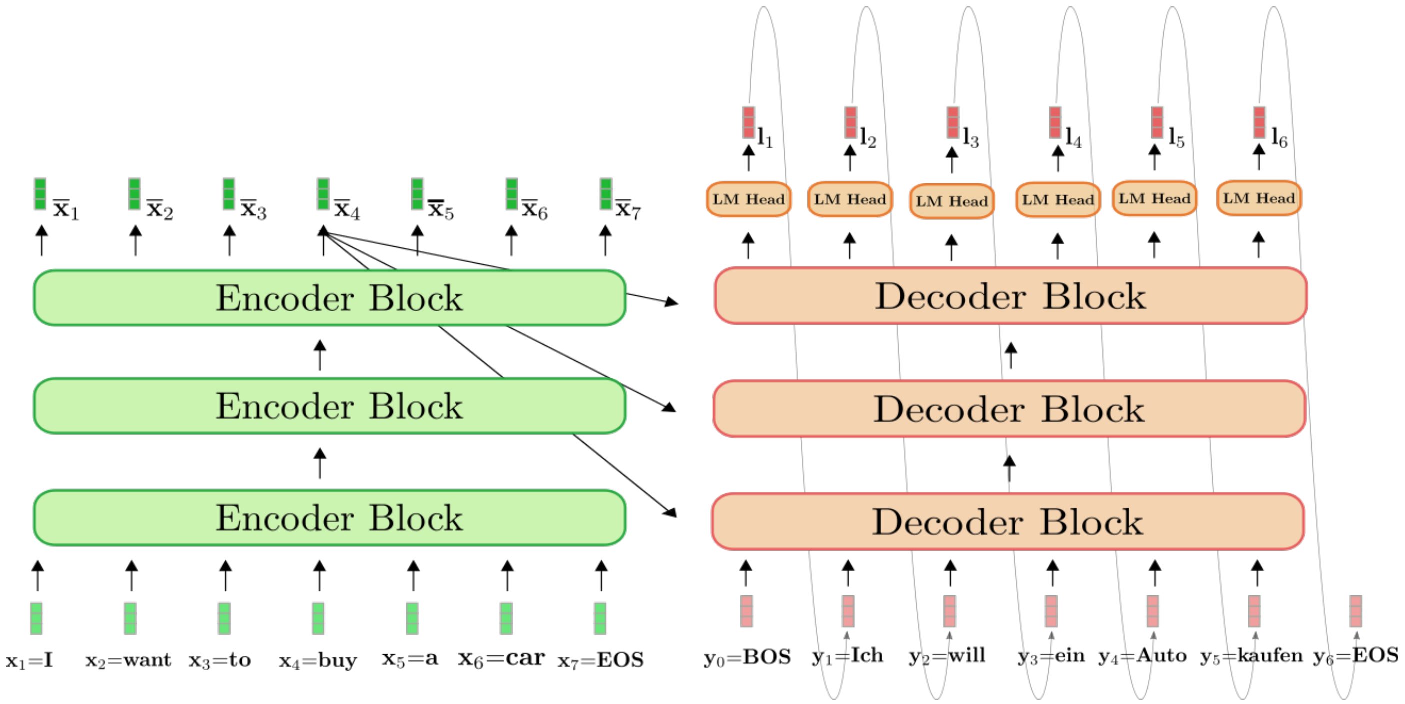
FGraDA: A Dataset and Benchmark for Fine-Grained Domain Adaptation in Machine Translation

Wenhao Zhu[1], Shujian Huang[1], Tong Pu[1], Pingxuan Huang[2], Xu Zhang[3], Jian Yu[3], Wei Chen[3], Yanfeng Wang[3], Jiajun Chen[1]

Content

Neural Machine Translation

- Neural machine translation (NMT) systems 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.

