2)**	23.5(+0.8)
BT+FT(DS)	34.8(+1.6)	32.3(+1.2)	16.7(+0.1)	23.5(+0.8)
BT+MFT	34.8(+1.6)	32.5(+1.4)	**16.8(+0.2)**	23.6(+0.9)
BT+MFT(DS)	**34.9(+1.7)**	**32.6(+1.5)**	16.7(+0.1)	**24.0(+1.3)**

Table 5: Baseline results on four fine-grained domains. The numbers in the brackets represents improvement compared with BASE. Bold text means the best result in each domain.

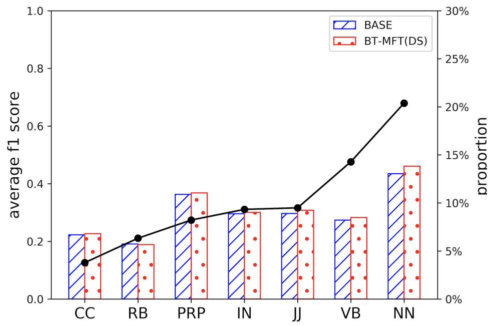

Figure 2: Average f1 scores on different POS (the bar chart) and proportion of each POS in corpus (the line chart).

Domain	Unknown	Infrequent	Precision
AV	42	2,679	63.4%
AIE	69	4,976	62.0%
RTN	245	5,648	49.9%
SP	46	3,435	58.6%

Table 6: Counts of unknown and infrequent words in test sets of four domains and overall average translation precision on them.

4.5 Benchmark Results

We present baseline results for four domains in Table 5. In the first row, BLEU scores in different ranges show that the difficulties of the four subsets vary, even though they are all IT sub-domains. BT-MFT(DS) achieves the best results on three out of four domains compared with other baselines. The result shows that back-translation, mixed fine-tuning strategy and data selection are all useful in our fine-grained setting. However, given that the fine-tuning approach in OPUS/IWSLT always leads to significant improvements (Zeng et al., 2019), the performance gap between BT-MFT(DS) and BASE is considerably weak, especially in the domain of real-time networks (RTN), which demonstrates that our proposed task is quite challenging and simply fine-tuning struggles to achieve satisfying performance in our real-world fine-grained setting. More effective usage of the resources provided in FDMT are waiting to be explored.

5 Remaining Challenges

5.1 Terrible Performance on Verbs

To better analyze the challenge that our dataset brings, we group words according to their Part-Of-Speech (POS) tags to measure baseline models' translation performance at the word level. Seven main types of words are taken into consideration: adjectives (JJ), adverbs (RB), nouns (NN), verbs (VB), coordinating conjunctions (CC), prepositions (IN) and personal pronouns (PRP). For the brevity and clarity, we take the smart phone domain as an example. For each POS, we calculate the average f1 score between output translation words and the corresponding reference words (Figure 2). The overall f1 scores are not satisfying (lower than 0.5). Especially for VB, with the second largest proportions across the corpus, the f1 score is terribly low in comparison with NN. Moreover, BT-MFT(DS) barely improves translation performance on VB comparing to the BASE model. Since VB is usually the core for overall semantics, we still need more effective approach to narrow the gap. Thus the first challenge that our dataset present is to improve translation performance on VB by a large margin.

5.2 Bad Translation of Rare Words

Another challenge of our dataset is brought by rare words because the target domain context is quite different from the general source domain. We use CWMT-17 training corpus for computing our lexical frequency counts and find that there are many unknown and infrequent words in our test sets. We calculate word translation precision in the same way as (Koehn and Knowles, 2017) and find that NMT models still have difficulty translating them (Table 6). Here is the second challenge that our dataset presents.

Error type	Wrongly translating 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 the *instant messaging technology of a web page*.
Reference	The only way in which you can send a message directly from one browser to the other is using *WebRTC*.
Error type	Wrongly translating common words which have domain specific meanings
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	There are a lot of relative *cards* on the left and a lot of fluency on the right and a lot of *carton* on the left, but overall fluency has increased dramatically.
Reference	The left is relatively *stutter*. The right is relatively smooth, and there are *stutters*, but the overall fluency is greatly improved.
Error type	Under-translation
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 notice that this *mode of service* was very important before.
Reference	However, we also notice that although this *delivery mode* used to be very important.

Table 7: Error types and examples. It's difficult for current baseline models to translate the source sentence correctly (especially sentence fragments in bold font). The error in the hypothesis and the correct translation in the reference are shown in red and italic font.

Error Type	BASE	BT-MFT(DS)
Domain-specific Words	18.18%	8.33%
Polysemy	6.31%	4.29%
Under-translation	15.40%	11.36%

Table 8: Proportion of three types of errors in the 10% worst translation results. "Domain-specific Words", "Polysemy" and "Under-translation" correspond to three categories of errors described in Section 5.3.

