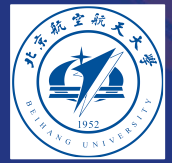
国际人工智能会议
AAAI 2021 论文北京预讲会


Revisiting Iterative Back-Translation from the Perspective of Compositional Generalization

郭一诺 (gyn0806@pku.edu.cn)

Our Team - Main Contributors

Peking University
Yinuo Guo

Beihang University
Hualei Zhu

Microsoft Asia
Zeqi Lin

Microsoft Asia
Bei Chen

Microsoft Asia
Jian-Guang

Microsoft Asia
Dongmei Zhang

Compositional Generalization

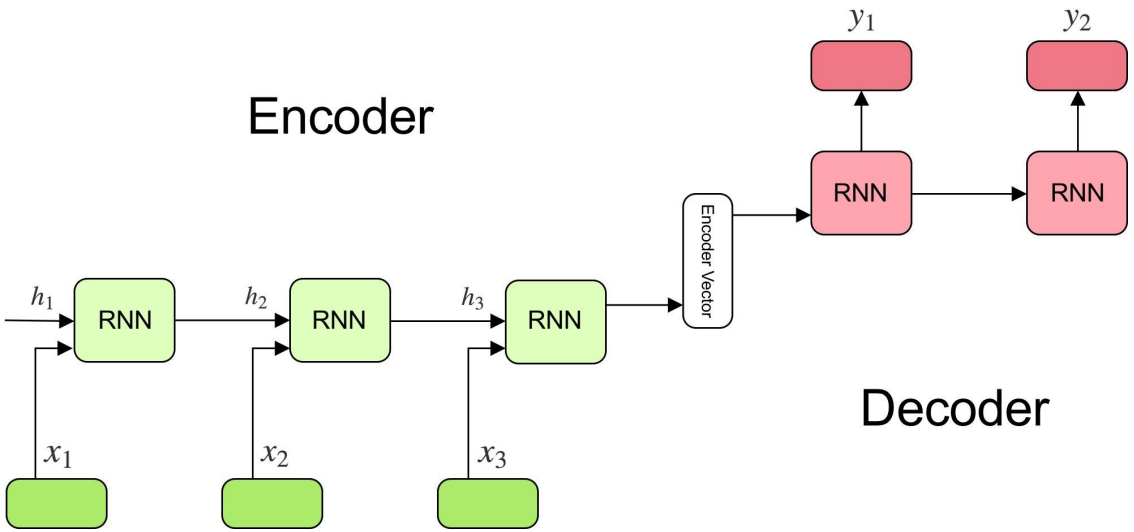
- The algebraic ability to understand and produce unseen combinations of seen atoms.

Infinite use of finite means. —— Chomsky

	Natural Language		Programming Language
Train	run twice	⟹	RUN RUN
	jump and walk	⟹	JUMP WALK
Test	jump twice and run	⟹	JUMP JUMP RUN