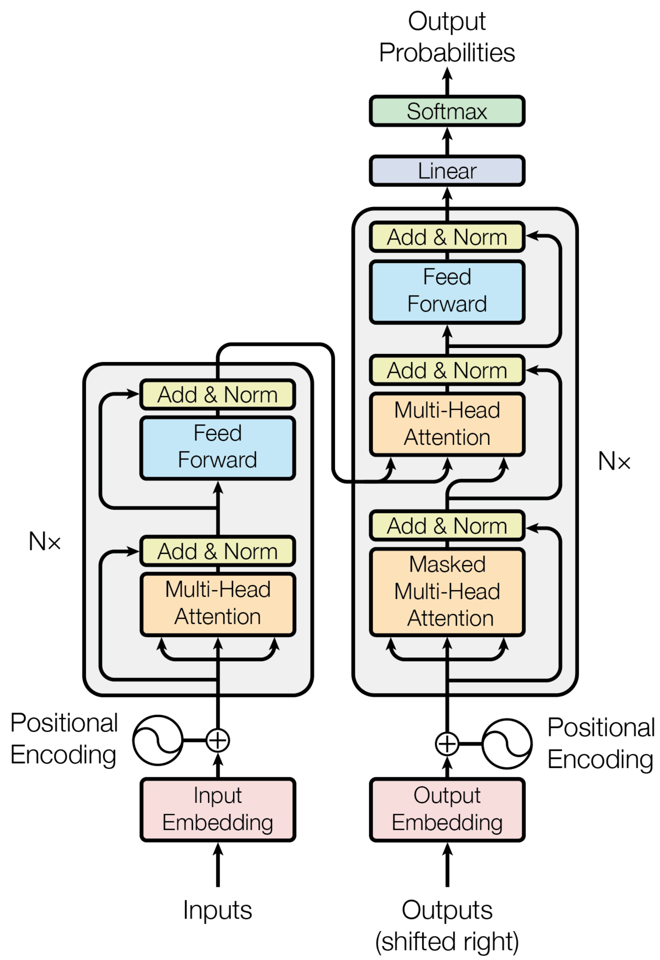


Translation probability distribution:

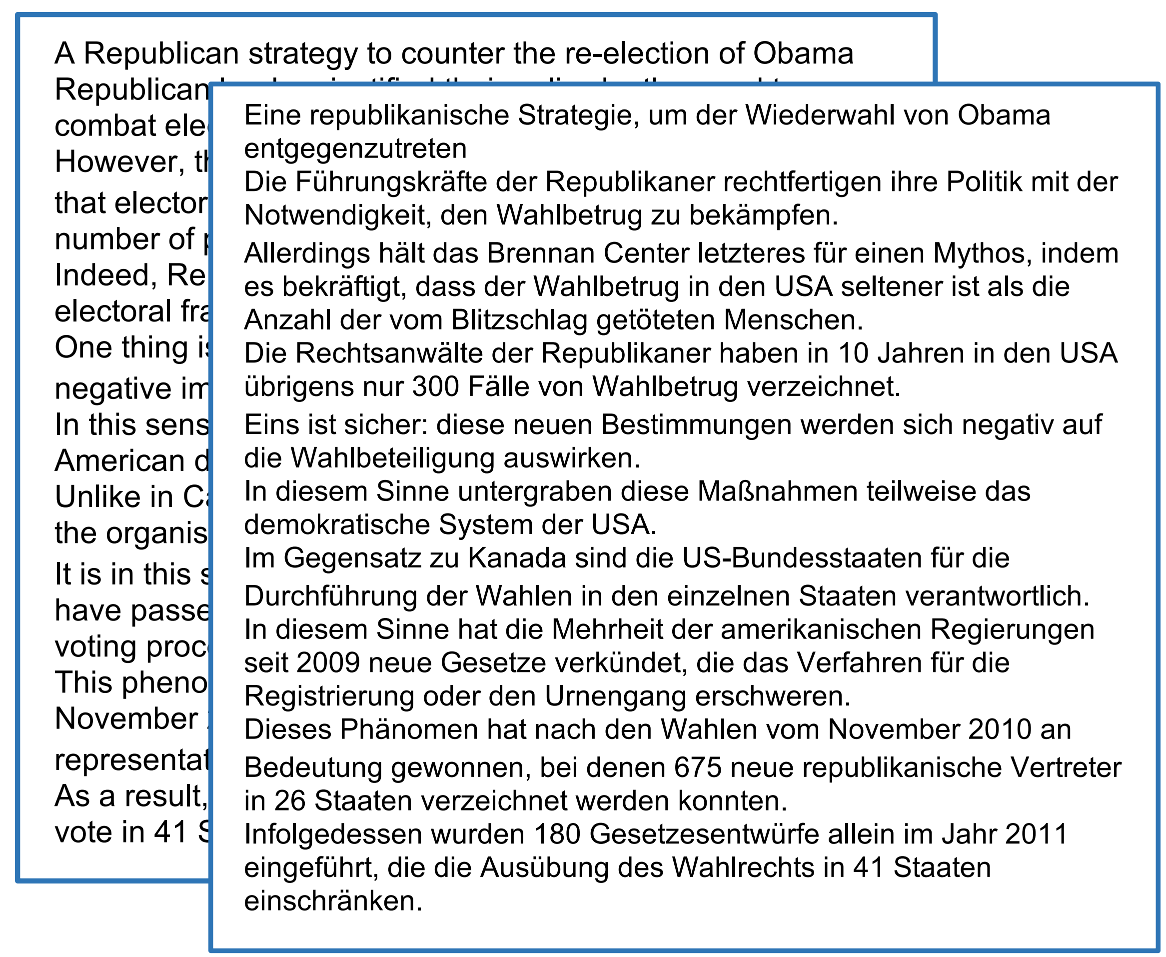
$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{|\mathbf{y}|} p(y_i|\mathbf{x}, \mathbf{y}_{<i}; \theta),$$

