Cross-Lingual Machine Reading Comprehension

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu

Research Center for Social Computing and Information Retrieval (SCIR), Harbin Institute of Technology, China
Joint Laboratory of HIT and iFLYTEK Research (HFL), Beijing, China

Nov 5, 2019

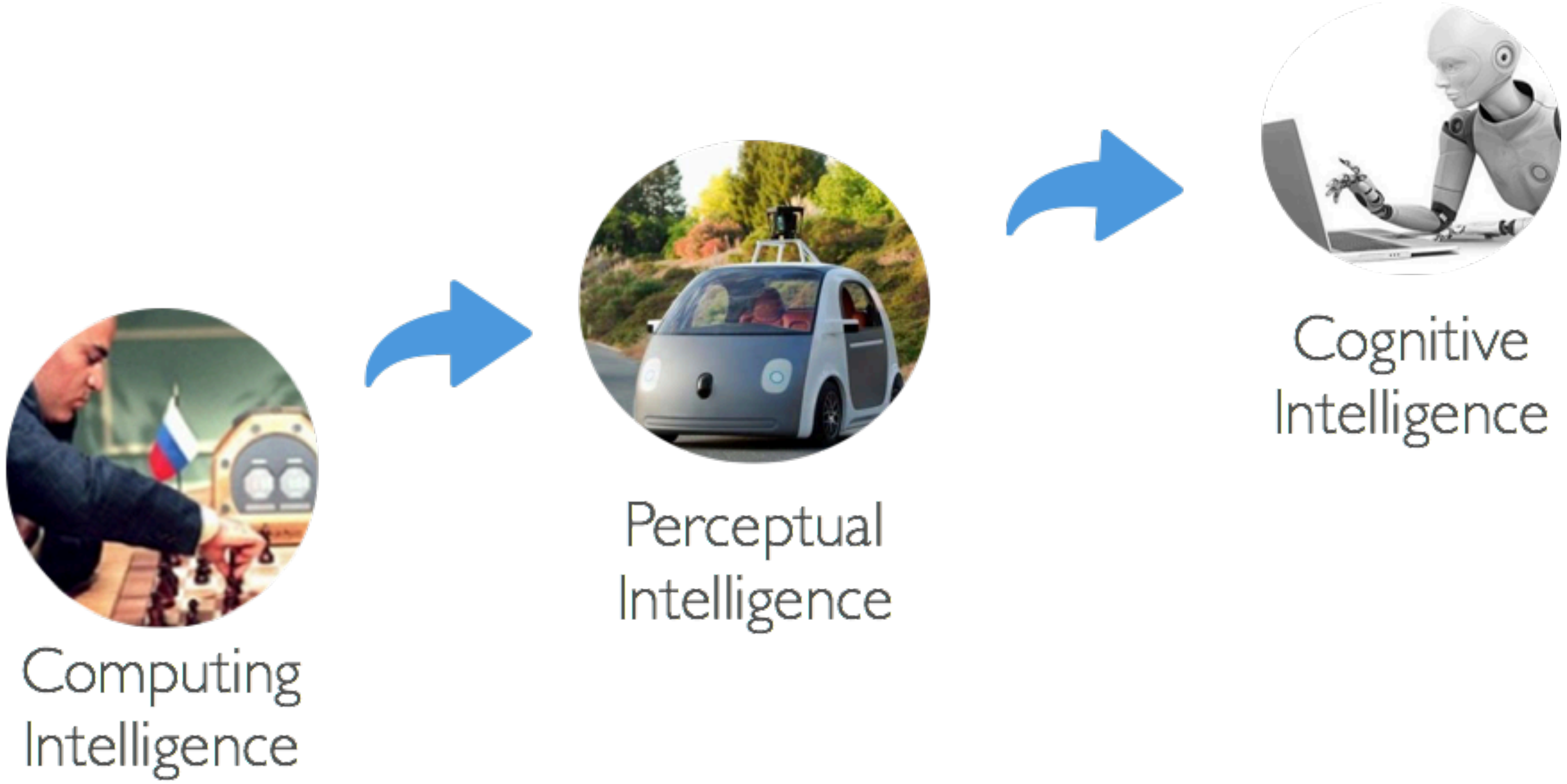
EMNLP-IJCNLP 2019, Hong Kong SAR, China

OUTLINE

- Introduction
- Related Work
- Preliminaries
- Back-Translation Approaches
- Dual BERT
- Experiments
- Discussion
- Conclusion & Future Work

INTRODUCTION

- To comprehend human language is essential in AI
- **M**achine **R**eading **C**omprehension (MRC) has been a trending topic in recent NLP research



INTRODUCTION

- **Machine Reading Comprehension (MRC)**
  - To read and comprehend a given article and answer the questions based on it
- **Type of MRC**
  - Cloze-style: CNN / Daily Mail (Hermann et al., 2015), CBT (Hill et al., 2015)
  - Span-extraction: SQuAD (Rajpurkar et al., 2016)
  - Choice-selection: MCTest (Richardson et al., 2013), RACE (Lai et al., 2017)
  - Conversational: CoQA (Reddy et al., 2018), QuAC (Choi et al., 2018)
  - …

INTRODUCTION

- **Problem: Most of the MRC research is mainly for English**
  - Languages other than English are not well-addressed due to the lack of data

TriviaQA
… NaturalQuestions HotpotQA
…
NarrativeQA CNN / DailyMail
MultiRC CLOTH
SQuAD
DuoRC ARC
MCTest
… QuAC RACE
DROP
MS MARCO CBT DREAM
SCT
NewsQA CoQA …
SearchQA
RecipeQA

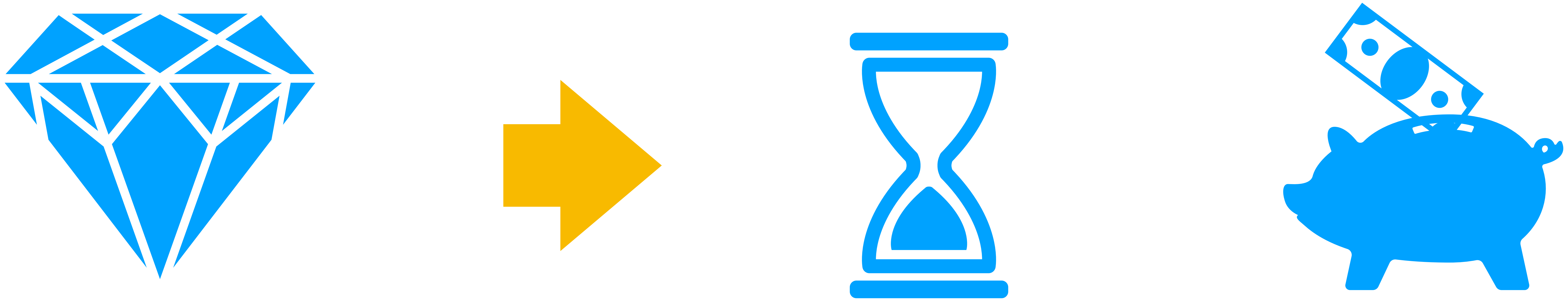
▲ English MRC Datasets

C^3 …
WebQA
PD&CFT CMRC 2018
DRCD
CJRC
DuReader
CMRC 2017
CMRC 2019
ChID
…

▲ Chinese MRC Datasets

INTRODUCTION

- **How to enrich the training data in low-resource language?**
    - Solution 1: Annotate by human experts