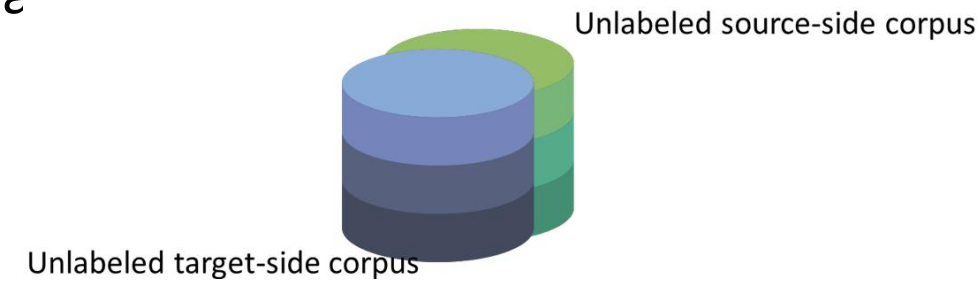
Background: Seq2seq Tasks in NLP

- Machine Translation
- Semantic Parsing
- Summarization
- …



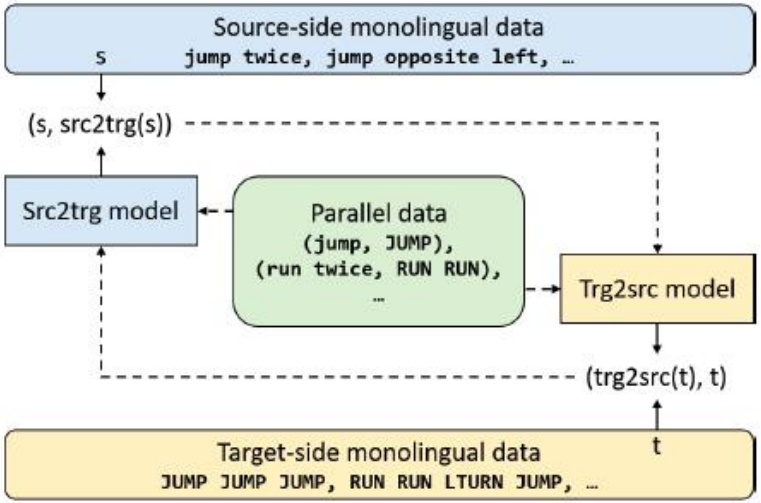
Semi-Supervised Learning

- Parallel data are limited and expensive
- Monolingual data are cheap and abundant, containing lots of unseen combinations
- Hypothesis: semi-supervised learning can enable models understand and produce much more combinations beyond labelled data, thus tackling the bottleneck of lacking compositional generalization



Iterative Back-Translation

- We focus on Iterative Back-Translation (IBT), a simple yet effective semi-supervised method that has been successfully applied in machine translation.



Three Research Questions

- RQ1:How does IBT affect compositional generalization of seq2seq models?
  - Yes
- RQ2: What is the key that contributes to the success of IBT?
  - Quality of pseudo parallel data & Perturbations
- RQ3: How to further improve the performance of IBT?
  - Curriculum Iterative Back-translation

Evaluate on CFQ & SCAN

- Substantially improves the performance on CG benchmarks.
- Better monolingual data, better results.

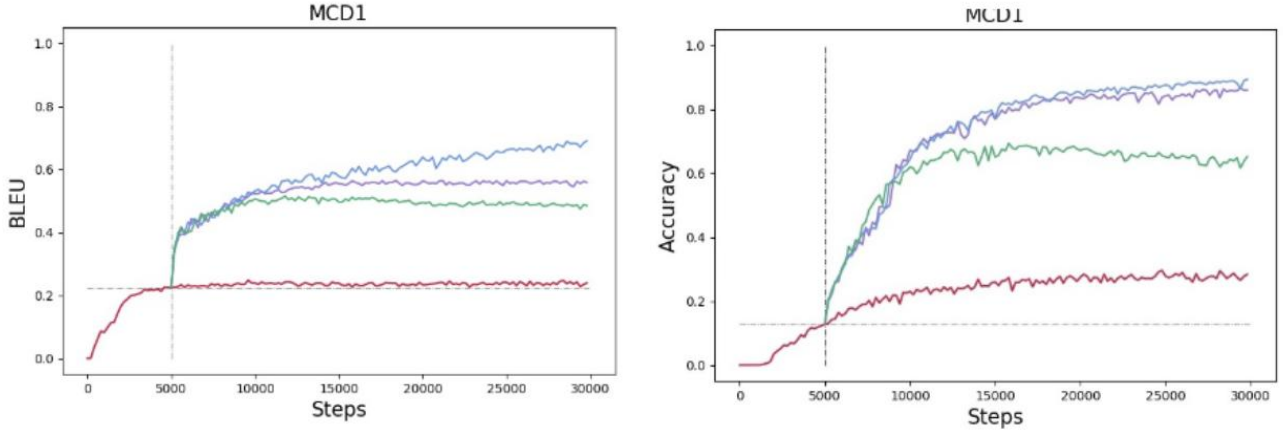
Which Swedish founder of [M0] produced [M2] ?

```
SELECT DISTINCT ?x0 WHERE {
          ?x0 ns:film.producer.film|ns:film.production_company.films [M2] .
          ?x0 ns:organization.organization_founder.organizations_founded [M0] .
          ?x0 ns:people.person.nationality ns:m.0d0vqn
}
```

Models	MCD1	MCD2	MCD3
LSTM+Attn	28.9 ± 1.8	5.0 ± 0.8	10.8 ± 0.6
Transformer	34.9 ± 1.1	8.2 ± 0.3	10.6 ± 1.1
Uni-Transformer	37.4 ± 2.2	8.1 ± 1.6	11.3 ± 0.3
CGPS	13.2 ± 3.9	1.6 ± 0.8	6.6 ± 0.6
T5-11B	61.4 ± 4.8	30.1 ± 2.2	31.2 ± 5.7
GRU+Attn (Ours)	32.6 ± 0.22	6.0 ± 0.25	9.5 ± 0.25
+mono30	$\mathbf{64.8 \pm 4.4}$	$\mathbf{57.8 \pm 4.9}$	$\mathbf{64.6 \pm 4.9}$
+mono100	83.2 ± 3.1	71.5 ± 6.9	81.3 ± 1.6
+transductive	88.4 ± 0.7	81.6 ± 6.5	88.2 ± 2.2

