Self-supervised Pre-training for Machine Reading Comprehension

Fangkai Jiao

Shandong University

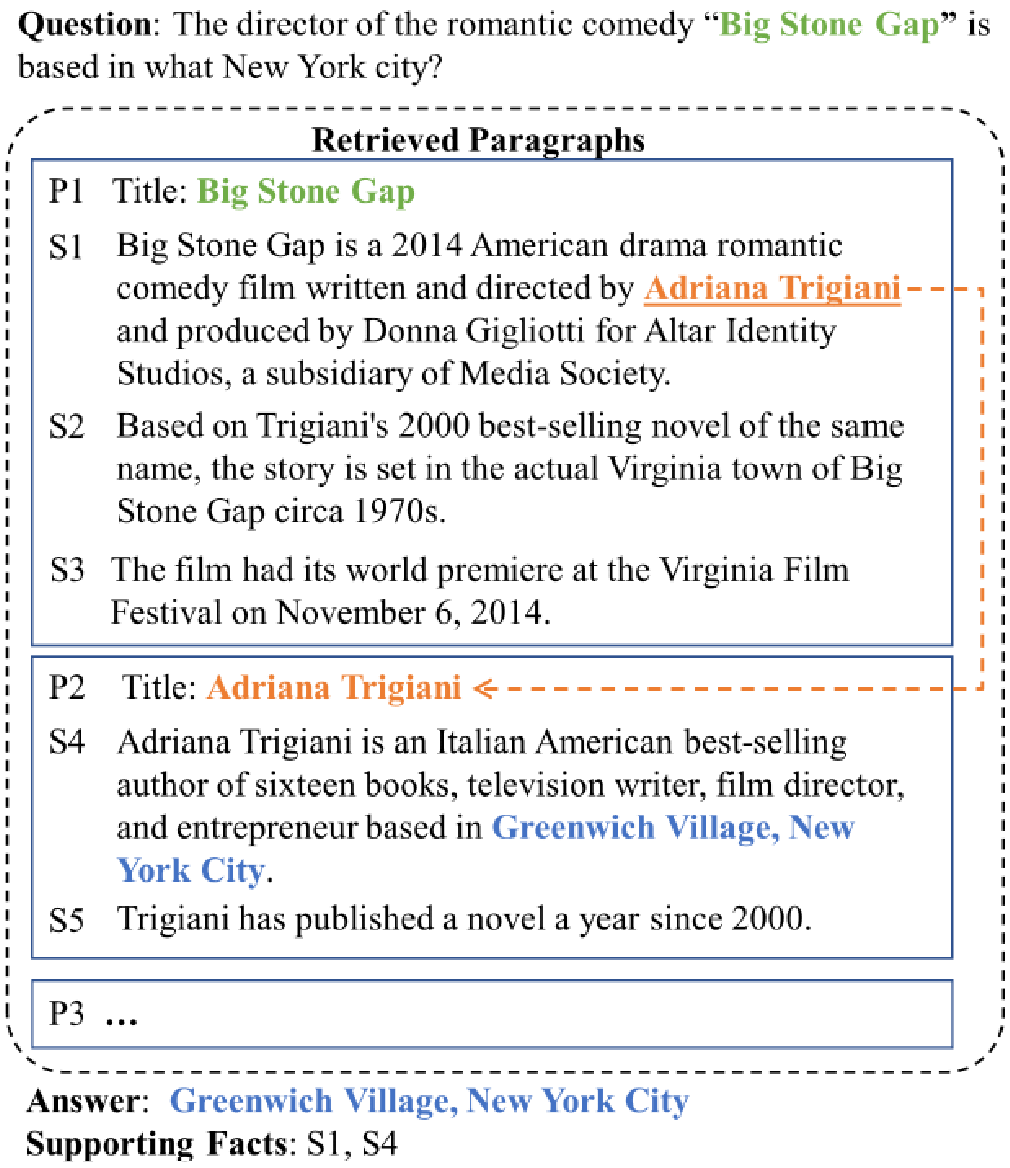
jiaofangkai.com

Outline

- Background
- General Framework
- REPT
- MERIt
- Conclusion & Future Work

Background

- MRC and its application
- Limitation of neural networks for MRC
  - Data hungry
  - Absence of complex reasoning
  - Gap between pre-trained LM and MRC
    - Evidence extraction
    - Reasoning, e.g., logical reasoning



Background

- MRC and its application
- Limitation of neural networks for MRC
  - Data hungry
  - Absence of complex reasoning
  - Gap between pre-trained LM and MRC
    - Evidence extraction
    - Reasoning, e.g., logical reasoning

Chatbot

Search Engine

Documents Analysis

Two General Frameworks

General Pre-training → In-domain Pre-training → Task-oriented Pre-training → Fine-tuning /Prompt-tuning → Knowledge Distillation

---

Pre-text Task Design → Shortcut Elimination → Post-training → Evaluation

REPT: Bridging Language Models and Machine Reading Comprehension via Retrieval-based Pre-training

Fangkai Jiao, Yangyang Guo, Yilin Niu, Feng Ji, Feng-Lin Li, Liqiang Nie. *Findings of ACL 2021*.

Motivation

- PLMs have a significant gap with MRC system.
  - Language modeling focuses on general contextual language representation
  - MRC system requires strong evidence extraction ability to perform reasoning across multiple sentences.

Question: The director of the romantic comedy "Big Stone Gap" is based in what New York city?

Retrieved Paragraphs

P1 Title: Big Stone Gap

S1 Big Stone Gap is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society.

S2 Based on Trigiani's 2000 best-selling novel of the same name, the story is set in the actual Virginia town of Big Stone Gap circa 1970s.

S3 The film had its world premiere at the Virginia Film Festival on November 6, 2014.

P2 Title: Adriana Trigiani

S4 Adriana Trigiani is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in Greenwich Village, New York City.

S5 Trigiani has published a novel a year since 2000.

P3 ...

Answer: Greenwich Village, New York City

Supporting Facts: S1, S4

An Intuitive Idea

- Evidence extraction

  → Sentence-level evidence retriever

  → Pre-training tasks for training the retriever
- Self-supervised tasks

  → Introduce input noise: **masking / shuffling / deleting …**

  → Shuffling can help learn the discourse knowledge in document level.
- Improve the difficulty of task

  → Common entities or nous (coreference) may lead to information leak.

  → Eliminate the information short-cut.

Pre-training Tasks

Given a Wikipedia document,

1. Select 30% sentences as query.
2. Mask the entities and nouns with pre-defined ratio to eliminate the information short cut.
3. **TO**:
    1. Predict the initial preceding and following sentences or each query.
    2. Recover the correct entities and nouns.

Query

1. History The Mentally Retarded Children s Society of SA Inc. was established in 1950 by a group of parents who wanted [MASK A] employment and accommodation opportunities for their children within the local community at a time when institutionalised [MASK B] in Adelaide was their only alternative.
2. Today [MASK C] [MASK D] provides assisted employment assisted accommodation and respite services to people with intellectual disabilities.

Document

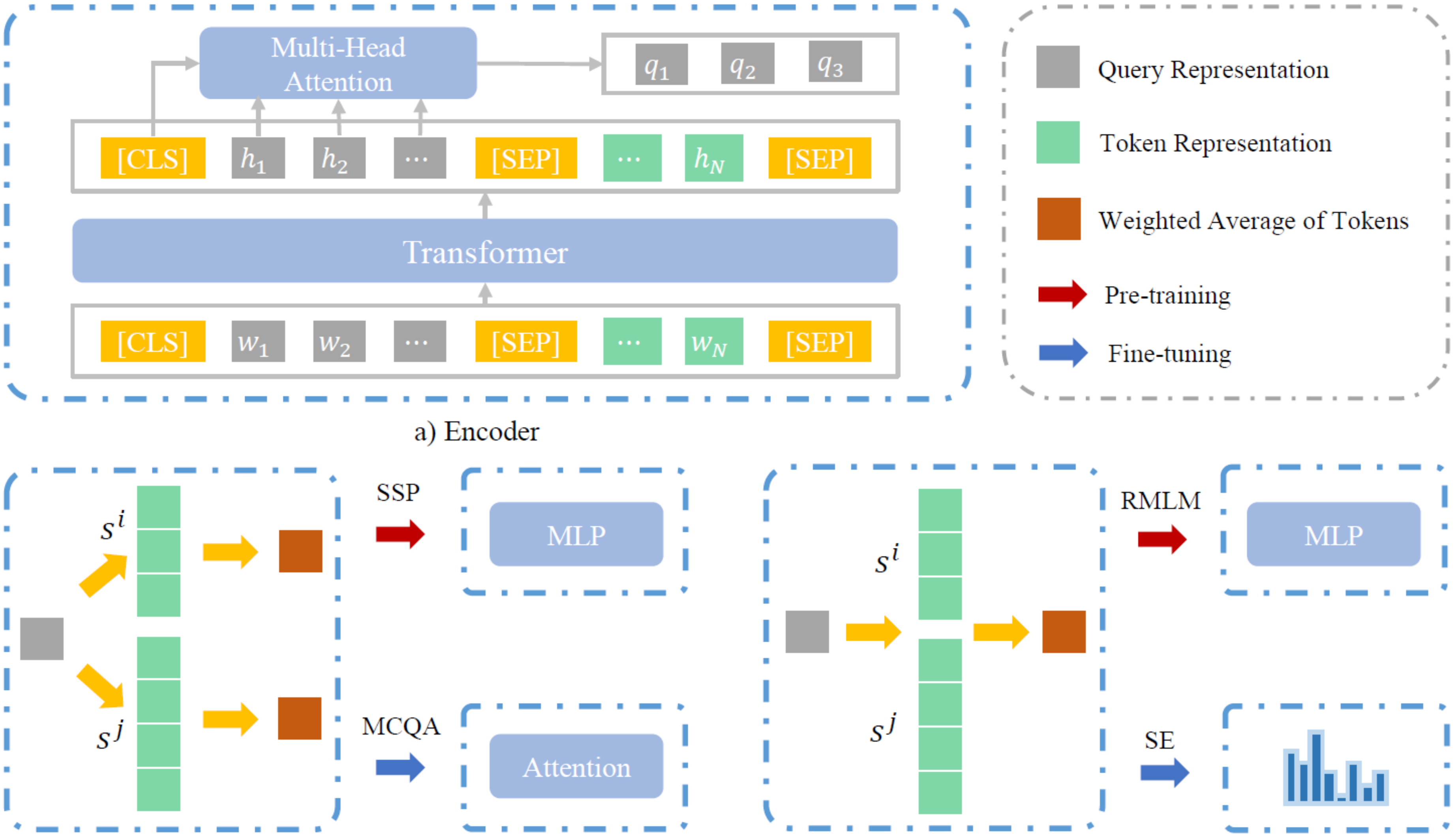
3. The society s aims were to seek education or training facilities for people with intellectual disabilities to establish sheltered workshops and to establish residential hostels.
4. A number of sheltered workshops were established and in 1980 the name was changed to the Aboriginal word Orana which means Welcome .
5. Orana s current and previous clients include Mitsubishi Motors Clipsal RAA Elders Limited and Billycart Kids.
6. Orana was one of the first disability service organisations to achieve Quality Accreditation.
7. After the unveiling of the Australian Government s Commonwealth Home Support Programme CHSP and seeing it as a natural step of progression Orana now provides quality tailored aged care at home.
8. The well resourced organization delivers help across a range of areas helping the elderly remain where they want to be in the comfort of their own home during their later years.
9. Orana continues with its mission to support people remain independent valued and productive members of the community.