Introduction

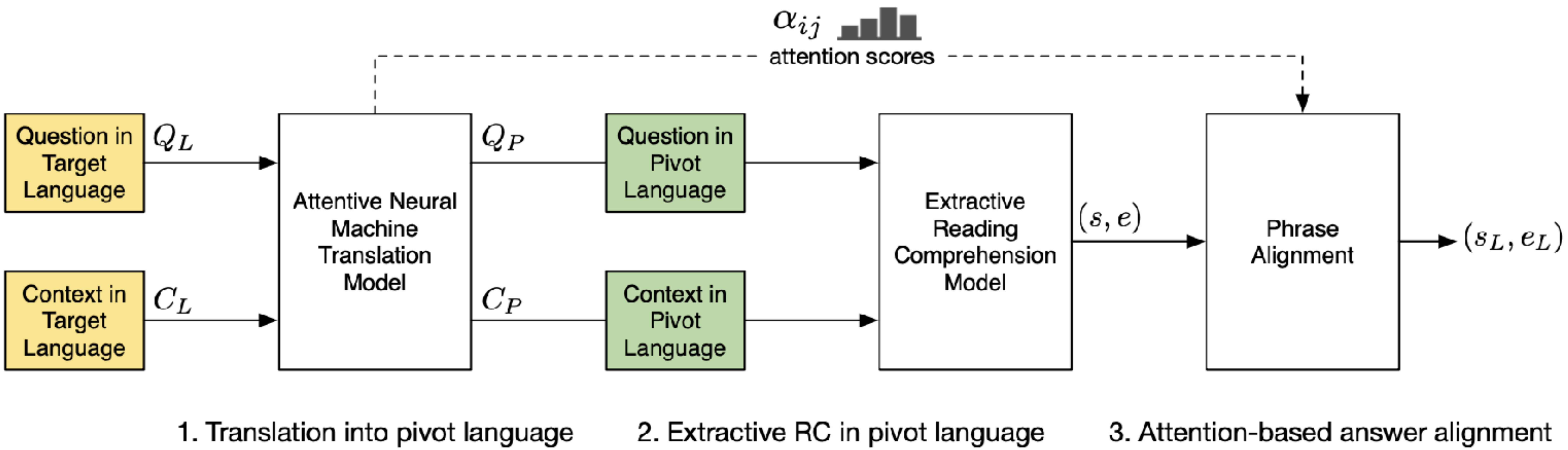
- **How to enrich the training data in low-resource language?**
  - Solution 2: Cross-lingual approaches
  - Multilingual representation, translation-based approaches, etc.

INTRODUCTION

- **Contributions**
  - We propose a new task called **Cross-Lingual Machine Reading Comprehension (CLMRC)** to address the MRC problems in low-resource language.
  - Several back-translation based approaches are presented for cross-lingual MRC and yield state-of-the-art performances on Chinese, Japanese, and French data.
  - Propose a novel model called **Dual BERT** to simultaneously model <Passage, Question> in both source and target language.
  - Dual BERT shows promising results on two public Chinese MRC datasets and set new state-of-the-art performances, indicating the potentials in CLMRC research.

Related Work

- Asai et al. (2018) propose to use runtime MT for multilingual MRC



Related Work

- **Contemporaneous Works (not in the paper)**
  - XQA: A Cross-lingual Open-domain Question Answering Dataset (Liu et al., ACL 2019)
    - Propose a cross-lingual QA dataset
  - Cross-Lingual Transfer Learning for Question Answering (Lee and Lee, arXiv 201907)
    - Propose transfer learning approaches for QA
  - Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model (Hsu et al., EMNLP 2019)
  - …

PRELIMINARIES

- **Task: Span-Extraction Machine Reading Comprehension**
- **SQuAD** (Rajpurkar et al., EMNLP 2016)
  - Passage: From Wikipedia pages, segment into several small paragraphs
  - Question: Human-annotated, including various query types (what/when/where/who/how/why, etc.)
  - Answer: Continuous segments (text spans) in the passage, which has a larger search space, and much harder to answer than cloze-style RC

SQuAD
The Stanford Question Answering Dataset

Oxygen
The Stanford Question Answering Dataset

In the meantime, on August 1, 1774, an experiment conducted by the British clergyman Joseph Priestley focused sunlight on mercuric oxide (HgO) inside a glass tube, which liberated a gas he named "dephlogisticated air". He noted that candles burned brighter in the gas and that a mouse was more active and lived longer while breathing it. After breathing the gas himself, he wrote: "The feeling of it to my lungs was not sensibly different from that of common air, but I fancied that my breast felt peculiarly light and easy for some time afterwards." Priestley published his findings in 1775 in a paper titled "An Account of Further Discoveries in Air" which was included in the second volume of his book titled Experiments and Observations on Different Kinds of Air. Because he published his findings first, Priestley is usually given priority in the discovery.

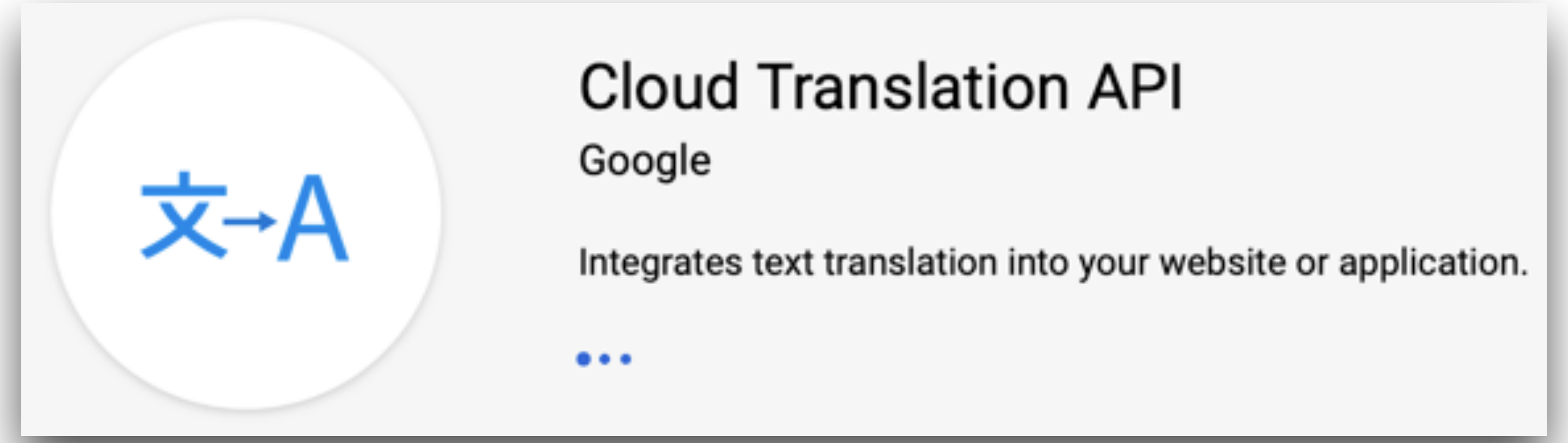
Why is Priestley usually given credit for being first to discover oxygen?
Ground Truth Answers: published his findings first | he published his findings first | he published his findings first | he published his findings first | Because he published his findings first