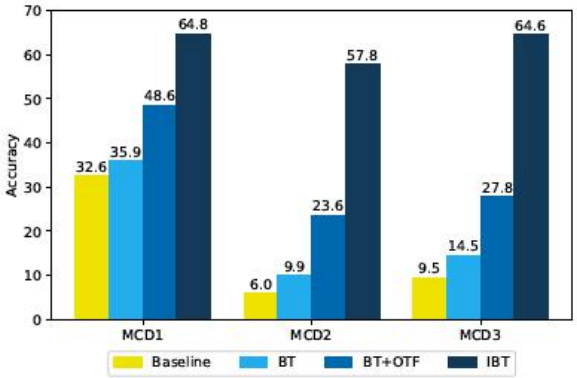
Quality of Pseudo Parallel Data

- Iterative back-translation can increasingly correct errors in pseudo-parallel data

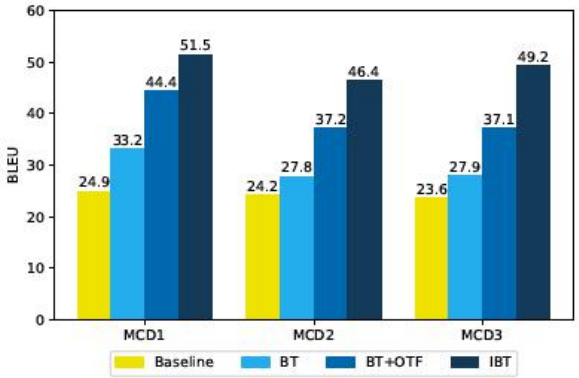


Impact of Error-Prone Data & Perturbations

- Even noise pseudo-parallel data can bring gains!
  - As they bring implicit knowledge of unseen combinations
- Perturbations brought by OTF (on-the-fly) is very important!
  - Pseudo-parallel data are generated dynamically, which prevent learning specific incorrect bias



(a) Accuracy of Src2trg models



(b) BLEU of Trg2src models

Curriculum Iterative Back-Translation

- We want to help reduce errors more efficiently
- CIBT: during the training process:
  - start out with easy monolingual data,
  - then gradually increase the difficulty.

Curriculum Iterative Back-Translation

- Curriculum learning benefits iterative back-translation.
- Curriculum learning is more beneficial to difficult data than simple data.

Table 3: Performance (accuracy) of curriculum iterative back-translation.

	IBT	CIBT with hyperparameter c (steps in each stage)				
		2000	2500	3000	3500	4000
MCD1	64.8 ± 4.4	66.1 ± 5.0	66.0 ± 4.8	$\mathbf{66.6} \pm 5.4$	65.9 ± 3.7	65.4 ± 3.8
MCD2	57.8 ± 4.9	68.6 ± 2.6	$\mathbf{69.1} \pm 3.1$	68.0 ± 1.9	66.8 ± 2.4	65.4 ± 3.1
MCD3	64.6 ± 4.9	70.2 ± 4.9	68.4 ± 7.0	$\mathbf{70.4} \pm 4.8$	69.2 ± 4.1	67.0 ± 6.3
Mean	62.4 ± 6.1	$\mathbf{68.3} \pm 4.1$	67.8 ± 4.7	$\mathbf{68.3} \pm 4.1$	67.3 ± 3.4	65.9 ± 4.1

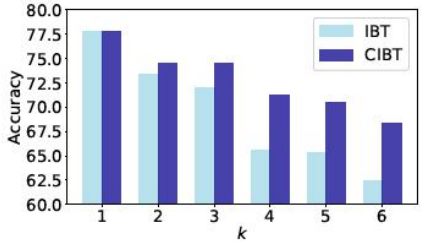


Figure 6: Performance on different subsets. This figure indicates that curriculum learning is more beneficial to difficult data (larger k) than simple data (smaller k).

Takeaways

- Iterative back-translation can significantly improve CG.
- Why IBT works well:
  - Unseen combinations
  - Increasingly improving the quality of pseudo-parallel data
  - Perturbations
- We propose curriculum iterative back-translation to further improving the performance.

国际人工智能会议
AAAI 2021 论文北京预讲会

THANKS

2020.12.19

Related papers from our team (MSRA DKI):

Hierarchical Poset Decoding for Compositional Generalization in Language (NeurIPS 2020)

Compositional Generalization by Learning Analytical Expressions (NeurIPS 2020 Spotlight)

Iterating Utterance Segmentation for Neural Semantic Parsing (AAAI 2021)