Correct order: 1 3 4 2 5 6 7 8 9

Recovery:

1. [MASK A] -> education
2. [MASK B] -> care
3. [MASK C] [MASK D] -> Orana Provides

Model Architecture



b) Sentence-Level Retrieval for SSP and MCQA

c) Document-Level Retrieval for RMLM, ODQA and Span Extraction (SE)

Encoder

- Joint encoding

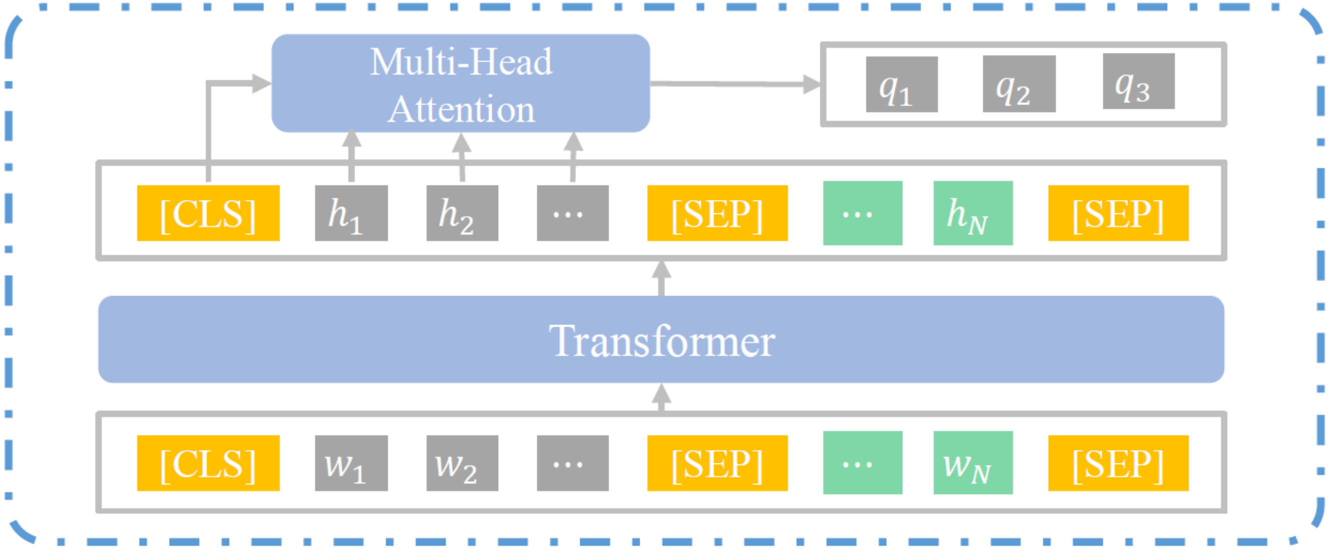
$$\mathbf{H} = [\mathbf{h}_{\text{cls}}, \cdots, \mathbf{h}_m, \mathbf{h}_{\text{sep}}] = \text{Encoder}(\tilde{\mathcal{S}}),$$

$$\mathbf{H} = [\mathbf{H}^1, \mathbf{H}^2, \cdots, \mathbf{H}^n], \ \mathbf{H}^i \in \mathbb{R}^{d \times l},$$

- Query representation

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{softmax}(\frac{\mathbf{Q}_i^\top \mathbf{K}_i}{\sqrt{d}})\mathbf{V}_i)$$

$$\mathbf{v}_0^{q\top} = \text{MHA}(\mathbf{h}_{\text{cls}}^\top, \mathbf{H}^q, \mathbf{H}^q).$$



Sentence-level Evidence Extraction

- Intra-sentence evidence retrieval

$$\mathbf{u}_q^{i\top} = \mathrm{Att}(\mathbf{v}^{q\top}, \mathbf{H}^i, \mathbf{H}^i),$$

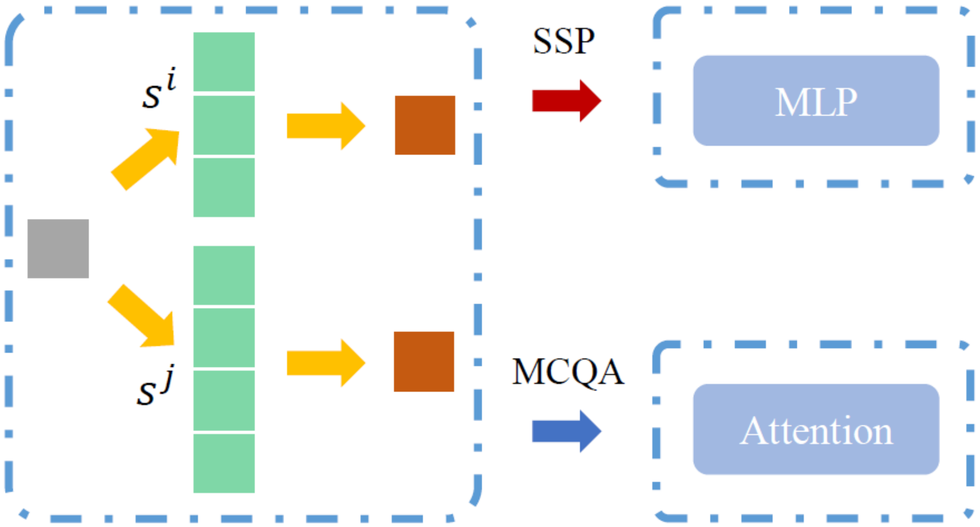
- Surrounding Sentence Prediction

$$\mathbf{o}_q^i = \mathbf{W}_2(\tanh(\mathbf{W}_1\mathbf{u}_q^i + \mathbf{b}_1)) + \mathbf{b}_2.$$

- Multiple Choice QA

$$\mathbf{v}^p = \mathrm{Att}(\mathbf{v}^{q\top}, \mathbf{U}, \mathbf{U}), \;\; \mathbf{U} = [\mathbf{u}_q^1, \cdots, \mathbf{u}_q^n]$$

$$p_c^{\mathrm{mc}} \propto \exp(\mathbf{W}_6(\tanh(\mathbf{W}_5[\mathbf{v}^q; \mathbf{v}^p] + \mathbf{b}_5)) + \mathbf{b}_6).$$



b) Sentence-Level Retrieval for SSP and MCQA

Document-level Evidence Extraction

- Document-level evidence retrieval

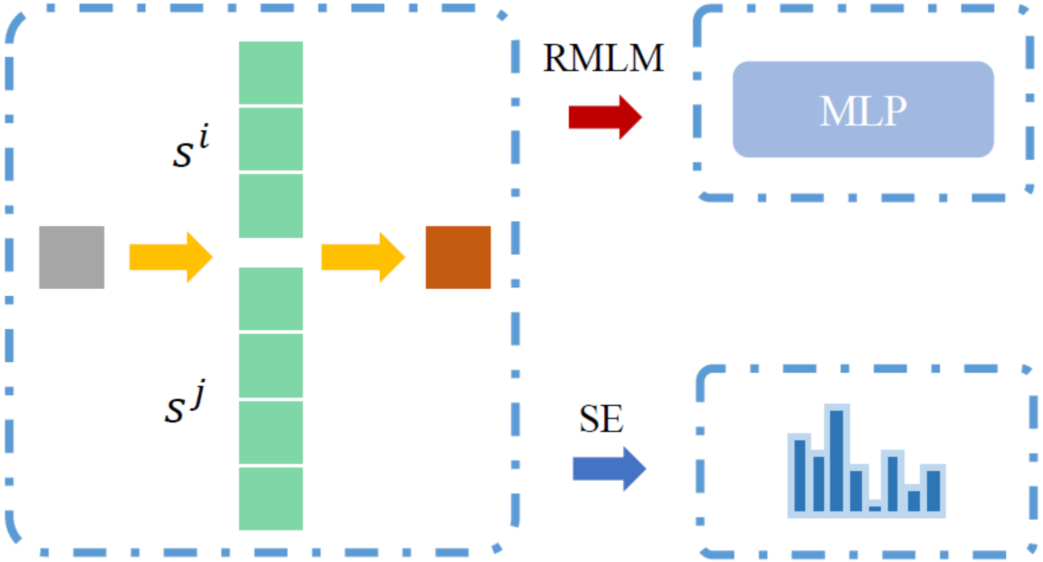
$$\mathbf{g}^{q\top} = \mathrm{Att}(\mathbf{v}^{q\top}, \mathbf{H}, \mathbf{H}).$$