PRELIMINARIES

- **Terminology**
  - Source Language (S): for extracting knowledge
    - Rich-resourced, large-scale training data
    - For example, English.
  - Target Language (T): to optimize on
    - Low-resourced, limited or no training data
    - For example, Japanese, French, Chinese, etc.
- We aim to improve Chinese (target language) MRC using English (source language) resource

Back-Translation Approaches

- **Google Neural Machine Translation (GNMT)**
    - Easy API for translation, language detection, etc.
    - Results on NIST MT02~08 show state-of-the-art performances



	MT02	MT03	MT04	MT05	MT06	MT08	Average
$AST_{feature}$ (Cheng et al., 2018)	46.10	**44.07**	**45.61**	44.06	**44.44**	34.94	43.20
GNMT (March 25, 2019)	**46.26**	43.40	44.17	**44.14**	43.86	**37.61**	**43.24**

▲ GNMT performance on NIST MT 02~08 datasets

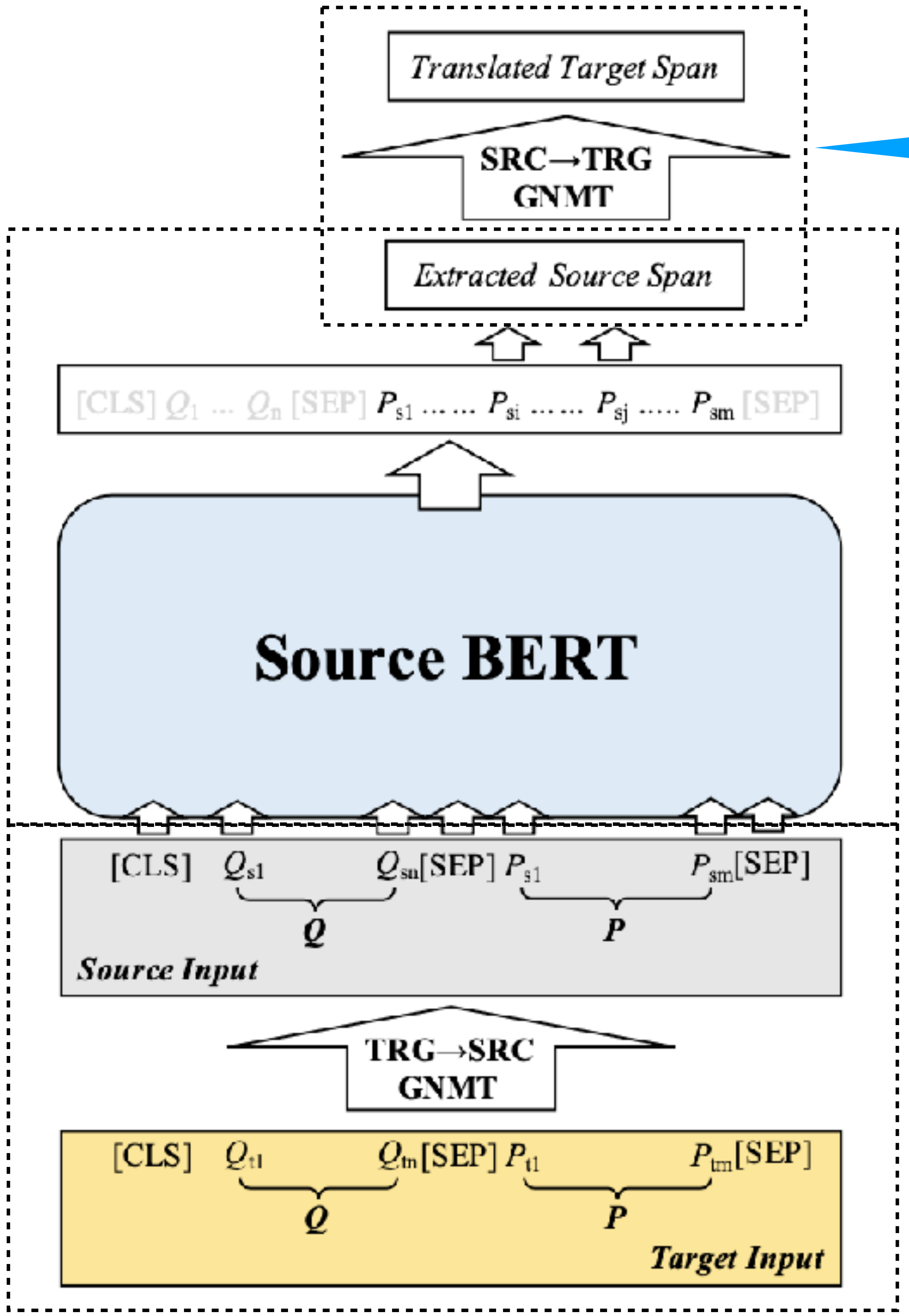
Back-Translation Approaches

- GNMT♠

Step3: Back-translate answer into target language