Neural Machine Translation (Cont.)

- Recent years have witnessed the great thrive in NMT, e.g., Transformer.
  - A general domain NMT model can be trained on large-scale general domain parallel data $D_g = \{(\mathbf{x}_g, \mathbf{y}_g)\}$.



Transformer(Vaswani et al. 2017)


A Republican strategy to counter the re-election of Obama

Eine republikanische Strategie, um der Wiederwahl von Obama entgegenzutreten
Die Führungskräfte der Republikaner rechtfertigen ihre Politik mit der Notwendigkeit, den Wahlbetrug zu bekämpfen.
Allerdings hält das Brennan Center letzteres für einen Mythos, indem es bekräftigt, dass der Wahlbetrug in den USA seltener ist als die Anzahl der vom Blitzschlag getöteten Menschen.
Die Rechtsanwälte der Republikaner haben in 10 Jahren in den USA übrigens nur 300 Fälle von Wahlbetrug verzeichnet.
Eins ist sicher: diese neuen Bestimmungen werden sich negativ auf die Wahlbeteiligung auswirken.
In diesem Sinne untergraben diese Maßnahmen teilweise das demokratische System der USA.
Im Gegensatz zu Kanada sind die US-Bundesstaaten für die Durchführung der Wahlen in den einzelnen Staaten verantwortlich.
In diesem Sinne hat die Mehrheit der amerikanischen Regierungen seit 2009 neue Gesetze verkündet, die das Verfahren für die Registrierung oder den Urnengang erschweren.
Dieses Phänomen hat nach den Wahlen vom November 2010 an Bedeutung gewonnen, bei denen 675 neue republikanische Vertreter in 26 Staaten verzeichnet werden konnten.
Infolgedessen wurden 180 Gesetzesentwürfe allein im Jahr 2011 eingeführt, die die Ausübung des Wahlrechts in 41 Staaten einschränken.

Parallel Data (En-De)

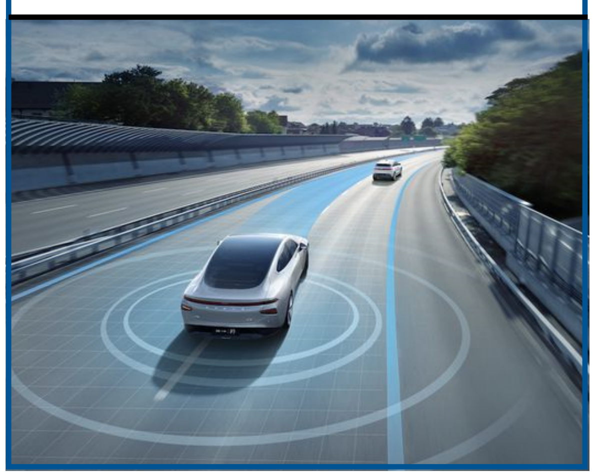
Domain Adaptation

- Domain Adaptation aims at adapting a general model NMT model to a target domain.