- Masked language modeling

$$\tilde{\mathbf{h}}_z^q = f(\mathbf{h}_z, \mathbf{g}^q),$$

- Span Extraction

$$\begin{cases} p_s^{\text{span}} \propto \exp(\mathbf{v}^{q\top}\mathbf{W}_7\mathbf{h}_s), \\ p_e^{\text{span}} \propto \exp(\mathbf{v}^{q\top}\mathbf{W}_8\mathbf{h}_e). \end{cases}$$



c) Document-Level Retrieval for RMLM, ODQA and Span Extraction (SE)

Optimization

- Surrounding Sentence Prediction

$$\begin{cases} p_{\text{ssp}}(a|q,\mathcal{S}) = \dfrac{\exp(\mathbf{o}_q^a)}{\sum_{j=1,j\notin\{b,q\}}^{n}\exp(\mathbf{o}_q^j)}, \\ p_{\text{ssp}}(b|q,\mathcal{S}) = \dfrac{\exp(\mathbf{o}_q^b)}{\sum_{j=1,j\notin\{a,q\}}^{n}\exp(\mathbf{o}_q^j)}. \end{cases} \qquad \mathcal{L}_{\text{ssp}} = \mathbb{E}(-\frac{1}{Q}\sum_q(\log p_{\text{ssp}}(a|q,\mathcal{S}) + \log p_{\text{ssp}}(b|q,\mathcal{S})))$$

- Retrieval based MLM

$$p_{\text{rmlm}}(x_z|z,q,\mathcal{S}) = \frac{\exp(\mathrm{e}(x_z)^\top \tilde{\mathbf{h}}_z^q)}{\sum_{x'}\exp(\mathrm{e}(x')^\top \tilde{\mathbf{h}}_z^q)} \qquad \mathcal{L}_{\text{rmlm}} = \mathbb{E}(-\frac{\sum_q\sum_z \log p_{\text{rmlm}}(x_z|z,q,\mathcal{S})}{\sum_q|\mathcal{Z}^q|})$$

Dataset

- Pre-training: Wikipedia (13 GB)
- Multiple Choice QA
  - DREAM
  - RACE
  - Multi-RC
  - ReClor
- Span Extraction QA
  - Hotpot QA

Baseline

- BERT & RoBERTa
- BERT-Q & RoBERTa-Q
- BERT-Q w. **R** & BERT-Q w. **S**
- BERT-Q w. **M** & BERT w. **M**
- BERT-Q w. R/S & RoBERTa-Q w. R/S **(Ours model)**

Results of Multiple Choice QA

Model / Dataset	RACE		DREAM		ReClor		Multi-RC		
	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.	Dev EM	Dev $F1_a$	Dev $F1_m$
BERT-base†	–	65.0	63.4	63.2	**54.6**	47.3	–	–	–
BERT w. M	67.7	66.3	62.9	63.2	51.6	45.1	26.6	71.8	74.2
BERT-Q	67.2	65.2	62.9	62.3	48.4	45.0	22.8	69.6	72.0
BERT-Q w. M	67.7	66.9	61.8	62.2	48.8	48.3	23.8	70.1	72.6
BERT-Q w. R	65.5	64.7	59.0	58.6	46.8	45.1	26.4	71.5	74.0
BERT-Q w. S	69.5	66.5	**64.8**	62.2	52.0	46.5	30.0	73.0	75.8
BERT-Q w. R/S	**70.1**	**68.1**	64.4	**64.0**	50.6	**49.2**	**31.9**	**73.8**	**76.3**
RoBERTa-base	76.0	75.5	**71.2**	69.8	54.8	**50.8**	38.7	77.1	79.2
RoBERTa-Q	76.8	**75.7**	70.9	69.5	**56.0**	49.7	34.6	75.4	77.4
RoBERTa-Q w. R/S	**77.1**	74.9	70.9	**70.8**	54.8	50.3	**40.4**	**77.6**	**80.0**

Table 1: Results on multiple choice question answering tasks. ($F1_a$: F1 score on all answer-options; $F1_m$: macro-average F1 score of all questions.) We ran all experiments using **four** different random seeds with the same hyper-parameters, and report the average performance, except for ReClor and Multi-RC. For ReClor, we submitted the best model on development set to the leaderboard to get the results on test set. For MultiRC, we merely reported the performance on development set since the test set is unavailable. †: The results are reported by the leaderboard.

Results of Multiple Choice QA

Model / Dataset	RACE Dev Acc.	RACE Test Acc.	DREAM Dev Acc.	DREAM Test Acc.	ReClor Dev Acc.	ReClor Test Acc.	Multi-RC Dev EM	Multi-RC Dev $F1_a$	Multi-RC Dev $F1_m$
BERT-base†	–	65.0	63.4	63.2	**54.6**	47.3	–	–	–
BERT w. M	67.7	66.3	62.9	63.2	51.6	45.1	26.6	71.8	74.2
BERT-Q	67.2	65.2	62.9	62.3	48.4	45.0	22.8	69.6	72.0
BERT-Q w. M	67.7	66.9	61.8	62.2	48.8	48.3	23.8	70.1	72.6
BERT-Q w. R	65.5	64.7	59.0	58.6	46.8	45.1	26.4	71.5	74.0
BERT-Q w. S	69.5	66.5	**64.8**	62.2	52.0	46.5	30.0	73.0	75.8
BERT-Q w. R/S	**70.1**	**68.1**	64.4	**64.0**	50.6	**49.2**	**31.9**	**73.8**	**76.3**
RoBERTa-base	76.0	75.5	**71.2**	69.8	54.8	**50.8**	38.7	77.1	79.2
RoBERTa-Q	76.8	**75.7**	70.9	69.5	**56.0**	49.7	34.6	75.4	77.4
RoBERTa-Q w. R/S	**77.1**	74.9	70.9	**70.8**	54.8	50.3	**40.4**	**77.6**	**80.0**

Table 1: Results on multiple choice question answering tasks. ($F1_a$: F1 score on all answer-options; $F1_m$: macro-average F1 score of all questions.) We ran all experiments using **four** different random seeds with the same hyper-parameters, and report the average performance, except for ReClor and Multi-RC. For ReClor, we submitted the best model on development set to the leaderboard to get the results on test set. For MultiRC, we merely reported the performance on development set since the test set is unavailable. †: The results are reported by the leaderboard.

Results of Multiple Choice QA

Model / Dataset	RACE Dev Acc.	RACE Test Acc.	DREAM Dev Acc.	DREAM Test Acc.	ReClor Dev Acc.	ReClor Test Acc.	Multi-RC Dev EM	Multi-RC Dev $F1_a$	Multi-RC Dev $F1_m$
BERT-base†	–	65.0	63.4	63.2	**54.6**	47.3	–	–	–
BERT w. M	67.7	66.3	62.9	63.2	51.6	45.1	26.6	71.8	74.2
BERT-Q	67.2	65.2	62.9	62.3	48.4	45.0	22.8	69.6	72.0
BERT-Q w. M	67.7	66.9	61.8	62.2	48.8	48.3	23.8	70.1	72.6
BERT-Q w. R	65.5	64.7	59.0	58.6	46.8	45.1	26.4	71.5	74.0
BERT-Q w. S	69.5	66.5	**64.8**	62.2	52.0	46.5	30.0	73.0	75.8
BERT-Q w. R/S	**70.1**	**68.1**	64.4	**64.0**	50.6	**49.2**	**31.9**	**73.8**	**76.3**
RoBERTa-base	76.0	75.5	**71.2**	69.8	54.8	**50.8**	38.7	77.1	79.2
RoBERTa-Q	76.8	**75.7**	70.9	69.5	**56.0**	49.7	34.6	75.4	77.4
RoBERTa-Q w. R/S	**77.1**	74.9	70.9	**70.8**	54.8	50.3	**40.4**	**77.6**	**80.0**

Table 1: Results on multiple choice question answering tasks. ($F1_a$: F1 score on all answer-options; $F1_m$: macro-average F1 score of all questions.) We ran all experiments using **four** different random seeds with the same hyper-parameters, and report the average performance, except for ReClor and Multi-RC. For ReClor, we submitted the best model on development set to the leaderboard to get the results on test set. For MultiRC, we merely reported the performance on development set since the test set is unavailable. †: The results are reported by the leaderboard.

Results of Multiple Choice QA

Model / Dataset	RACE		DREAM		ReClor		Multi-RC		
	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.	Dev Acc.	Test Acc.	Dev EM	Dev $F1_a$	Dev $F1_m$
BERT-base†	–	65.0	63.4	63.2	**54.6**	47.3	–	–	–
BERT w. M	67.7	66.3	62.9	63.2	51.6	45.1	26.6	71.8	74.2
BERT-Q	67.2	65.2	62.9	62.3	48.4	45.0	22.8	69.6	72.0
BERT-Q w. M	67.7	66.9	61.8	62.2	48.8	48.3	23.8	70.1	72.6
BERT-Q w. R	65.5	64.7	59.0	58.6	46.8	45.1	26.4	71.5	74.0
BERT-Q w. S	69.5	66.5	**64.8**	62.2	52.0	46.5	30.0	73.0	75.8
BERT-Q w. R/S	**70.1**	**68.1**	64.4	**64.0**	50.6	**49.2**	**31.9**	**73.8**	**76.3**
RoBERTa-base	76.0	75.5	**71.2**	69.8	54.8	**50.8**	38.7	77.1	79.2
RoBERTa-Q	76.8	**75.7**	70.9	69.5	**56.0**	49.7	34.6	75.4	77.4
RoBERTa-Q w. R/S	**77.1**	74.9	70.9	**70.8**	54.8	50.3	**40.4**	**77.6**	**80.0**

Table 1: Results on multiple choice question answering tasks. ($F1_a$: F1 score on all answer-options; $F1_m$: macro-average F1 score of all questions.) We ran all experiments using **four** different random seeds with the same hyper-parameters, and report the average performance, except for ReClor and Multi-RC. For ReClor, we submitted the best model on development set to the leaderboard to get the results on test set. For MultiRC, we merely reported the performance on development set since the test set is unavailable. †: The results are reported by the leaderboard.