Step2: Answer the question using RC system in source language

Step1: Translate target sample into source language

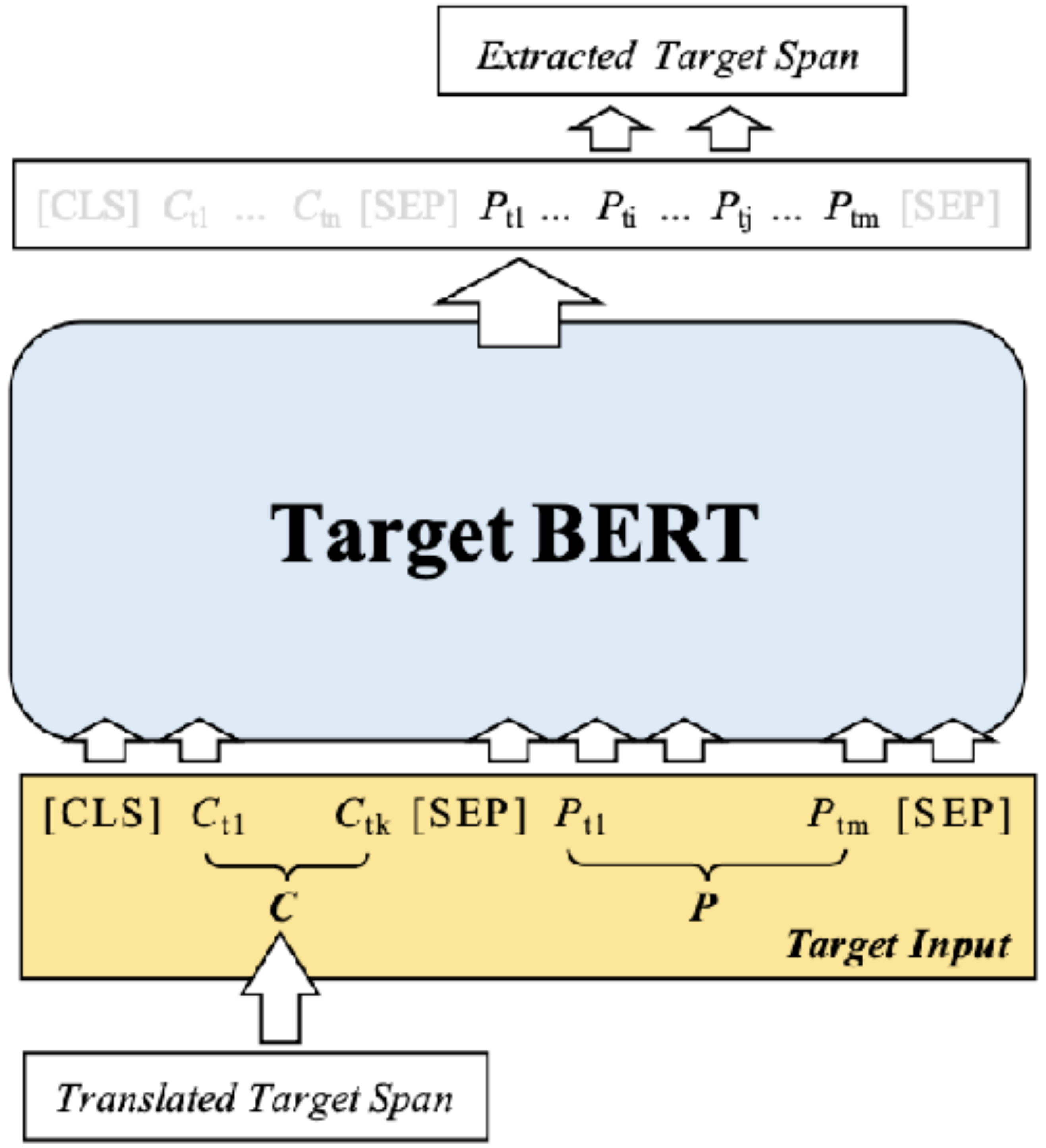


Back-Translation Approaches

- **Simple Match♠**
  - Motivation
    - recover translated answer into EXACT passage span
  - Approach
    - calculate character-level text overlap between translated Answer A_{trans} and arbitrary sliding window in target passage $P_{T[i:j]}$
    - Length of window: len(A_{trans}) ± δ, δ∈[0, 5]
  - We treat the window $P_{T[i:j]}$ that has largest F1-score as the final answer

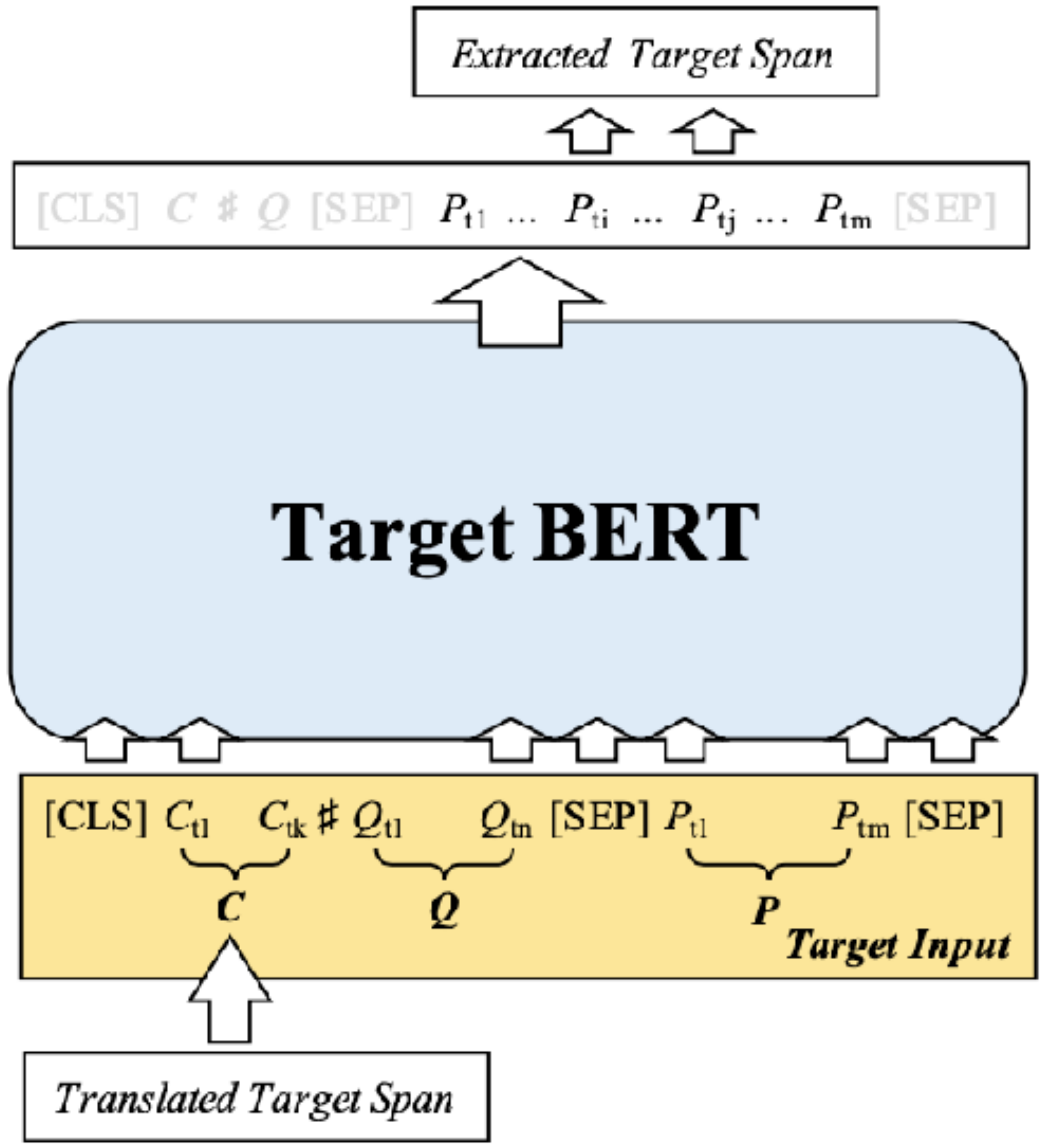
Back-Translation Approaches

- **Answer Aligner**
  - SimpleMatch stops at token-level and lacks semantic awareness between src/trg answers
  - If we have a few annotated data, we could further improve the answer span
  - Condition: A few training data available
  - Solution: Using translated answer and target passage to extract the exact span