$$\boxed{p_\theta(y_g \mid x_g) \rightarrow p_\theta(y_{in} \mid x_{in})}$$

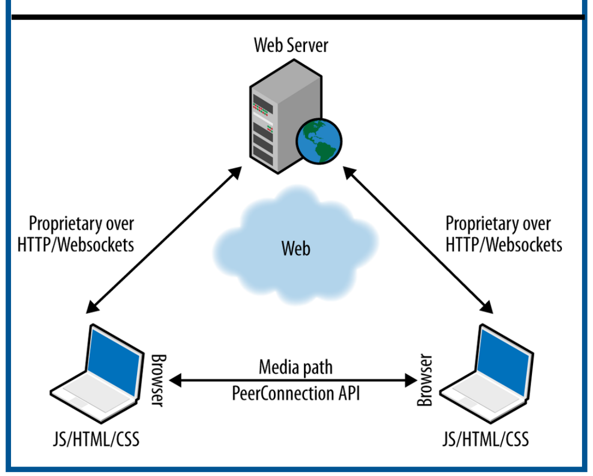
  - Current research of domain adaptation usually considers very broad target domains, e.g., medical, law, IT, subtitles.
  - We suggest that there are fine-grained sub-domains within coarse domains.

Autonomous Vehicles



AI Education

Real-Time Networks



Smart Phone

IT sub-domains

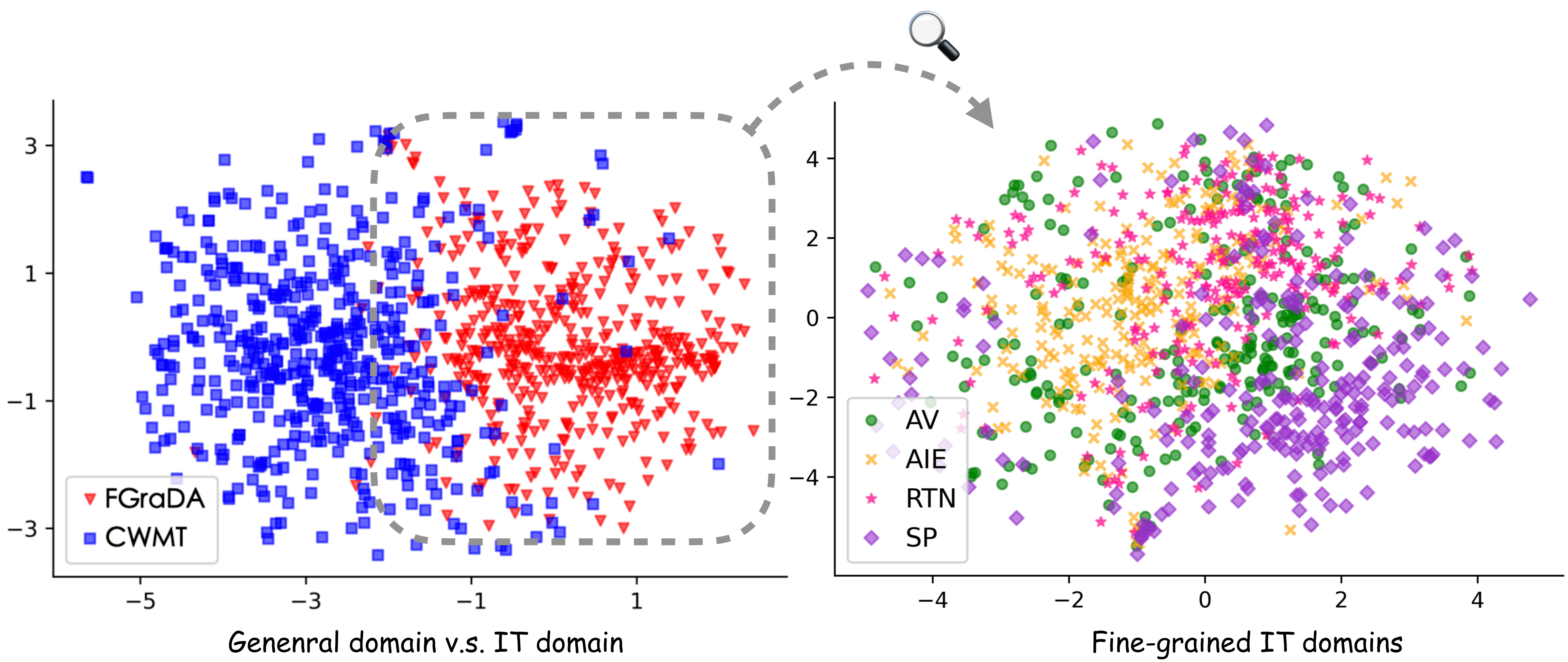
Diversities within Coarse Domains

- The words or sentences in different sub-domains may have different language phenomena.
  - Word-level: The same Chinese word "卡" corresponds to different English translations in different fine-grained IT domains.

