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

[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, chenweibbj8871, wangyanfeng}@sogou-inc.com

Background & Motivation

- Domain Adaptation
  - Domain Adaptation aim at adapting a general domain NMT model to a target domain.
  - Current research of domain adaptation usually considers very broad target domains.
  - The words or sentences in different sub-domains have different language phenomena.

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 in different sub-domains of IT.

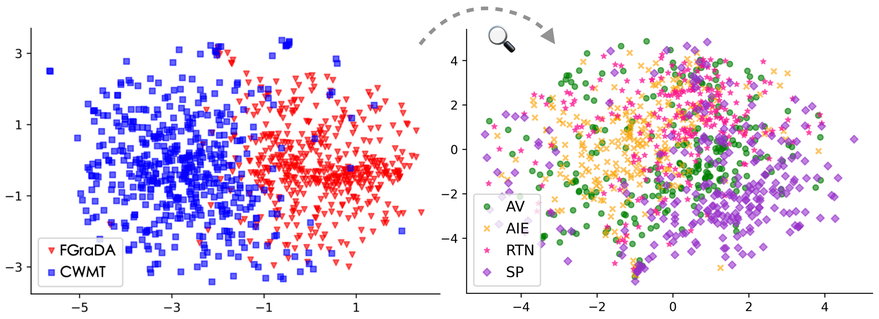

Figure 1: Visualization of sentences from general domain and IT fine-grained domains.

- Fine-grained Domain Adaptation
  - There are limited time and budget to collect data (especially parallel data).
  - Speicific research may be needed to model fine-grained domains with other heterogeneous resources that are more availble.

Our Contribution

- We build a fine-grained domain adaptation dataset for machine translation, FGraDA, to motivate wider investigation in such scenario.
- We compare different existing domain adaptation approaches and benchmark the FGraDA dataset.
- We present in-depth analyses showing that there are still challenging problems to further improve the performance with heterogeneous resources.

FGraDA Dataset

- Dataset Overview
  - Adaptation resource: bilingual dictionary and wiki knowledge base
  - Evalutation resource: development and test set

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.

- Bilingual Dictionary
  - It is easier and cheaper to obtain.
  - It contains domain-specific word-level correspondences between the two langauges.
  - We manually build a small set of bilingual dictionaries.

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.

- Wiki knowledge base
  - It is publicly available resource.
  - It contain rich monolingual resources and have additional structural knowledge.
  - We collect domain related English wikipages with the help of link relations.

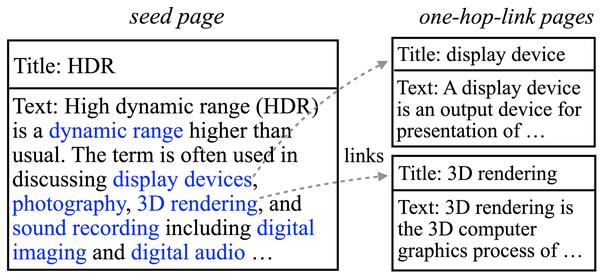

Figure 2: Illustration of the wiki knowledge base provided in our dataset.

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.

- Development and test set
  - We collect and label parallel data as development and test set.

Benchmarks

- Baselines
  - Base: directly using a general domain Transformer
  - Dict_{GBS}: Performing constrained decoding for Base with in-domain dictionary
  - Dict_{FT}: fine-tuning Base on the in-domain dictionary
  - Wiki_{BT}: using sentences of wiki seed pages for back-translation and fine-tune Base on it
  - Wiki_{BT}+Dict_{GBS}: Applying constrained decoding on Wiki_{BT}
- Bechmark Results
  - Dict_{GBS} and Wiki_{BT} improve the baseline to some extent.
  - Dict_{FT} barely brings any improvement.
  - With both resources, Wiki_{BT}+Dict_{FT} achieves the best performance. However, the translation performance of it on a large portion of test sentences is not satisfactory, e.g., under 20.

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

Table 5: Translation results (BLEU scores) on four fine-grained domains.

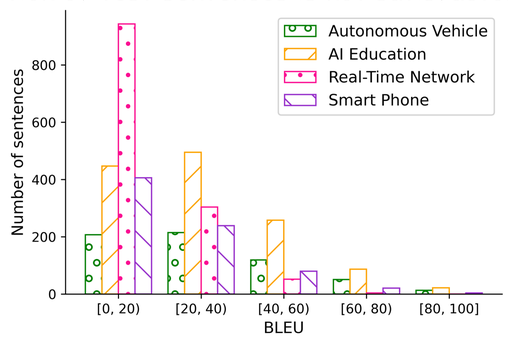

Figure 3:Distribution of sentence BLEU scores on four fine-grained domains.

Remaining Challenges

- Mining from the Dictionary
  - The domain dictionary contains accurate translation knowledge about the domain specific words
  - However, a large portion of dictionary items are still mis-translated.
  - Simply forcing the models to generate infrequent in-domain words is not sufficient.
- Mining from Wiki knowledge Base
  - Wiki knowledge base contain rich structural knowledge that may help the NMT model to "understand" domain specific words.
  - The first sentence in the page is usually the definition for title word.
  - Words that have link pages are closely related to the current title word.
- Mining from the Domain Hierarchy
  - Leveraing resources from other related sub-domains.

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 6: The translation accuracy (%) of items in the dictionary.

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 7: The performance on all four test sets.

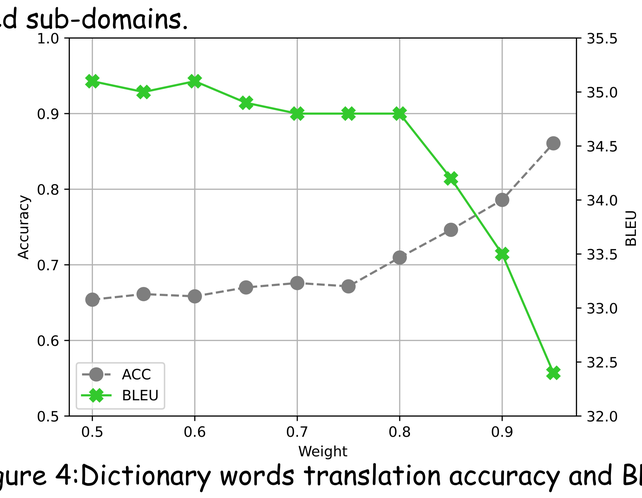

Figure 4:Dictionary words translation accuracy and BLEU w.r.t. different weights in GBS on AV test set.