Results of Span Extraction QA

Model / Dataset	Dev		Test	
	EM	F1	EM	F1
Transformer-XH (Zhao et al., 2020)	54.0	66.2	51.6	64.7
HGN (Fang et al., 2020)	–	–	56.7	69.2
GRR + BERT-wwm-Large*	**60.5**	**73.3**	**60.0**	**73.0**
GRR + BERT-base*	52.7	65.8	–	–
GRR + BERT-Q w. R/S	**55.2**	**68.4**	–	–
GRR + RoBERTa-base	56.8	69.6	–	–
GRR + RoBERTa-Q w. R/S	**58.4**	**71.3**	58.1	71.0

Table 2: Results of our method and other strong baselines on Hotpot QA. *GRR* means the Graph Recurrent Retriever proposed by Asai et al. (2020), *GRR + BERT-base* means the system whose retriever is GRR and reader is built on BERT-base. *: The results are reported by Asai et al. (2020).

Model / Dataset	EM	F1
BERT-Q	71.7	74.9
BERT-Q w. R/S	**77.2**	**80.4**
RoBERTa-Q	80.3	83.7
RoBERTa-Q w. R/S	**81.7**	**85.0**

Table 3: Results of our method and other baselines on the dev set of SQuAD2.0.

Analysis

Model	**P@1**	**R@1**	**P@2**	**R@2**
BERT-Q	21.83	9.66	20.24	17.73
BERT-Q w. R/S	**45.30**	**20.38**	**38.51**	**34.55**
RoBERTa-Q	28.25	12.45	26.93	23.74
RoBERTa-Q w. R/S	**35.34**	**15.76**	**30.33**	**26.85**

Table 3: Results of evidence extraction on the development set of Multi-RC.

Model/Dataset	RACE		Multi-RC		
	Dev	Test	Dev		
	Acc.	Acc.	EM	$F1_a$	$F1_m$
B.Q w.R/S (30%)	70.1	68.1	31.9	**73.8**	**76.3**
B.Q w.R/S (60%)	70.2	67.3	**32.0**	**73.8**	**76.3**
B.Q w.R/S (90%)	**70.4**	**68.2**	31.0	73.5	76.2
B.Q w.S (No Mask)	69.0	67.2	29.0	72.7	75.4

Table 4: Results on RACE and Multi-RC using models pre-trained with different mask ratios. *B.Q* means *BERT-Q*.

Performance under Low Resource

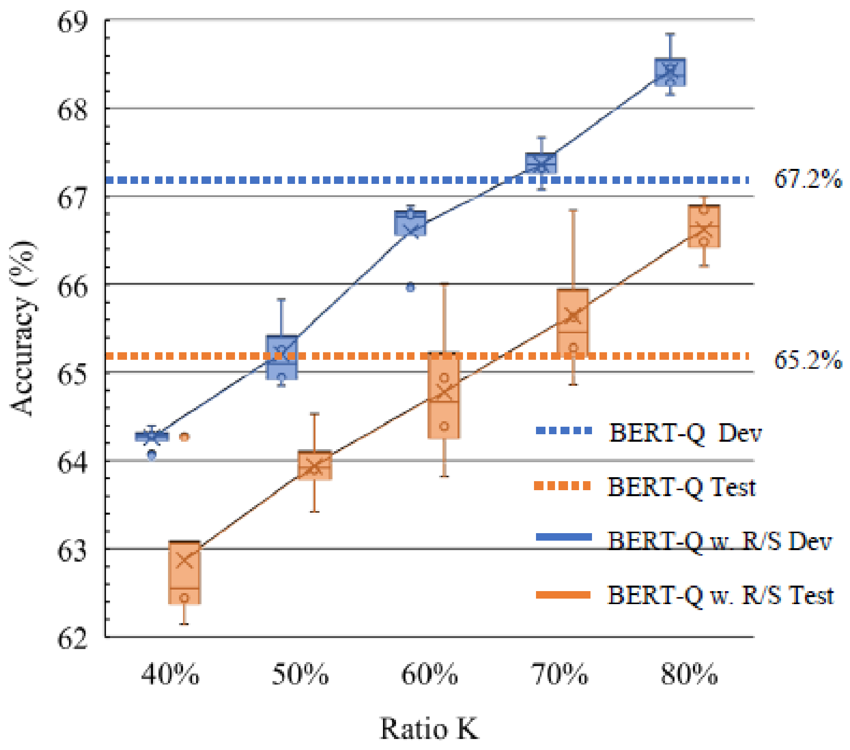


Figure 3: The accuracy of BERT-Q w. R/S on the development and test of RACE. The horizontal axis refers to the ratio K of training data compared to the original training set.

Case 1

Passage:

(0)A group of researchers at a remote jungle island outpost discover the natives are practicing voodoo and black magic. … (4)She returns years later as an adult with a group of mercenaries to attempt to uncover what happened to her parents. (5)Shortly after arriving at the island their boat's engine dies, stranding them. (6)Meanwhile elsewhere on the island a trio of hikers discover a cave, the same cave leading to the underground temple where the original curse was created. (7)After accidentally reviving the curse, the dead once again return to kill any who trespass on their island. (8)The mercenaries encounter their first zombie, who injures a member of the team. (9)Taking shelter in the remains of the old research facilities medical quarters they are soon joined by Chuck, the only surviving hiker. (10)Arming themselves with weapons left behind by the long dead research team, they make their stand as the dead once again rise. (11)One by one they are injured or killed, one of whom sacrifices himself to blow up the medical facility and his newly undead team members. (12)Jenny and Chuck flee, the only survivors remaining. (13)They stumble upon the cave once again, where the zombies appear and attack.

Question: Where did Chuck find weapons?

Option: From the previous research team.

Sentences Used: 9, 10.

BERT-Q: Answer: False Evidence: 0

BERT-Q w. R/S: Answer: True Evidence: 10, 9

Case 2

Passage:

(0)The film opens with Sunita, a medical student , and her friends working on a project about the human brain. (1)She wants to investigate the curious case of Sanjay Singhania, a notable city businessman, who is reported to have anterograde amnesia. (2)Her professor denies access to Sanjay's records as it is currently under criminal investigation. (3)Sunita, nonetheless, decides to investigate the matter herself. (4)Sanjay is introduced as he brutally murders a man. (5)He takes a Polaroid picture of the man, and writes on it ``done". (6)It is revealed that Sanjay has anterograde amnesia where he loses his memory every 15 minutes. (7)Sanjay uses a system of photographs, notes, and tattoos on his body to recover his memory after each cycle. (8)It is revealed that Sanjay is ultimately out to avenge the death of his sweetheart Kalpana , and that he is systematically killing the people who were responsible for it. (9)His main target is ``Ghajini", a notable social personality in the city. (10)Police Inspector Arjun Yadav, on the case of the serial murders, tracks Sanjay down to his flat and attacks and disables him. (11)Yadav finds two diaries where Sanjay has chronicled the events of 2005 and 2006. …

Question: Who denies Sunita access to Sanjay's records, who is reported to have anterograde amnesia, because they are under criminal investigation?

Option: Sunita's professor&Arjun Yadav.

Sentences Used: 1, 2.

RoBERTa-Q: Answer: False Evidence: 0

RoBERTa-Q w. R/S: Answer: True Evidence: 2, 1

Limitation

- The upper bound by evidence extraction[1].
- The gap between the different queries in SSP and MRC.
- Better solution[2 3].

Model / Dataset	MultiRC Dev		
	$F1_m$	$F1_a$	EM_0
GPT+DPL	70.5	67.8	13.3
BERT-MLP	71.8	69.1	21.2
BERT-HA	70.1	68.1	19.9
BERT-HA+RL	72.1	69.5	21.1
BERT-HA+Rule	69.5	66.7	17.9
BERT-HA+STM	**74.0**‡	**70.9**‡	**22.0**†
BERT-HA+Gold	73.7	70.9	27.2

Model/Metric	Ans. Acc	Evi. Acc
RoBERTa-HA	92.6	13.8
RoBERTa-HA+STM	**92.7**	**19.3**(+40%)

Table 6: Answer prediction accuracy (Ans. Acc) and evidence extraction accuracy (Evi. Acc) on the development set of CoQA.

1 A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction. Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, Minlie Huang. ACL 2020.
2 ReasonBERT: Pre-trained to Reason with Distant Supervision. Xiang Deng, Yu Su, Alyssa Lees, You Wu, Cong Yu, Huan Sun. EMNLP 2021.
3 Few-Shot Question Answering by Pretraining Span Selection. Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, Omer Levy. ACL 2021.

MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning

Fangkai Jiao, Yangyang Guo, Xuemeng Song, Liqiang Nie. *Under review.*

Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text

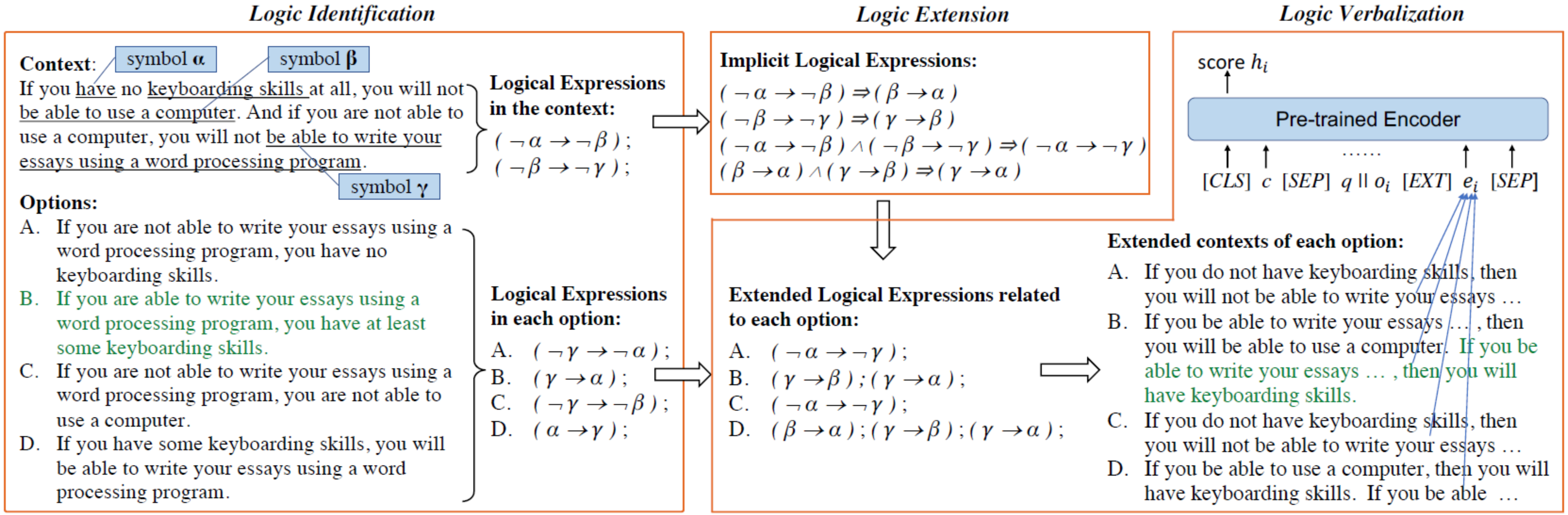


Figure 2: The overall architecture of our proposed logic-driven context extension framework. c, q, o_i and e_i are the context, question, i-th option and the extended context for i-th option, respectively. The texts in green mean that the option B is matched against its extended context which has the highest score.

Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text

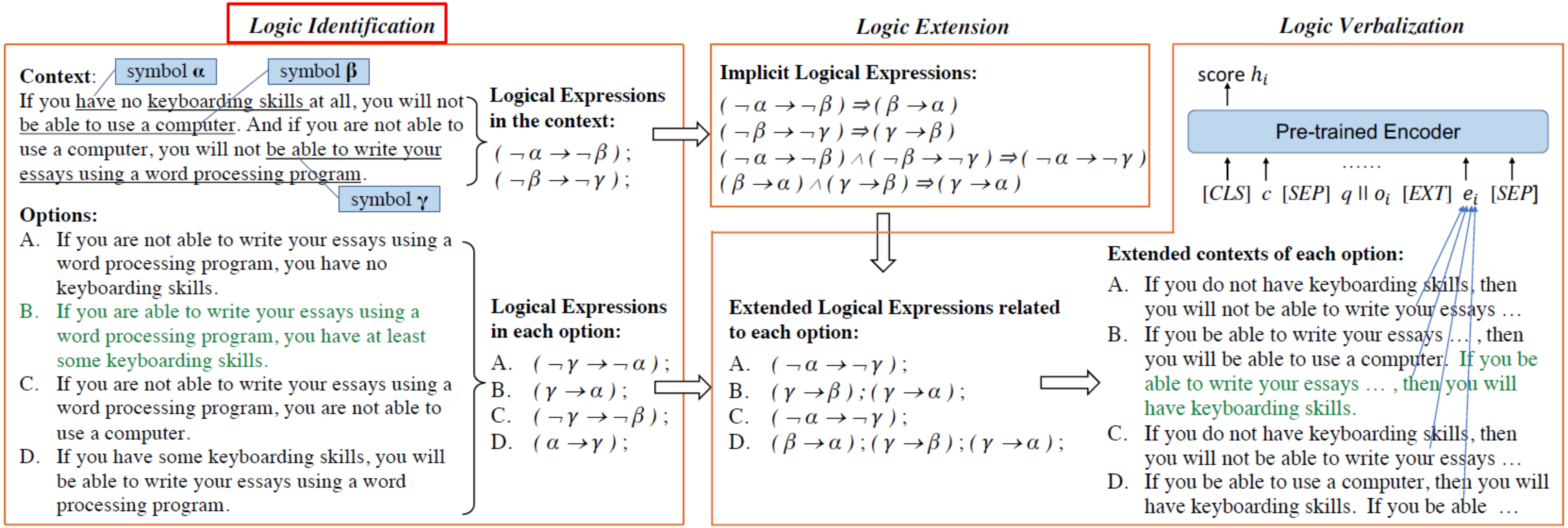


Figure 2: The overall architecture of our proposed logic-driven context extension framework. c, q, o_i and e_i are the context, question, i-th option and the extended context for i-th option, respectively. The texts in green mean that the option B is matched against its extended context which has the highest score.

Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text

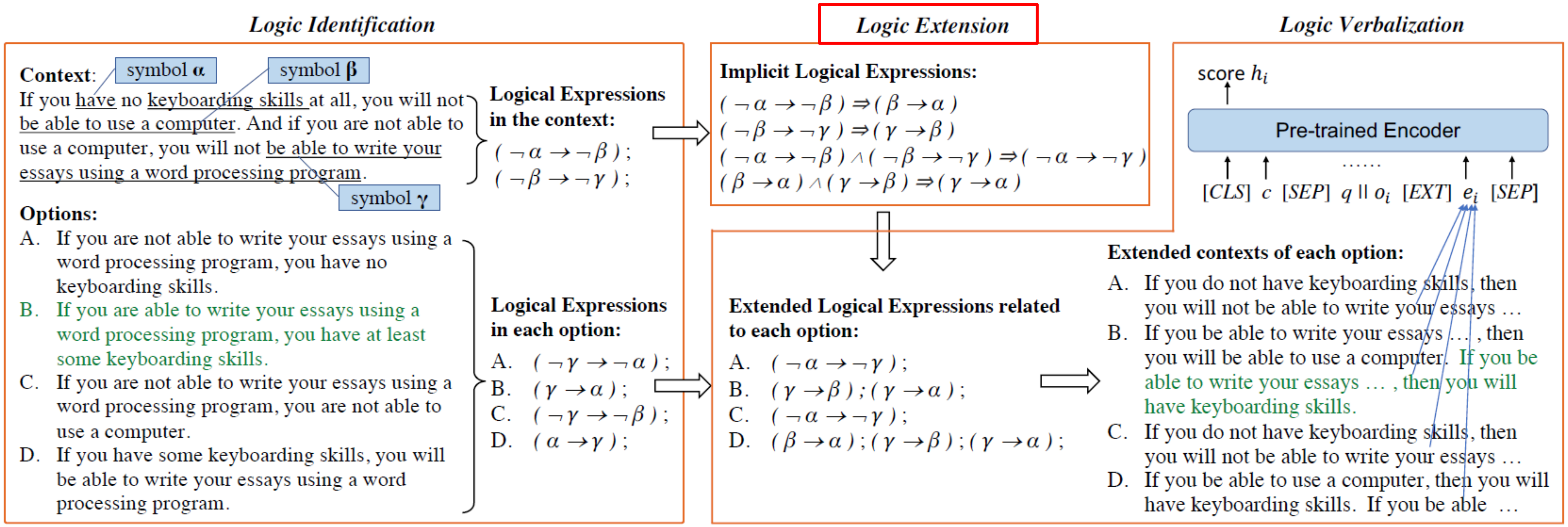


Figure 2: The overall architecture of our proposed logic-driven context extension framework. c, q, o_i and e_i are the context, question, i-th option and the extended context for i-th option, respectively. The texts in green mean that the option B is matched against its extended context which has the highest score.

Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text

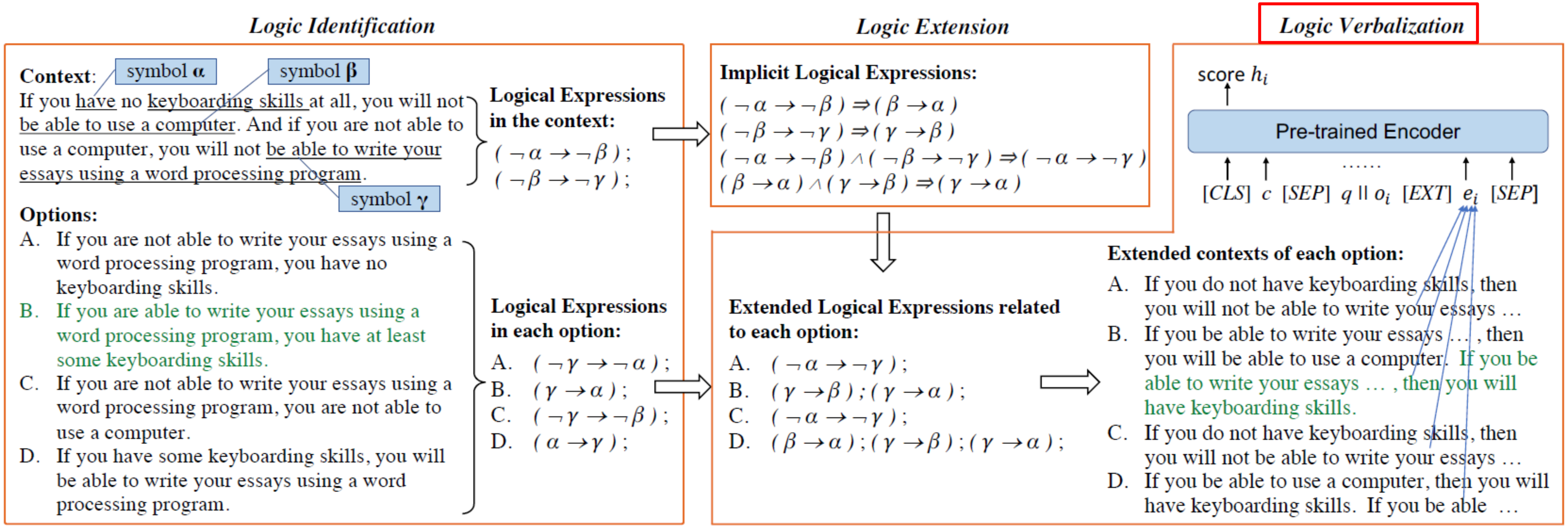


Figure 2: The overall architecture of our proposed logic-driven context extension framework. c, q, o_i and e_i are the context, question, i-th option and the extended context for i-th option, respectively. The texts in green mean that the option B is matched against its extended context which has the highest score.