5.3 Inaccurate Understanding of Complex Lexical Meaning

We also dive into the translation text. We compute the BLEU score for each translation result generated by BT-MFT(DS) and go through the worst 10% results carefully. We find that words/phrases with complex lexical meaning tend be inaccurately translated. Specifically, translation errors can be categorized into three types:

Domain-specific words For example, "WebRTC" is a domain-specific word but the system generates an inaccurate translation for it (the first case in Table 7).

Common words with domain specific meaning Taking the Chinese word "卡" as an example, it means "card" in most cases, but is expected to be translated as "stutter" in the second case in Table 7.

Under-translation of the source sentence Part of the source sentence information is missed in the translation output. For example, in the third case in Table 7, the translation model can not completely capture the semantic meaning of "送达模式" and only translates it as "mode of service" rather than "delivery mode".

For comparison, we go through the translation generated by the BASE model and count the proportion of three types of errors occurred in the worst 10% translation results (Table 8). We find that BT-MFT(DS) has lower error rate on all three categories. Compared with BASE, BT-MFT(DS) substantially ameliorates domain-specific words translation which is easy to understand because target side monolingual data contains massive domain-specific words. However, for other two types of errors, BT-MFT(DS) only brings limited improvement. Generally, three problems mentioned above frequently occur in the fine-grained domain adaptation which wait to be solved. Thus the third challenge that our dataset presents is to better understand the complex lexical meaning.

6 Fine-grain vs. Coarse-grain

6.1 Coarse-grained Adaptation

What if we ignore fine-grained domains and treat them as a single coarse-grained domain like before? To our knowledge, the OPUS dataset is constructed in this way and also has an IT domain which has the

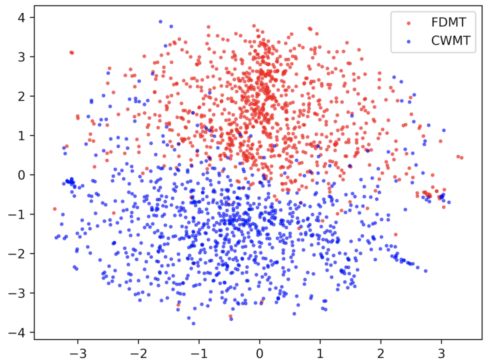

Figure 3: Visualization of our dataset FDMT and CWMT. The colors represent the dataset that each sentence belongs to.

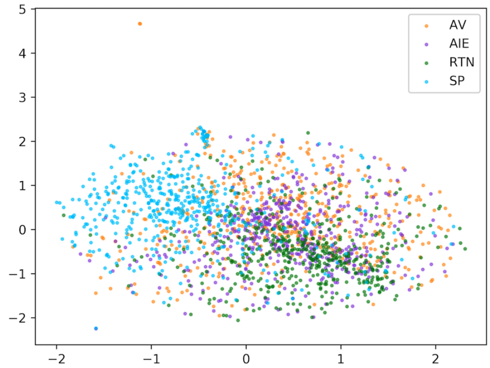

Figure 4: Visualization of four fine-grained domain contained in FDMT. The colors represent the domain for each sentence.

	AV	AIE	RTN	SP
F-adapt.	34.9	32.6	16.8	24.0
C-adapt.	33.5(-1.4)	31.8(-0.8)	16.7(-0.1)	22.7(-1.3)

Table 9: Comparison results on four fine-grained domains. "F-adapt." and "C-adapt." stands for fine-grained and coarse-grained adaptation separately. In the "F-adapt." row, we report the best performance that baseline models achieve. The numbers in the brackets represents the gap between two methods.

FT \ Test	AV	AIE	RTN	SP
AV	**34.9**	32.3(-0.3)	16.4(-0.3)	23.1(-0.9)
AIE	34.0(-0.9)	**32.6**	16.4(-0.3)	22.8(-1.2)
RTN	34.9(-0.0)	32.3(-0.3)	**16.7**	22.4(-1.6)
SP	34.8(-0.1)	32.2(-0.4)	16.4(-0.3)	**24.0**

Table 10: Adapted BT-MFT(DS) models' performance on all four test sets. The model is fine-tuned on "FT" and evaluated on "Test". Bold text means the best result achieved on each test set. The numbers in the brackets represents degradation compared with the best result.

same topic as our dataset. Thus we use the IT data from OPUS to imitate a coarse-grained adaptation. English monolingual resources provided in OPUS-IT is back-translated to generate pseudo parallel training data. Then we follow the steps of training BT-MFT to obtain the adapted model. Compared with fine-grained adaptation, coarse-grained adaptation can only achieve poorer results on our fine-grained test set (Table 9). Experimental results demonstrate the necessity of figuring out solution for the fine-grained setting.

6.2 Visualization of Sentence Representations

To get a deeper insight of our dataset, we use average-pooled BERT hidden-states as sentence embedding and visualize samples contained in FDMT with t-SNE (Maaten and Hinton, 2008). At first, we plot the distribution of CWMT and our datset FDMT. As shown in Figure 3, the two distribution are almost separated in the semantic vector space, which means that the target domain is quite different from the source domain. Then we plot four fine-grained domains in Figure 4. Each domain presents a unique distribution but also has overlaps with other IT sub-domains. This phenomena proves that even in a coarse-grained domain, there still exists different fine-grained sub-domains.