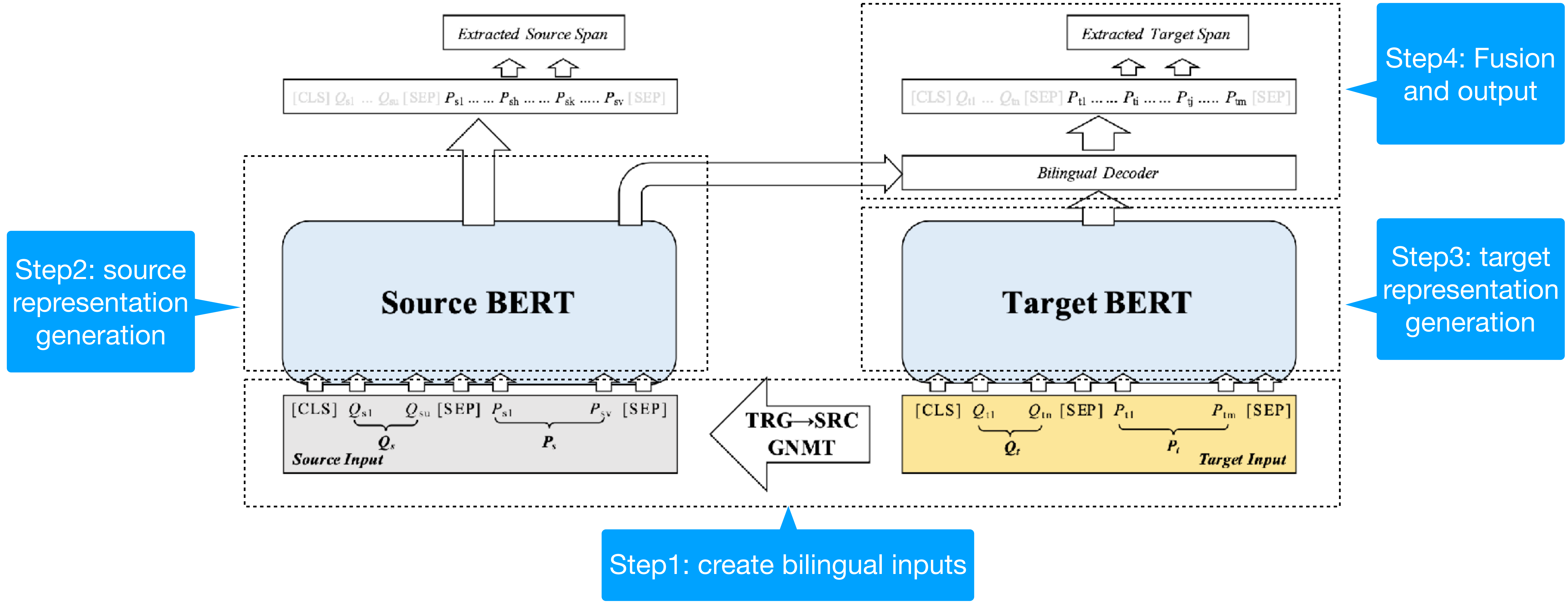
Back-Translation Approaches

- **Answer Verifier**
  - Answer Aligner does not utilize question information
  - Condition: A few training data available
  - Solution: Feed translated target span, target question, and target passage to extract target span



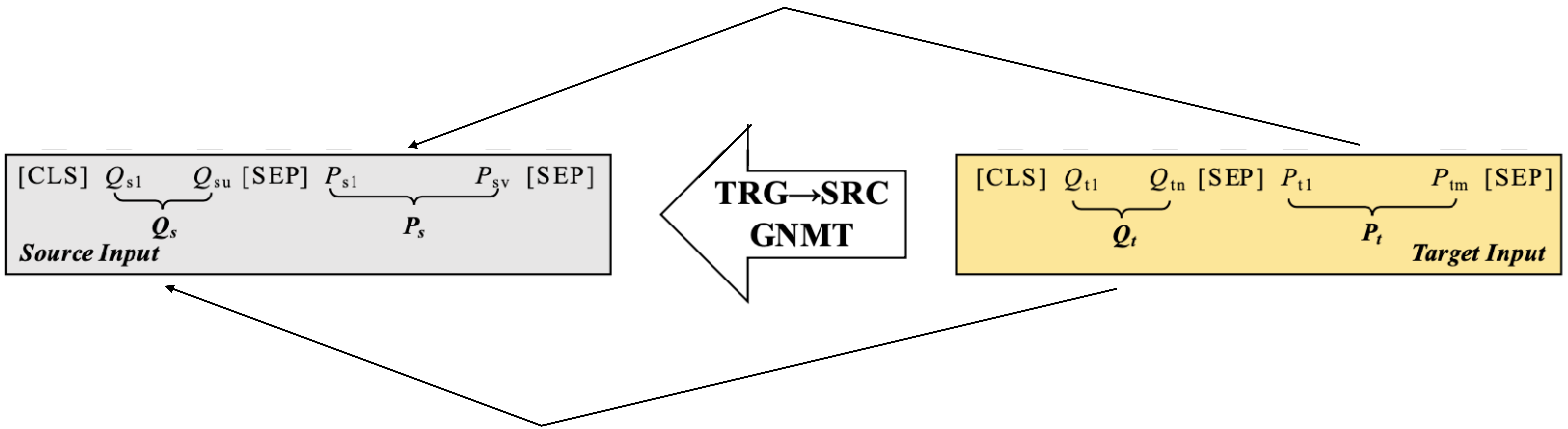
DUAL BERT

- Overview



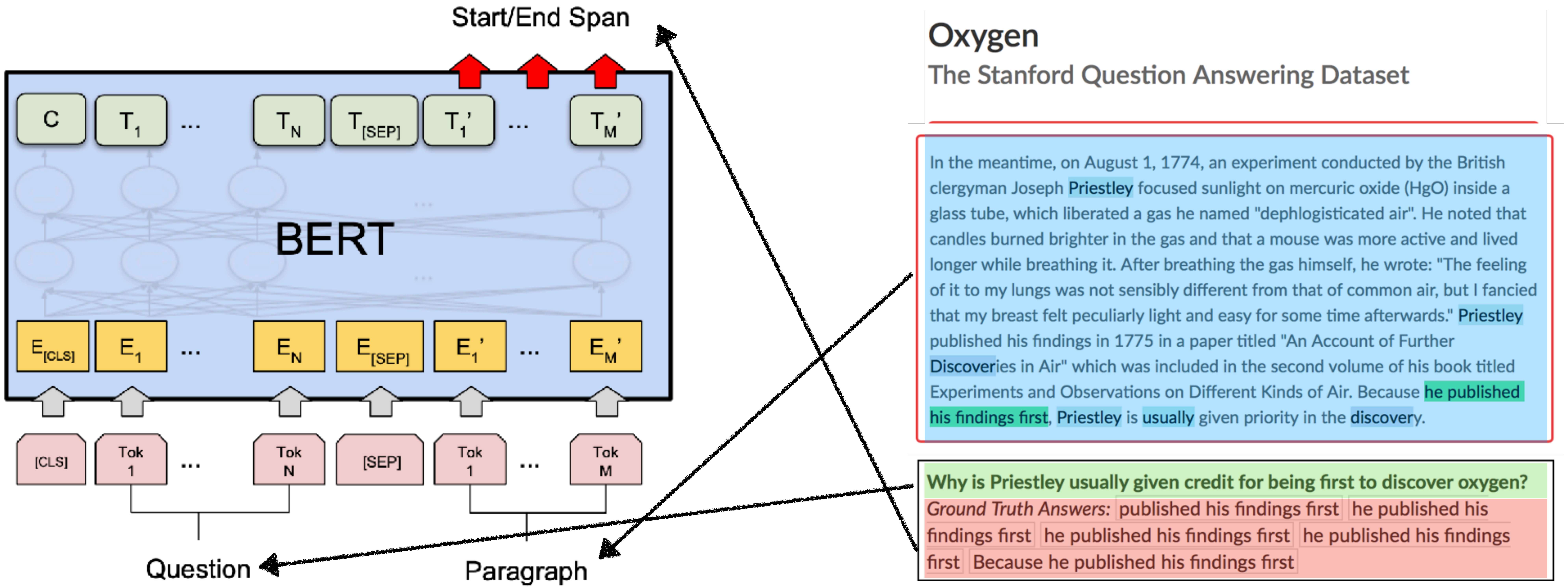
DUAL BERT

- Dual Encoder



Dual BERT

- **Dual Encoder**
  - We use BERT (Devlin et al., NAACL 2019) for RC system



Dual BERT

- **Bilingual Decoder**
  - Raw dot attention

↓ *BERT representation*

$$A_{TS} = B_T \cdot B_S^\top, \;\; A_{TS} \in \mathbb{R}^{L_T * L_S}$$

  - Self-Adaptive Attention (SAA)

$$A_T = \mathbf{softmax}(B_T \cdot B_T^\top)$$
$$A_S = \mathbf{softmax}(B_S \cdot B_S^\top)$$

$$\tilde{A}_{TS} = A_T \cdot A_{TS} \cdot {A_S}^\top$$
$$R' = \mathbf{softmax}(\tilde{A}_{TS}) \cdot B_S$$