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 ...

Diversities within Coarse Domains (Cont.)

- The words or sentences in different fine-grained domains may have different language phenomena.
  - Sentence-level: The four fine-grained IT domains have overlaps but present a unique distribution.



Low-resource Scenario

- Adapting to fine-grained domains often faces challenges as a low-resource scenario.
  - There 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 that are more available.

Challenge: Modeling Fine-grained Domains

- Modeling fine-grained domains with heterogeneous resources is the key challenge for fine-grained domain adaptation.
  - When the target domain is insufficiently modeled, various translation errors will happen.

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.

Our Contribution

- We build a fine-grained domain adaptation dataset for machine translation, FGraDA, to motivate wider investigation in such a 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.

Please note that we are here only presenting and benchmarking this task and calling for attention and solution.

Overview of FGraDA Dataset

- We select four real-world conferecens as representatives to construct the dataset. Each conference is organized for a particular topic of IT, which could be seen as four fine-grained IT domains.
  - Global AI and Robotics Conference (CCF-GAIR2019) -> Autonoumous Vehicles
  - GIIS China Education Industry Innovation Summit (GIIS2019) -> AI Education
  - Real-Time Internet Conference (RTC2019) -> Real-Time Networks
  - Apple-events (held in 2018 and 2019) -> Smart Phone

Overview of FGraDA Dataset (Cont.)

- FGraDA dataset

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

  - Adaptation resource: heterogeneous but more available resources: bilingual dictionaries and wiki knowledge base, which contain rich domain information.
  - Evaluation resource: development and test set.

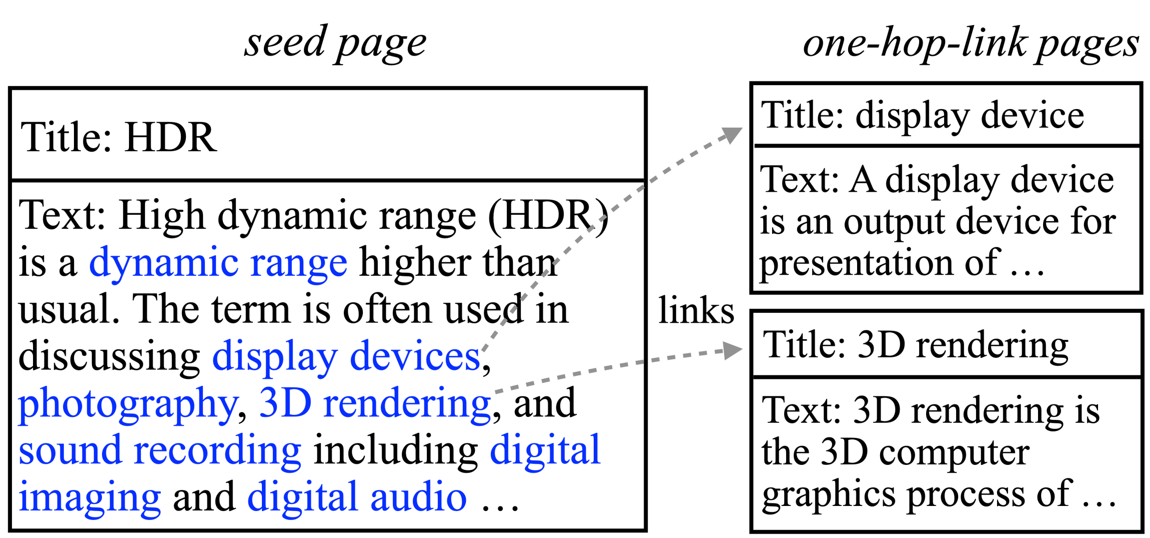
Bilingual Dictionary

- Bilingual dictionary
  - is much easier or cheaper to obtain than parallel data.
  - contains domain-specific word-level correspondences between the two languages.
- We manually build a small set of bilingual dictionaries and ask the linguistic experts to check them.