Logic-Driven Context Extension and Data Augmentation for Logical Reasoning of Text

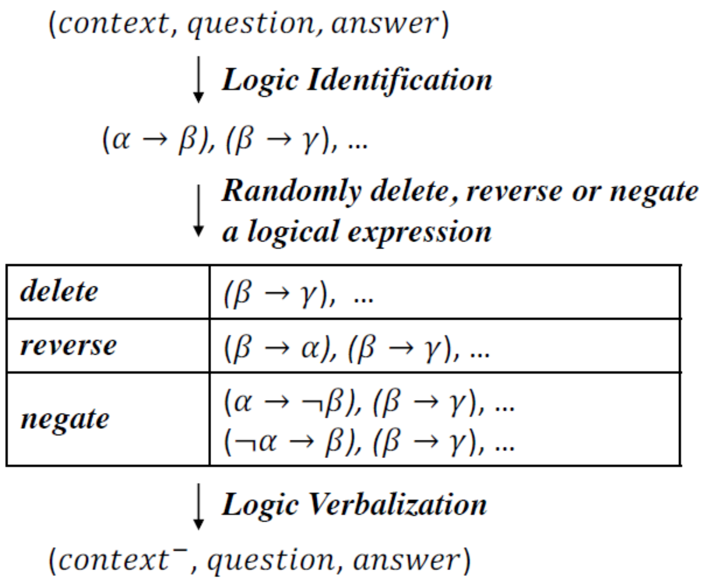


Figure 3: Procedure to construct a logical negative sample.

Logic-Driven Contrastive Learning In our question answering setting, we alter the score function from measuring the similarity between two representations towards calculating the score that the question can be solved by the correct answer under a given context:

$$s^{'}(c^{+}, q, o_a) \gg s^{'}(c^{-}, q, o_a) \tag{5}$$

Questions

1. Is it possible to employ the framework on unlabeled data?
2. How to discover the potential logical structure in raw text instead of particular text, e.g., a document from LSAT?
3. How to construct positive and negative data pairs to facilitate contrastive learning?
4. Does any trivial solution exist? How to avoid the trivial solution?

Preliminary

- Contrastive Learning:

$$\mathcal{L} = L(x, x^+, \mathcal{X}^-) = -\log \frac{\exp f(x, x^+)}{\sum_{x' \in \mathcal{X}^- \cup \{x^+\}} \exp f(x, x')} \quad (1)$$

- Symbolic Logic Reasoning

$$\langle v_i, r_{i,j}, v_j \rangle \leftarrow \left(v_i \xrightarrow{r_{i,i+1}} v_{i+1} \xrightarrow{r_{i+1,i+2}} \cdots \xrightarrow{r_{j-1,j}} v_j \right)$$

- Meta-Path
    - Given a knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
    - A meta-path connecting the entity pair $\langle e_i, e_j \rangle$:

$$e_i \xrightarrow{r_{i,i+1}} e_{i+1} \xrightarrow{r_{i+1,i+2}} \cdots \xrightarrow{r_{j-1,j}} e_j$$

From Logical Reasoning to Meta-Path

- A typical logical structure:
  - $\langle v_i, r_{i,j}, v_j \rangle \leftarrow \left(v_i \xrightarrow{r_{i,i+1}} v_{i+1} \xrightarrow{r_{i+1,i+2}} \cdots \xrightarrow{r_{j-1,j}} v_j \right)$ (2)
- Take entity as logical variable:
  - $\langle e_i, r_{i,j}, e_j \rangle \leftarrow \left(e_i \xrightarrow{r_{i,i+1}} e_{i+1} \xrightarrow{r_{i+1,i+2}} \cdots \xrightarrow{r_{j-1,j}} e_j \right)$ (3)
  - The right side is a **meta-path** connecting $\langle e_i, e_j \rangle$.
- A assumption for logical consistency:
  - *Under the same context (in the same passage), the definite relation between a pair of entities can be inferred from the contextual indirect one, or at least not logically contradict to it.*
- Eqn. (3) is weaker than Eqn. (2) in a segment of plain text, but can be further enhanced by negative candidates violating the logics.

Context: Economist: **(1)** A country's rapid emergence from an economic recession (r_1) requires **(2)** substantial new investment in that country's economy. Since **(3)** people's confidence in the economic policies of their country (r_2) is a precondition for **(2)** any new investment, **(4)** countries that put collective goals before individuals' goals (r_3) cannot **(1)** emerge quickly from an economic recession.

Question:

Which one of the following, if assumed, enables the economist's conclusion to be properly drawn?

Options:

A. People in **(4)** countries that put collective goals before individuals' goals (r_4) lack **(3)** confidence in the economic policies of their countries.

B. A country's economic policies are the most significant factor determining whether that country's economy will experience a recession.

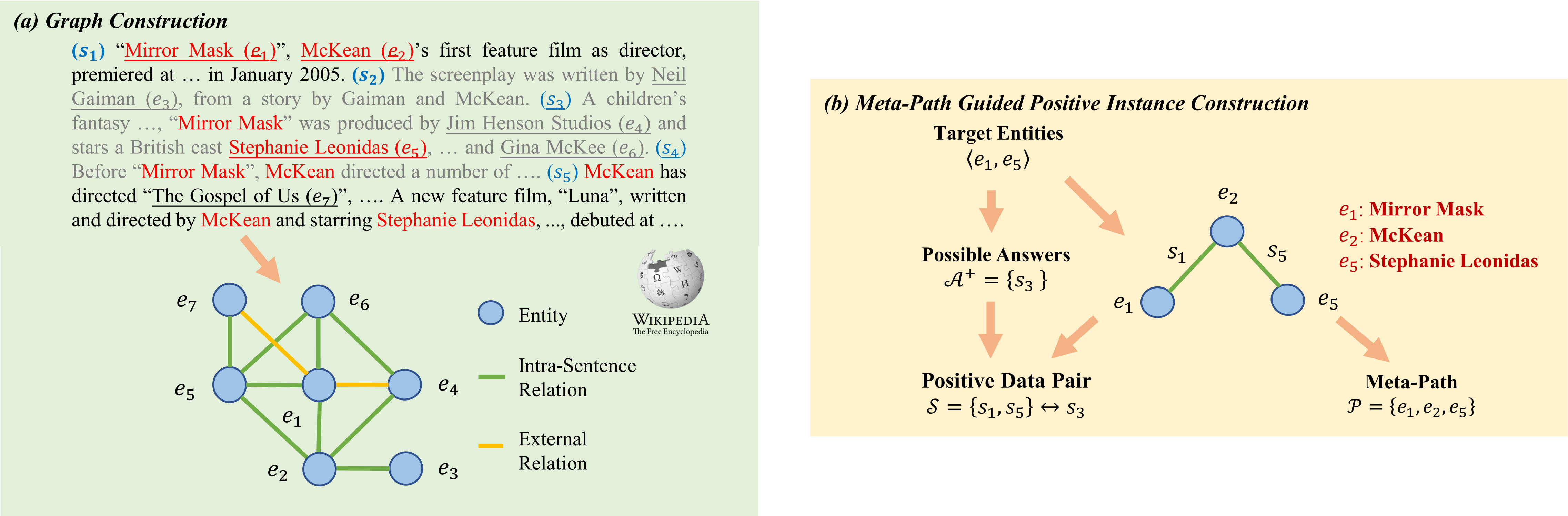
C. If the people in a country that puts individuals' goals first are willing to make new investments in their country's economy, their country will emerge quickly from an economic recession.

D. No new investment occurs in any country that does not emerge quickly from an economic recession.

Answer: A

Logic Structure: $(4) \xrightarrow{r_4} (3) \xrightarrow{r_2} (2) \xrightarrow{\bar{r}_1} (1) \Leftrightarrow (4) \xrightarrow{r_3} (1)$

Meta-Path Guided Positive Instance Construction



- $\{s_1, s_5\}$: The **director** *McKean* has **cooperated with** the **actor** *Stephanie Leonidas*. *Mirror Mask* **is directed by** *McKean*.
- No logical contradiction between $\{s_1, s_5\}$ and s_3.

Negative Instance Generation

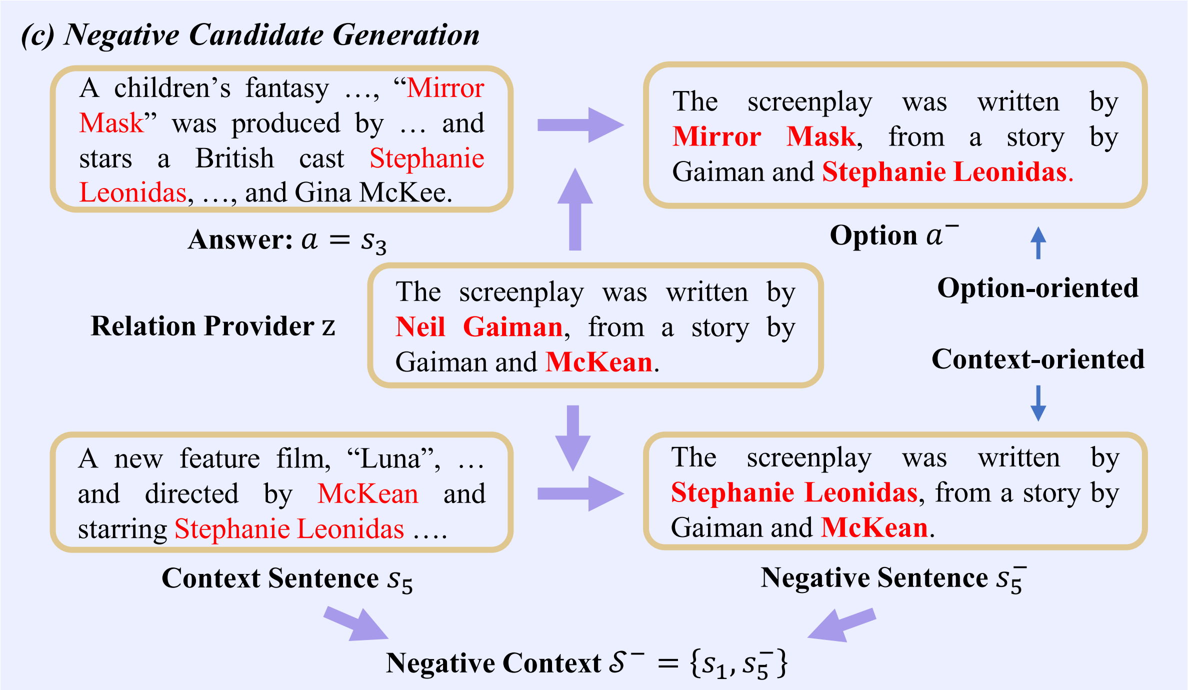
- Randomly sampling
  - Trivial solution by checking the involved entities or context.
- Modification of relations