DUAL BERT

- **Bilingual Decoder**
  - Fully connected layer with residual layer normalization

$$R = W_r R' + b_r, \quad W_r \in \mathbb{R}^{h*h}$$
$$H_T = concat[B_T, \mathbf{LayerNorm}(B_T + R)]$$

  - Final output for start/end position in the target language

$$P_T^s = \mathbf{softmax}(W_T^\top H_T + b), \quad W_T \in \mathbb{R}^{2h}$$

  - Training objective

Loss for target prediction ↓

$$\mathcal{L} = \mathcal{L}_T + \lambda \mathcal{L}_{aux}$$

↑ *Loss for source prediction*

DUAL BERT

- **How to decide λ?**
  - Idea: measure how the translated samples assemble the real target samples
  - Approach: calculate cosine similarity between ground truth span in the source and target language

Start/End Representation ↓ ↓ ↓ *Span Representation*

$$\tilde{H}_S = concat[B_S^s, B_S^e, B_S^{att}]$$

$$\tilde{H}_T = concat[B_T^s, B_T^e, B_T^{att}]$$

$$\lambda = \max\{0, \cos < \tilde{H}_S, \tilde{H}_T >\}$$

λ → 1, translated samples are good, thus we'd like to use L_{aux}

λ → 0, translated samples are bad, thus we'd rather NOT use L_{aux}

Experiments: Datasets

- Task: Span-Extraction MRC
- Source Language: English
  - SQuAD (Rajpurkar et al., EMNLP 2016)
- Target Language: Chinese
  - CMRC 2018 (Cui et al., EMNLP 2019)
  - DRCD (Shao et al., 2018)

	Train	Dev	Test	Challenge
CMRC 2018				
Question #	10,321	3,219	4,895	504
Answer #	1	3	3	3
DRCD				
Question #	26,936	3,524	3,493	-
Answer #	1	2	2	-

▲ Statistics of CMRC 2018 & DRCD

EXPERIMENTS: SETUPS

- **Tokenization**
  - WordPiece tokenizer (Wu et al., 2016) for English, character-level tokenizer for Chinese
- **BERT**
  - Multilingual BERT (base): 12-layers, 110M parameters
- **Translation**
  - Google Neural Machine Translation (GNMT) API (March, 2019)
- **Optimization**
  - AdamW / lr 4e-5 / cosine lr decay / batch 64 / 2 epochs
- **Implementation**
  - TensorFlow (Abadi et al., 2016) / Cloud TPU v2 (64G HBM)

Experiments: Results

- **Zero-shot Approaches♠**
  - zero-shot: no training data for target language
  - Better source BERT, better target performance
  - Multi-lingual models exceed all other approaches

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
	Human Performance	*91.1*	*97.3*	*92.4*	*97.9*	*90.4*	*95.2*	-	-	*80.4*	*93.3*
	P-Reader (single model)†	59.9	81.5	65.2	84.4	15.1	39.6	-	-	-	-
	Z-Reader (single model)†	79.8	92.7	74.2	88.1	13.9	37.4	-	-	-	-
	MCA-Reader (ensemble)†	66.7	85.5	71.2	88.1	15.5	37.1	-	-	-	-
	RCEN (ensemble)†	76.3	91.4	68.7	85.8	15.3	34.5	-	-	-	-
	r-net (single model)†	-	-	-	-	-	-	-	-	29.1	44.4
	DA (Yang et al., 2019)	49.2	65.4	-	-	-	-	55.4	67.7	-	-
1	GNMT+BERT$_{SQ-B_{en}}$♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6
6	BERT$_{B_{cn}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	BERT$_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3
9	BERT$_{SQ-B_{mul}}$♠	56.5	77.5	59.7	79.9	18.6	41.4	66.7	81.0	65.4	80.1
10	BERT$_{SQ-B_{mul}}$ + Cascade Training	66.6	87.3	71.8	89.4	25.6	52.3	85.2	91.4	84.4	90.8
11	BERT$_{B_{mul}}$ + Mixed Training	66.8	87.5	72.6	89.8	26.7	53.4	85.3	91.6	84.7	91.2
12	**Dual BERT (w/ SQuAD)**	68.0	88.1	73.6	90.2	27.8	55.2	86.0	92.1	85.4	91.6

EXPERIMENTS: RESULTS

- **Back-Translation Approaches**
  - SimpleMatch significantly improves performance
  - SimpleMatch → Aligner → Verifier: The more information we use, better performance we get
- **Without SQuAD Weights**
  - Modeling input in bilingual space could substantially improves performance

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
	Human Performance	*91.1*	*97.3*	*92.4*	*97.9*	*90.4*	*95.2*	-	-	*80.4*	*93.3*
	P-Reader (single model)†	59.9	81.5	65.2	84.4	15.1	39.6	-	-	-	-
	Z-Reader (single model)†	79.8	92.7	74.2	88.1	13.9	37.4	-	-	-	-
	MCA-Reader (ensemble)†	66.7	85.5	71.2	88.1	15.5	37.1	-	-	-	-
	RCEN (ensemble)†	76.3	91.4	68.7	85.8	15.3	34.5	-	-	-	-
	r-net (single model)†	-	-	-	-	-	-	-	-	29.1	44.4
	DA (Yang et al., 2019)	49.2	65.4	-	-	-	-	55.4	67.7	-	-
1	GNMT+BERT$_{SQ-B_{en}}$♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6
6	BERT$_{B_{en}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	BERT$_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3
9	BERT$_{SQ-B_{mul}}$♠	56.5	77.5	59.7	79.9	18.6	41.4	66.7	81.0	65.4	80.1
10	BERT$_{SQ-B_{mul}}$ + Cascade Training	66.6	87.3	71.8	89.4	25.6	52.3	85.2	91.4	84.4	90.8
11	BERT$_{B_{mul}}$ + Mixed Training	66.8	87.5	72.6	89.8	26.7	53.4	85.3	91.6	84.7	91.2
12	**Dual BERT (w/ SQuAD)**	68.0	88.1	73.6	90.2	27.8	55.2	86.0	92.1	85.4	91.6

EXPERIMENTS: RESULTS

- **With SQuAD Weights**
  - Cascade Training
    - SQuAD → CMRC/DRCD
  - Mixed Training
    - SQuAD + CMRC/DRCD
  - Mixed > Cascade
  - Dual BERT again outperforms all previous methods