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

Wiki Knowledge Base

- Wiki knowledge base
  - is a publicly available resource.
  - not only contains rich monolingual resources, but also have additional structural knowledge, e.g., link relations.
- We collect English wikipages containing annotated dictionary keywords in their titiles (seed pages) and wikipages directly linked by links in the seed pages (one-hop-link pages).



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

Development and Test Set

- We collect and label parallel data as development and test set:

1. collect 70 hours of audio recordings from the four conferences mentioned above.
2. transcript the audio recordings with the in-house tools
3. filter out domain-irrelevant sentences
4. annotate them into 4,767 parallel parallel pairs
5. conduct data desensitization to hide human names and company names to protect privacy.
6. split annotated data in each domain into two parts: 200 sentence pairs as the development set, and the rest as the test set.

Existining Domain Adaptation Approaches

- Using (pseudo) parallel data
  - Luong and Manning (2015) fine-tune a general domain NMT model on target domain data parallel data

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

  - Sennrich et al. (2016) propose back-translation to construct pseudo parallel data with target language monolingual data.

$$\boxed{\hat{D}_{in} = \{(\hat{x}_{in}, y_{in})\}, \hat{x}_{in} \sim p_{\phi}(x_g \,|\, y_g)}$$

Existing Domain Adaptation Approaches (Cont.)

- Using dictionaries
  - Hokamp and Liu (2017) propose grid beam search (GBS) to incorporate subsequences into the NMT model's output.
  - Kothur et al. (2018) treat bilingual dictionaries as pseudo bitext and fine-tune the NMT model on them.
  - Hu et al. (2019) use large-scale bilingual dictionaries for word-by-word translation to generate pseudo bitext.

Existing Domain Adaptation Approaches (Cont.)

- Using knowledge base
  - Bilingual knowledge base could be used for extracting bilingual lexicons (Zhao et al., 2020)
  - To our best knowledge, there is no previous attempts in exploring domain related information in monolingual wikipages.

Benchmark Systems

- We implement the following systems as benchmark baselines
  - Base: directly using a general domain Transformer
  - $\mathrm{Dict}_{\mathrm{GBS}}$: performing constrain decoding for Base with in-domain dictionary
  - $\mathrm{Dict}_{\mathrm{FT}}$: fine-tuning Base on the in-domain dictionary
  - $\mathrm{Wiki}_{\mathrm{BT}}$: using sentences of wiki seed pages for back-translation and fine-tune Base on it
  - $\mathrm{Wiki}_{\mathrm{BT}}$+$\mathrm{Dict}_{\mathrm{GBS}}$: Applying constrained decoding on $\mathrm{Wiki}_{\mathrm{BT}}$

Experiment Settings

- Data
  - General Domain: WMT-CWMT-17 Chinese-English Dataset
  - Target Domain: FGraDA Chinese-English Dataset
- Model
  - Transformer
- Evaluation Metric
  - BLEU (Papineni et al., 2002)