  $\rightarrow$ Entity replacement
- Given z containing $\langle e_a, e_b \rangle$ as the relation provider, a containing $\langle e_i, e_j \rangle$ as the answer, the negative candidate can be obtained by replace $\langle e_a, e_b \rangle$ in z as $\langle e_i, e_j \rangle$, defined as:

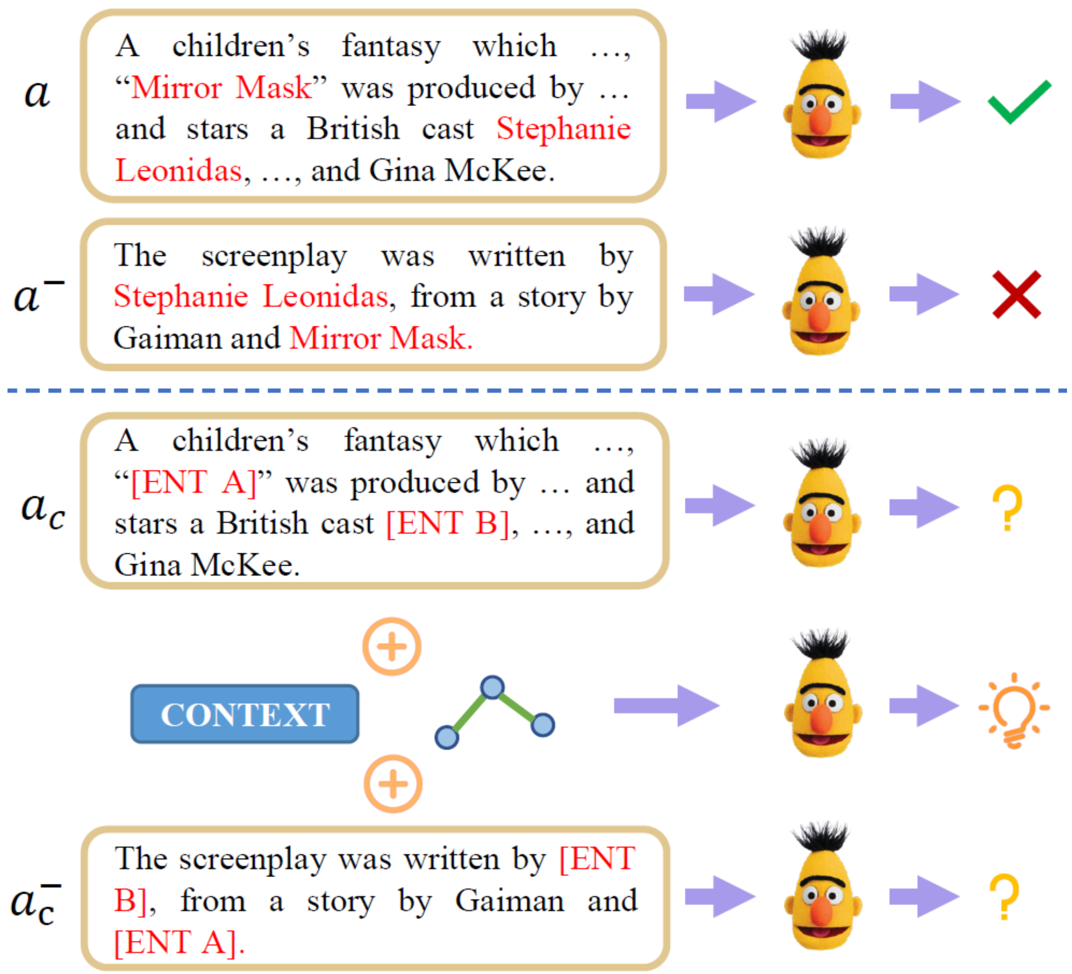
$$a^- = \text{Relation\_Replace}(z \rightarrow a)$$

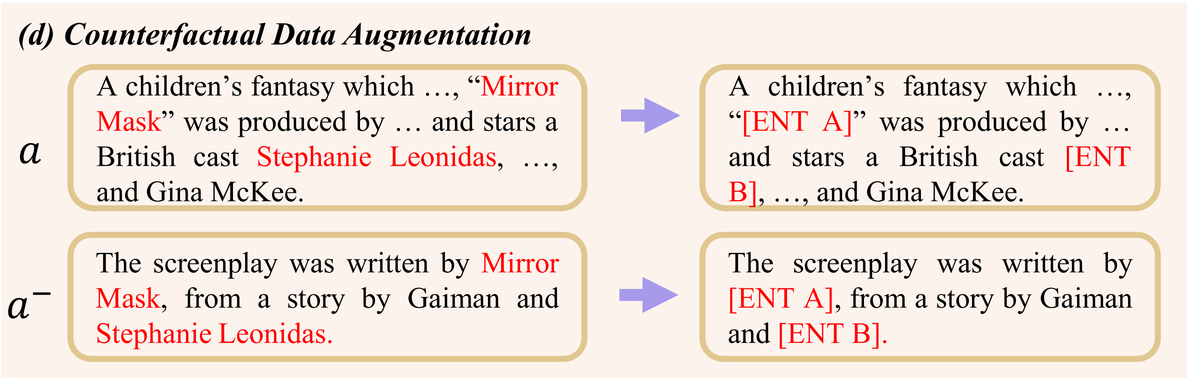
- For context-oriented negative instance generation:

$$s_i^- = \text{Relation\_Replace}(z \rightarrow s_i)$$



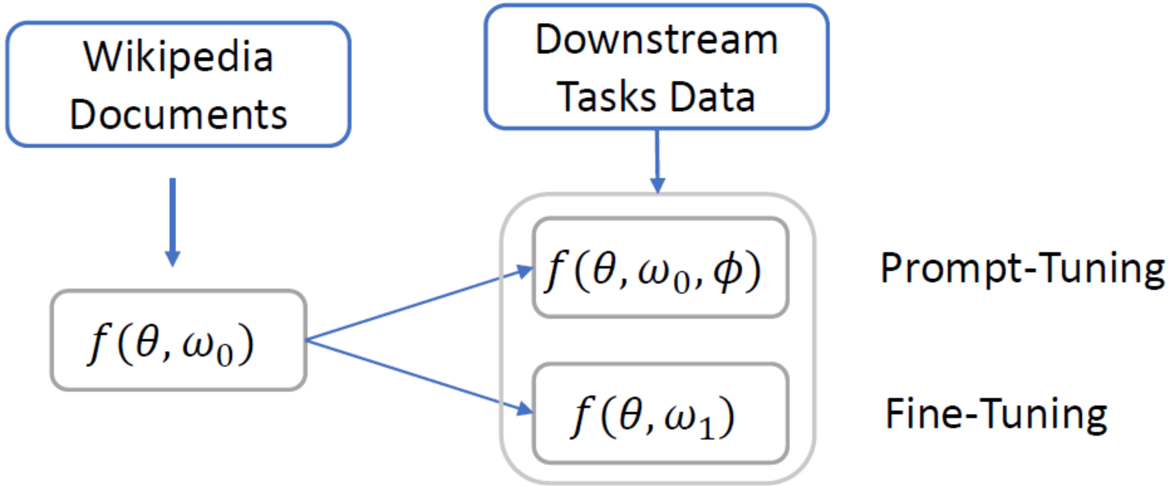
Counterfactual Data Augmentation





Training

- Contrastive Learning:
  - $\mathcal{L}_{OCL} = L(\mathcal{S}, a, \mathcal{A}^-)$
  - $\mathcal{L}_{CCL} = L(a, \mathcal{S}, \mathcal{C}^-)$
- Pre-training: $\mathcal{L} = \mathcal{L}_{OCL} + \mathcal{L}_{CCL} + \mathcal{L}_{MLM}$
- Fine-tuning: $\mathcal{L}_{QA} = -\log \frac{\exp f(P,Q,O_y)}{\sum_i \exp f(P,Q,O_i)}$
- Prompt-tuning:
  - Input sequence: $[Q, [\text{prefix}], O_i, P]$



Experiment Setup

- Backbone
  - RoBERTa-large (2080Ti * 4, 32 hours)
  - ALBERT-v2-xxlarge (Tesla T4 * 2, 3 days)
  - DeBERTa-v2-xlarge (A100 * 4, 20 hours)
  - DeBERTa-v2-xxlarge (A100 * 4, 1 day)
- Pre-training corpus: Wikipedia
- Dataset
  - ReClor
  - LogiQA
- Baseline
  - DAGN
  - Focal Reasoner
  - LReasoner

Experiment

- Overall performance
- Ablation study
- Performance with limited training data
- Effect of pre-training steps
- Linear probing
- Performance on DREAM

Experiments

Model / Dataset	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
RoBERTa	62.6	55.6	75.5	40.0	35.0	35.3
DAGN	65.2	58.2	76.1	44.1	35.5	38.7
DAGN (Aug)	65.8	58.3	75.9	44.5	36.9	39.3
LReasoner (RoBERTa)‡	64.7	58.3	77.6	43.1	—	—
Focal Reasoner	66.8	58.9	77.1	44.6	**41.0**	40.3
MERIt	66.8	59.6	78.1	45.2	40.0	38.9
MERIt + LReasoner	67.4	60.4	78.5	46.2	—	—
MERIt + Prompt	**69.4**	**61.6**	79.3	**47.8**	39.9	**40.7**
MERIt + Prompt + LReasoner	67.3	61.4	**79.8**	46.9	—	—
ALBERT	69.1	66.5	76.7	58.4	38.9	37.6
MERIt (ALBERT)	74.2	70.1	81.6	61.0	43.7	**42.5**
MERIt (ALBERT) + Prompt	**74.7**	**70.5**	**82.5**	**61.1**	**46.1**	41.7
max						
LReasoner (RoBERTa)	66.2	62.4	81.4	47.5	38.1	40.6
MERIt	67.8	60.7	79.6	45.9	**42.4**	41.5
MERIt + Prompt	**70.2**	**62.6**	80.5	**48.5**	39.5	**42.4**
LReasoner (ALBERT)	73.2	70.7	81.1	62.5	41.6	41.2
MERIt (ALBERT)	73.2	71.1	83.6	61.3	43.9	**45.3**
MERIt (ALBERT) + Prompt	**75.0**	**72.2**	**82.5**	**64.1**	**45.8**	43.8

Table 1: The overall results on ReClor and LogiQA. We adopt the **accuracy** as the evaluation metric and all the baselines are based on RoBERTa except specific statement. For each model we repeated training for 5 times using different random seeds and reported the average results. ‡: The results are reproduced by ourselves. *max*: The results of the model achieving the best accuracy on the test set.