#	System	CMRC 2018 Dev EM	CMRC 2018 Dev F1	CMRC 2018 Test EM	CMRC 2018 Test F1	CMRC 2018 Challenge EM	CMRC 2018 Challenge F1	DRCD Dev EM	DRCD Dev F1	DRCD Test EM	DRCD Test F1
	Human Performance	*91.1*	*97.3*	*92.4*	*97.9*	*90.4*	*95.2*	-	-	*80.4*	*93.3*
	P-Reader (single model)†	59.9	81.5	65.2	84.4	15.1	39.6	-	-	-	-
	Z-Reader (single model)†	79.8	92.7	74.2	88.1	13.9	37.4	-	-	-	-
	MCA-Reader (ensemble)†	66.7	85.5	71.2	88.1	15.5	37.1	-	-	-	-
	RCEN (ensemble)†	76.3	91.4	68.7	85.8	15.3	34.5	-	-	-	-
	r-net (single model)†	-	-	-	-	-	-	-	-	29.1	44.4
	DA (Yang et al., 2019)	49.2	65.4	-	-	-	-	55.4	67.7	-	-
1	GNMT+BERT$_{SQ-B_{en}}$ ♠	15.9	40.3	20.8	45.4	4.2	20.2	28.1	50.0	26.6	48.9
2	GNMT+BERT$_{SQ-L_{en}}$ ♠	16.8	42.1	21.7	47.3	5.2	22.0	28.9	52.0	28.7	52.1
3	GNMT+BERT$_{SQ-L_{en}}$+SimpleMatch♠	26.7	56.9	31.3	61.6	9.1	35.5	36.9	60.6	37.0	61.2
4	GNMT+BERT$_{SQ-L_{en}}$+Aligner	46.1	66.4	49.8	69.3	16.5	40.9	60.1	70.5	59.5	70.7
5	GNMT+BERT$_{SQ-L_{en}}$+Verifier	64.7	84.7	68.9	86.8	20.0	45.6	83.5	90.1	82.6	89.6
6	BERT$_{B_{cn}}$	63.6	83.9	67.8	86.0	18.4	42.1	83.4	90.1	81.9	89.0
7	BERT$_{B_{mul}}$	64.1	84.4	68.6	86.8	18.6	43.8	83.2	89.9	82.4	89.5
8	**Dual BERT**	65.8	86.3	70.4	88.1	23.8	47.9	84.5	90.8	83.7	90.3
9	BERT$_{SQ-B_{mul}}$ ♠	56.5	77.5	59.7	79.9	18.6	41.4	66.7	81.0	65.4	80.1
10	BERT$_{SQ-B_{mul}}$ + Cascade Training	66.6	87.3	71.8	89.4	25.6	52.3	85.2	91.4	84.4	90.8
11	BERT$_{B_{mul}}$ + Mixed Training	66.8	87.5	72.6	89.8	26.7	53.4	85.3	91.6	84.7	91.2
12	**Dual BERT (w/ SQuAD)**	68.0	88.1	73.6	90.2	27.8	55.2	86.0	92.1	85.4	91.6

Experiments: Results

- **Japanese and French SQuAD**
  - Better MT + Better RC = Better CLMRC
  - Translation attention is not essential for extracting answer span
  - Still, multi-lingual BERT (w/ SQuAD) yields best performance

	Japanese		French	
	EM	F1	EM	F1
Back-Translation†	24.8	42.6	23.5	44.0
+Runtime MT†	37.0	52.2	40.7	61.9
GNMT+BERT$_{L_{en}}$	26.9	46.2	39.1	67.0
+SimpleMatch	37.3	58.0	47.4	71.5
BERT$_{SQ-B_{mul}}$	61.3	73.4	57.6	77.1

▲ Results on Japanese and French SQuAD

EXPERIMENTS: ABLATIONS

- **Ablations on CMRC 2018 data**
  - Pre-training with SQuAD is essential for improving performance
  - With source BERT (cascade training), simultaneously modeling input will have positive impact
  - The other modifications seem to also decrease the performance but not that salient

	EM	F1
Dual BERT (w/ SQuAD)	**68.0**	**88.1**
w/o Auxiliary Loss	67.5 (-0.5)	87.7 (-0.4)
w/o Dynamic Lambda	67.3 (-0.7)	87.5 (-0.6)
w/o Self-Adaptive Att.	67.2 (-0.8)	87.5 (-0.6)
w/o Source BERT	66.6 (-1.4)	87.3 (-0.8)
w/o SQuAD Pre-Train	65.8 (-2.2)	86.3 (-1.8)

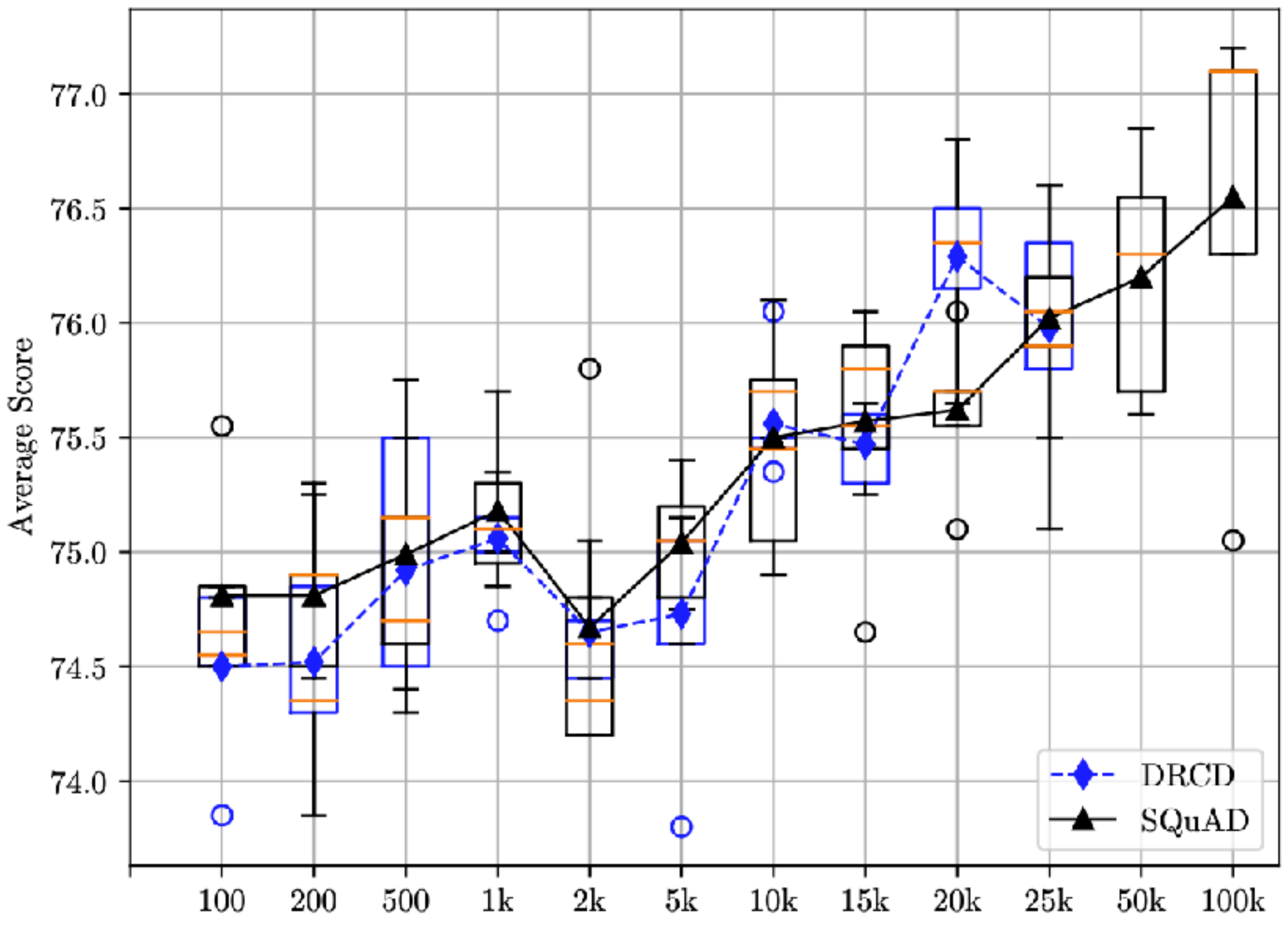
▲ Ablation of Dual BERT on CMRC 2018 dev set

DISCUSSION

- **Question: larger data vs. closer language**
  - Target Language: Simplified Chinese
  - Source Language: ?