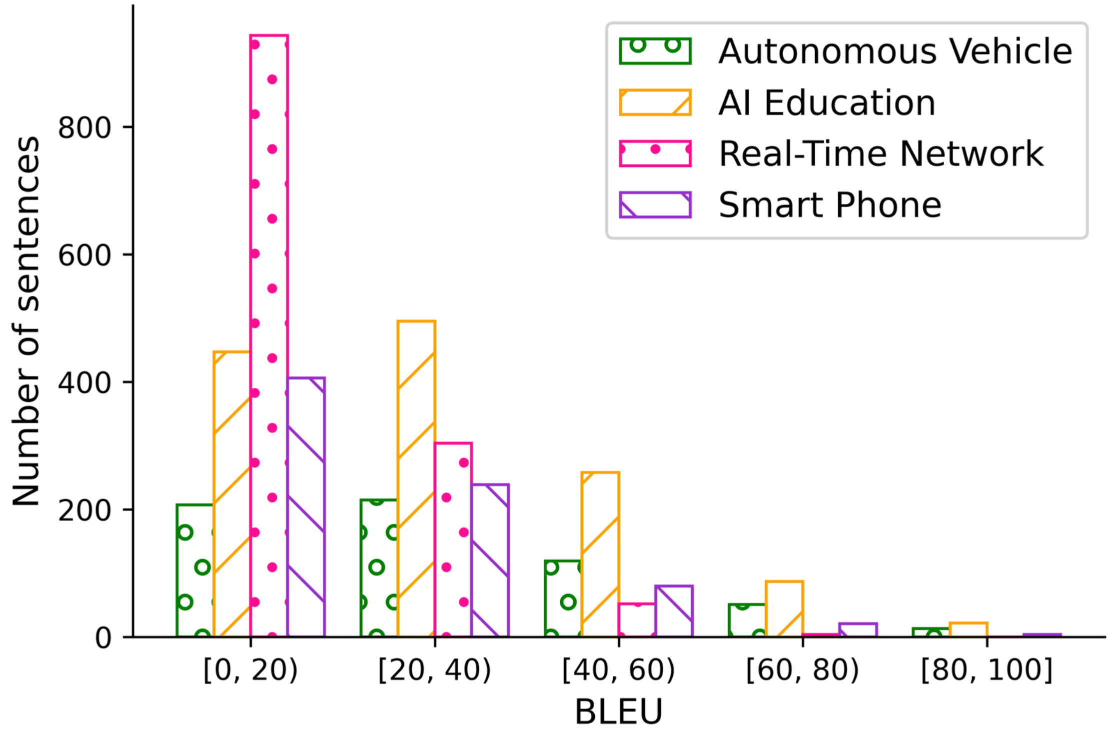
Benchmark Results

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

- $Dict_{GBS}$ and $Wiki_{BT}$ improve the baseline model to some extent
- $Dict_{FT}$ barely brings any improvement.
- With both resources, $Wiki_{BT}$+$Dict_{FT}$ achieves the best performance.

However, the translation quality does not improve as greatly as reported in other research; the performance on RTN and SP are much lower than other two domains

Benchmark Results (Cont.)



- The translation performance of Wiki_{BT}+Dict_{GBS} on a large portion of test sentences is not satisfactory, e.g., under 20, leaving a large room for improvement.

Mining from the Dictionary

- The domain dictionary contains accurate translation knowledge about the domain specific words.

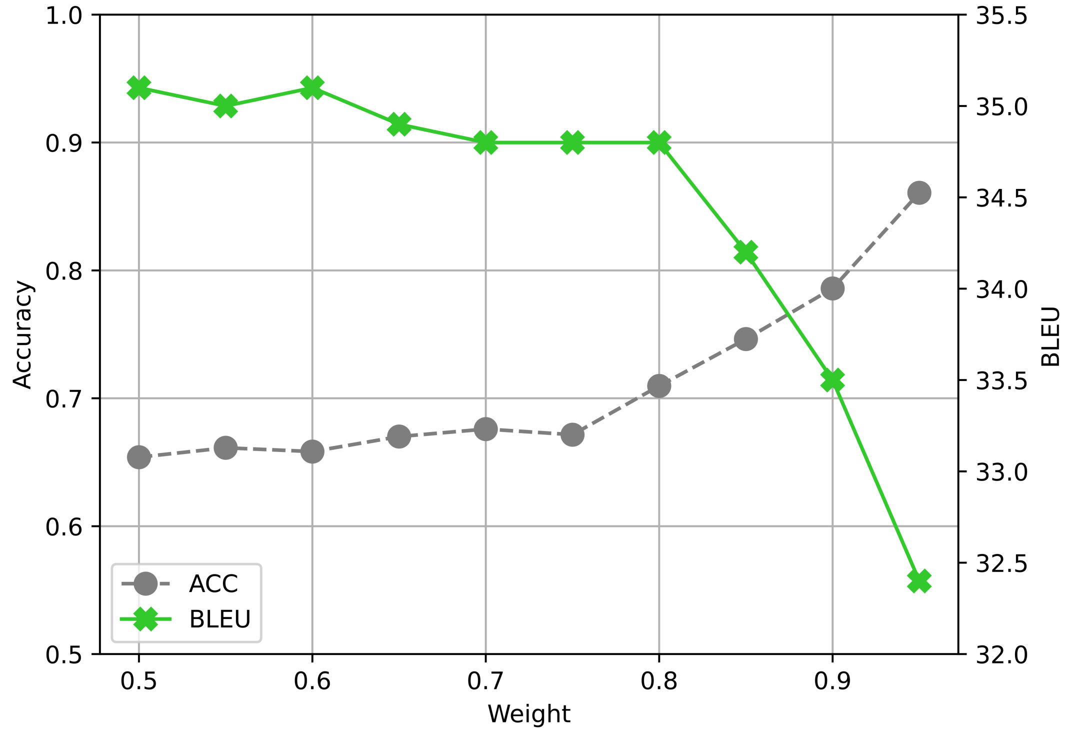
Model	AV	AIE	RTN	SP
Base	63.04	57.81	65.86	59.42
$\mathbf{Dict_{GBS}}$	**65.84**	59.69	76.94	61.85
$\mathbf{Wiki_{BT}}$	63.93	59.38	67.30	58.97
$\mathbf{Wiki_{BT}+Dict_{GBS}}$	**65.84**	**64.22**	**87.84**	**63.07**