6.3 Diversity among Fine-grained Domains

To qualify the diversity between fine-grained domains, we test four adapted BT-MFT(DS) model on other three unseen domains (Table 10). We find that the best result can only be achieved by the model fine-tuned on the corresponding target domain data. Otherwise the performance on the test set will degrade which again demonstrates that there exists a wide diversity in fine-grained domains.

7 Conclusion

In this paper, we introduce the first fine-grained domain-adaptation dataset for machine translation, FDMT, which simulates a real-world adaptation problem, i.e. automatic translation services at international conferences. We benchmark this dataset and demonstrate the need of fine-grained adaptation. We revealed several challenges: current baseline models perform poorly on verbs, still have low translation accuracy on rare words and can't accurately understand the complex lexical meaning.

The challenging setting of FDMT encourages further exploration of monolingual data, dictionaries and wiki knowledge graphs, which might be more available than in-domain parallel data.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. *arXiv preprint arXiv:1707.09920*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Conference of the European Association for Machine Translation (EAMT)*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Annual Meeting of the Association for Computational Linguistics(ACL)*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*.

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

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

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Workshop on Neural Machine Translation and Generation (WMT)*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime G Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Annual Meeting of the 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 *Workshop on Neural Machine Translation and Generation (WMT)*.

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

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

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ndapandula Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 221–227.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *International Conference on Language Resources and Evaluation (LREC)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics (ACL)*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation (SemEval 2014)*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Conference on Machine Translation (WMT)*.

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

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

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot nmt adaptation. *arXiv preprint arXiv:2004.02745*.

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

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Conference on Machine Translation (WMT*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *International Conference on Language Resources and Evaluation (LREC)*.

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 (NeurIPS)*.

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

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiali Zeng, Yang Liu, Yaojie Lu, Yongjing Yin, et al. 2019. Iterative dual domain adaptation for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. 2017. Regularization techniques for fine-tuning in neural machine translation. *arXiv preprint arXiv:1707.09920*.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Conference of the European Association for Machine Translation (EAMT)*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. 2020. Dynamic data selection and weighting for iterative back-translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Annual Meeting of the Association for Computational Linguistics(ACL)*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*.

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

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

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Workshop on Neural Machine Translation and Generation (WMT)*.

Yedid Hoshen and Lior Wolf. 2018. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime G Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Annual Meeting of the 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 *Workshop on Neural Machine Translation and Generation (WMT)*.

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

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

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. *arXiv preprint arXiv:1805.01817*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ndapandula Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 221–227.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *International Conference on Language Resources and Evaluation (LREC)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics (ACL)*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation (SemEval 2014)*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Conference on Machine Translation (WMT)*.

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

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

Amr Sharaf, Hany Hassan, and Hal Daumé III. 2020. Meta-learning for few-shot nmt adaptation. *arXiv preprint arXiv:2004.02745*.

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

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. 2018. Freezing subnetworks to analyze domain adaptation in neural machine translation. In *Conference on Machine Translation (WMT*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *International Conference on Language Resources and Evaluation (LREC)*.

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 (NeurIPS)*.

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

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jiali Zeng, Yang Liu, Yaojie Lu, Yongjing Yin, et al. 2019. Iterative dual domain adaptation for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Zhao, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2020. Knowledge graphs enhanced neural machine translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

A Appendices

Detailed description of our data selection method in Section 3.3.

Algorithm 1 Data Selection

Input: English sentences $\{\bar{\mathbf{y}}_1, \cdots, \bar{\mathbf{y}}_{|D_{\text{dev}}|})\}$ in dev set D_{dev}, monolingual set $M_{\text{in}} = \{(\mathbf{y}_{\text{in}})\}$, A pre-trained BERT model, the size of selected subset M'_{in}

Output: Selected monolingual subset M'_{in}

for $i = 1 \rightarrow |D_{dev}|$ **do**
  Generate semantic representations $[h_1, \cdots, h_n] = \text{BERT}(\bar{\mathbf{y}}_i)$
  Compute the sentence embedding $s_i = \frac{1}{n}\sum_{i=1}^{n} h_i$
end for
Compute the domain center: $C = \frac{1}{|D_{\text{dev}}|}\sum_{i=1}^{|D_{\text{dev}}|} s_i$
for $j = 1 \rightarrow |M_{\text{in}}|$ **do**
  Generate semantic representations $[h_1, \cdots, h_n] = \text{BERT}(\mathbf{y}_j)$
  Compute the sentence embedding $t_j = \frac{1}{n}\sum_{i=j}^{n} h_i$
  Compute the domain similarity $d_j = cosine(t_j, C)$
end for
Sort $\{d1, d2, \cdots, d_{|M_{\text{in}}|}\}$ in a descent order
Select top $|M'_{in}|$ sentences according to the rank
return M'_{in}