Discussion

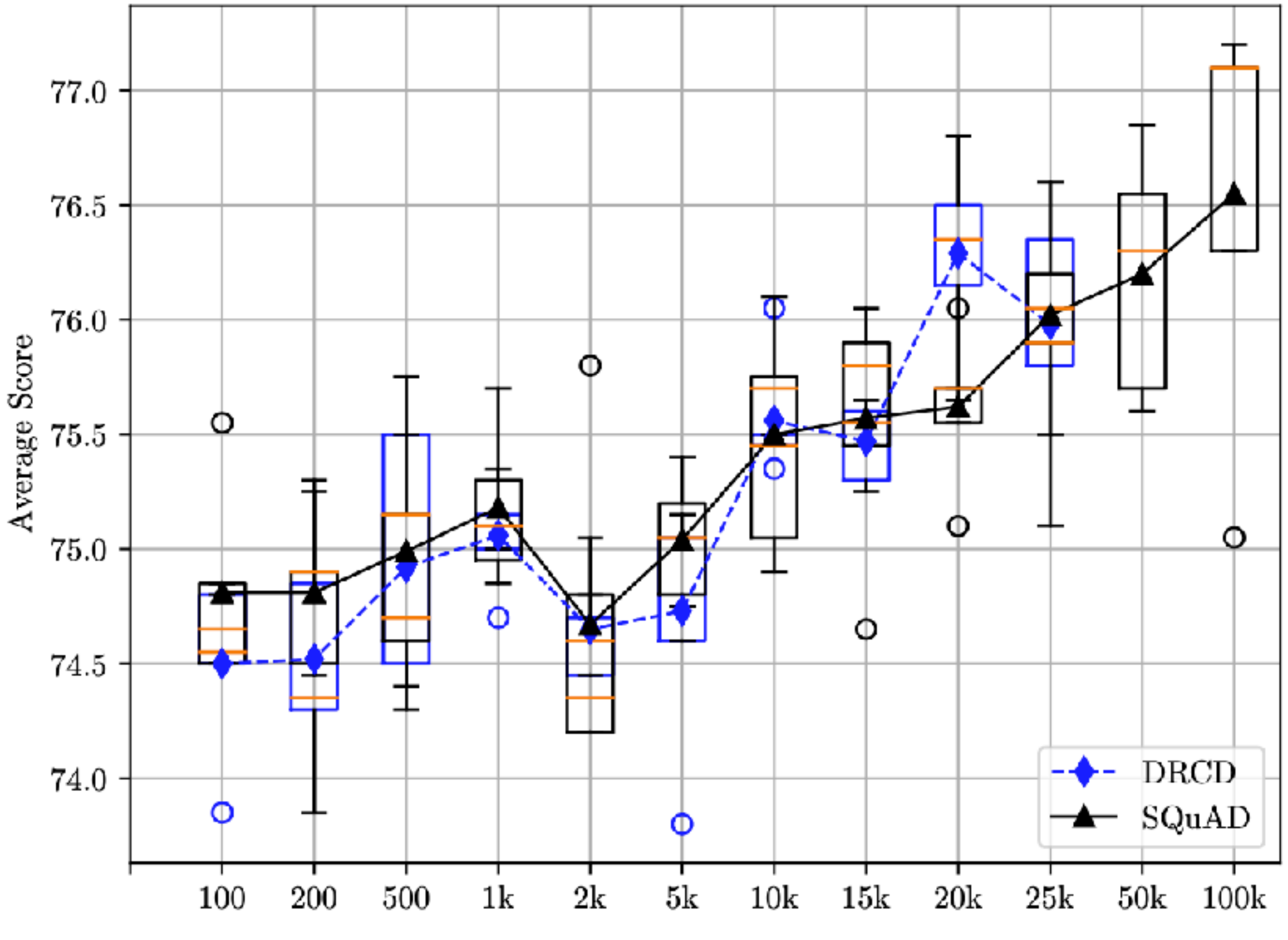
- **Question: larger data vs. closer language**
  - < 25k pre-training data
    - There is no much difference
    - Even English pre-trained models are better than Chinese ones
  - > 25k pre-training data
    - Down-stream task continues to improve significantly



▲ Performance (average of EM and F1) using different amount of pre-training data

Discussion

- **Question: larger data vs. closer language**
  - If the pre-training data is not abundant, there is no preference on the selection of the source language
  - If there are large-scale training data available, use the one that has bigger data, rather than closer to the target language
  - One may also make use of the data in various languages to further exploit knowledge, and we leave this for future work



▲ Performance (average of EM and F1) using different amount of pre-training data

Conclusion & Future Work

- **Conclusion**
  - Propose Cross-Lingual Machine Reading Comprehension (CLMRC)
  - Back-translation approaches for basic cross-lingual MRC purpose
  - Dual BERT for modeling text in bilingual space and enrich representations
  - State-of-the-art performances on Chinese (Simp./Trad.), Japanese, French MRC data
- **Future Work**
  - Utilize various types of English reading comprehension data
  - CLMRC without machine translation process

ACKNOWLEDGMENT

- We would like to thank
  - Google **TensorFlow Research Cloud (TFRC)** Program
  - Anonymous reviewers for theirs valuable comments on our work
- Supporting Funds
  - NSFC 61976072
  - NSFC 61632011
  - NSFC 61772153

USEFUL RESOURCES

- CMRC 2018 (Cui et al., EMNLP 2019)
  - https://github.com/ymcui/cmrc2018
- DRCD (Shao et al., 2018)
  - https://github.com/DRCKnowledgeTeam/DRCD
- Multilingual BERT (Devlin et al., NAACL 2019)
  - https://github.com/google-research/bert/blob/master/multilingual.md
- Google Neural Machine Translation
  - https://cloud.google.com/translate/

References

- Mart´ın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In OSDI, volume 16, pages 265–283.
- Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. arXiv preprint arXiv:1809.03275.
- Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1756–1766. Association for Computational Linguistics.
- Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-Attention neural networks for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 593–602. Association for Computational Linguistics.
- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References

- Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1832–1846. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pages 1684–1692.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301.
- Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read + verify: Machine reading comprehension with unanswerable questions. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):6529–6537.
- Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 908–918. Association for Computational Linguistics.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 102–111. Association for Computational Linguistics.

References

- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392. Association for Computational Linguistics.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hananneh Hajishirzi. 2016. Bi-directional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.
- Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.
- 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, pages 5998–6008.
- Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.
- Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. arXiv preprint arXiv:1611.01604.
- Wei Yang, Yuqing Xie, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. Data augmentation for bert fine-tuning in open-domain question answering. arXiv preprint arXiv:1904.06652.

THANK YOU !

https://github.com/ymcui/Cross-Lingual-MRC

me@ymcui.com