- However, a large portion of dictionary items are still mis-translated.

Mining from the Dictionary (Cont.)

Re-weighted log-likelihood in GBS:

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

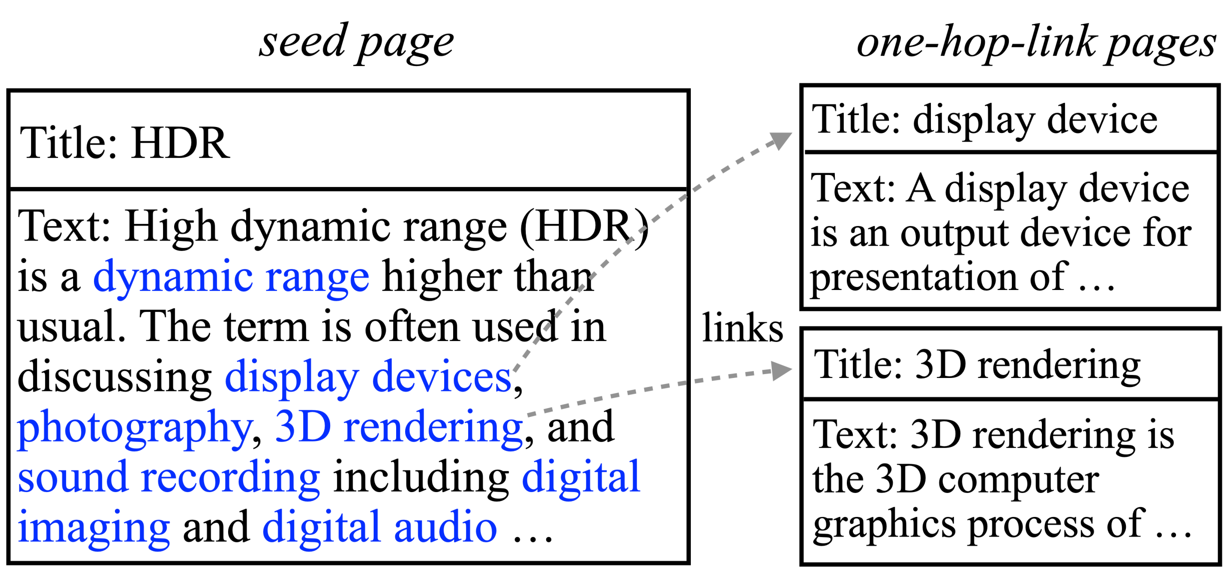


- In grid beam search, higher weight w ensures more domain specific words are translated, but BLEU score drops significantly.

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

- Leveraging resources from other related sub-domains for adaptation might be beneficial.
  - There is a close relation between these sub-domains.

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**

Conclusion

- We introduce FGraDA and benchmark the dataset.
- We find that current adaptation approaches are not satisfactroy.
- We suggest that provided hetergeneous resources may contain useful information for the adaptation and encourages further exploration of modeling fine-grained domains with these resources.

Paper: https://owennju.github.io/archieve/LREC2022_paper.pdf

Dataset: https://github.com/OwenNJU/FGraDA

Please feel free to contact me (zhuwh@smail.nju.edu.cn) if you have any questions.