Experiments

Model / Dataset	ReClor Dev	ReClor Test	ReClor Test-E	ReClor Test-H	LogiQA Dev	LogiQA Test
RoBERTa	62.6	55.6	75.5	40.0	35.0	35.3
DAGN	65.2	58.2	76.1	44.1	35.5	38.7
DAGN (Aug)	65.8	58.3	75.9	44.5	36.9	39.3
LReasoner (RoBERTa)‡	64.7	58.3	77.6	43.1	—	—
Focal Reasoner	66.8	58.9	77.1	44.6	**41.0**	40.3
MERIt	66.8	59.6	78.1	45.2	40.0	38.9
MERIt + LReasoner	67.4	60.4	78.5	46.2	—	—
MERIt + Prompt	**69.4**	**61.6**	79.3	**47.8**	39.9	**40.7**
MERIt + Prompt + LReasoner	67.3	61.4	**79.8**	46.9	—	—
ALBERT	69.1	66.5	76.7	58.4	38.9	37.6
MERIt (ALBERT)	74.2	70.1	81.6	61.0	43.7	**42.5**
MERIt (ALBERT) + Prompt	**74.7**	**70.5**	**82.5**	**61.1**	**46.1**	41.7
max						
LReasoner (RoBERTa)	66.2	62.4	81.4	47.5	38.1	40.6
MERIt	67.8	60.7	79.6	45.9	**42.4**	41.5
MERIt + Prompt	**70.2**	**62.6**	80.5	**48.5**	39.5	**42.4**
LReasoner (ALBERT)	73.2	70.7	81.1	62.5	41.6	41.2
MERIt (ALBERT)	73.2	71.1	83.6	61.3	43.9	**45.3**
MERIt (ALBERT) + Prompt	**75.0**	**72.2**	**82.5**	**64.1**	**45.8**	43.8

Table 1: The overall results on ReClor and LogiQA. We adopt the **accuracy** as the evaluation metric and all the baselines are based on RoBERTa except specific statement. For each model we repeated training for 5 times using different random seeds and reported the average results. ‡: The results are reproduced by ourselves. *max*: The results of the model achieving the best accuracy on the test set.

Experiments

Model / Dataset	ReClor				LogiQA	
	Dev	Test	Test-E	Test-H	Dev	Test
RoBERTa	62.6	55.6	75.5	40.0	35.0	35.3
DAGN	65.2	58.2	76.1	44.1	35.5	38.7
DAGN (Aug)	65.8	58.3	75.9	44.5	36.9	39.3
LReasoner (RoBERTa)‡	64.7	58.3	77.6	43.1	—	—
Focal Reasoner	66.8	58.9	77.1	44.6	**41.0**	40.3
MERIt	66.8	59.6	78.1	45.2	40.0	38.9
MERIt + LReasoner	67.4	60.4	78.5	46.2	—	—
MERIt + Prompt	**69.4**	**61.6**	79.3	**47.8**	39.9	**40.7**
MERIt + Prompt + LReasoner	67.3	61.4	**79.8**	46.9	—	—
ALBERT	69.1	66.5	76.7	58.4	38.9	37.6
MERIt (ALBERT)	74.2	70.1	81.6	61.0	43.7	**42.5**
MERIt (ALBERT) + Prompt	**74.7**	**70.5**	**82.5**	**61.1**	**46.1**	41.7
max						
LReasoner (RoBERTa)	66.2	62.4	81.4	47.5	38.1	40.6
MERIt	67.8	60.7	79.6	45.9	**42.4**	41.5
MERIt + Prompt	**70.2**	**62.6**	80.5	**48.5**	39.5	**42.4**
LReasoner (ALBERT)	73.2	70.7	81.1	62.5	41.6	41.2
MERIt (ALBERT)	73.2	71.1	83.6	61.3	43.9	**45.3**
MERIt (ALBERT) + Prompt	**75.0**	**72.2**	**82.5**	**64.1**	**45.8**	43.8

Table 1: The overall results on ReClor and LogiQA. We adopt the **accuracy** as the evaluation metric and all the baselines are based on RoBERTa except specific statement. For each model we repeated training for 5 times using different random seeds and reported the average results. ‡: The results are reproduced by ourselves. *max*: The results of the model achieving the best accuracy on the test set.

Experiments

Model / Dataset	ReClor Dev	ReClor Test	ReClor Test-E	ReClor Test-H	LogiQA Dev	LogiQA Test
RoBERTa	62.6	55.6	75.5	40.0	35.0	35.3
DAGN	65.2	58.2	76.1	44.1	35.5	38.7
DAGN (Aug)	65.8	58.3	75.9	44.5	36.9	39.3
LReasoner (RoBERTa)‡	64.7	58.3	77.6	43.1	—	—
Focal Reasoner	66.8	58.9	77.1	44.6	**41.0**	40.3
MERIt	66.8	59.6	78.1	45.2	40.0	38.9
MERIt + LReasoner	67.4	60.4	78.5	46.2	—	—
MERIt + Prompt	**69.4**	**61.6**	79.3	**47.8**	39.9	**40.7**
MERIt + Prompt + LReasoner	67.3	61.4	**79.8**	46.9	—	—
ALBERT	69.1	66.5	76.7	58.4	38.9	37.6
MERIt (ALBERT)	74.2	70.1	81.6	61.0	43.7	**42.5**
MERIt (ALBERT) + Prompt	**74.7**	**70.5**	**82.5**	**61.1**	**46.1**	41.7
max						
LReasoner (RoBERTa)	66.2	62.4	81.4	47.5	38.1	40.6
MERIt	67.8	60.7	79.6	45.9	**42.4**	41.5
MERIt + Prompt	**70.2**	**62.6**	80.5	**48.5**	39.5	**42.4**
LReasoner (ALBERT)	73.2	70.7	81.1	62.5	41.6	41.2
MERIt (ALBERT)	73.2	71.1	83.6	61.3	43.9	**45.3**
MERIt (ALBERT) + Prompt	**75.0**	**72.2**	**82.5**	**64.1**	**45.8**	43.8

Table 1: The overall results on ReClor and LogiQA. We adopt the **accuracy** as the evaluation metric and all the baselines are based on RoBERTa except specific statement. For each model we repeated training for 5 times using different random seeds and reported the average results. ‡: The results are reproduced by ourselves. *max*: The results of the model achieving the best accuracy on the test set.

Experiments

Model / Dataset	ReClor Dev	ReClor Test	ReClor Test-E	ReClor Test-H	LogiQA Dev	LogiQA Test
RoBERTa	62.6	55.6	75.5	40.0	35.0	35.3
DAGN	65.2	58.2	76.1	44.1	35.5	38.7
DAGN (Aug)	65.8	58.3	75.9	44.5	36.9	39.3
LReasoner (RoBERTa)‡	64.7	58.3	77.6	43.1	—	—
Focal Reasoner	66.8	58.9	77.1	44.6	**41.0**	40.3
MERIt	66.8	59.6	78.1	45.2	40.0	38.9
MERIt + LReasoner	67.4	60.4	78.5	46.2	—	—
MERIt + Prompt	**69.4**	**61.6**	79.3	**47.8**	39.9	**40.7**
MERIt + Prompt + LReasoner	67.3	61.4	**79.8**	46.9	—	—
ALBERT	69.1	66.5	76.7	58.4	38.9	37.6
MERIt (ALBERT)	74.2	70.1	81.6	61.0	43.7	**42.5**
MERIt (ALBERT) + Prompt	**74.7**	**70.5**	**82.5**	**61.1**	**46.1**	41.7
max						
LReasoner (RoBERTa)	66.2	62.4	81.4	47.5	38.1	40.6
MERIt	67.8	60.7	79.6	45.9	**42.4**	41.5
MERIt + Prompt	**70.2**	**62.6**	80.5	**48.5**	39.5	**42.4**
LReasoner (ALBERT)	73.2	70.7	81.1	62.5	41.6	41.2
MERIt (ALBERT)	73.2	71.1	83.6	61.3	43.9	**45.3**
MERIt (ALBERT) + Prompt	**75.0**	**72.2**	**82.5**	**64.1**	**45.8**	43.8

Table 1: The overall results on ReClor and LogiQA. We adopt the **accuracy** as the evaluation metric and all the baselines are based on RoBERTa except specific statement. For each model we repeated training for 5 times using different random seeds and reported the average results. ‡: The results are reproduced by ourselves. *max*: The results of the model achieving the best accuracy on the test set.

Experiments

Model	Dev	Dev (P.)	Test	Test (P.)
MERIt	66.8	69.4	59.6	61.6
- DA	63.0	64.5	57.9	59.8
+ DA^2	65.3	67.8	**60.2**	61.3
+ DA^3	66.2	68.0	59.3	**61.9**
- Option-oriented CL	63.8	65.4	58.9	61.5
- Context-oriented CL	64.0	66.5	58.8	60.2
- Meta-Path	64.8	65.1	58.0	60.8

Table 2: Performance comparisons on ReClor between different variants of MERIt. *DA* means data augmentation and *DA*^{N} refers to 1:N ratio of the original data to the augmented data. *P.* is short for *Prompt Tuning*.

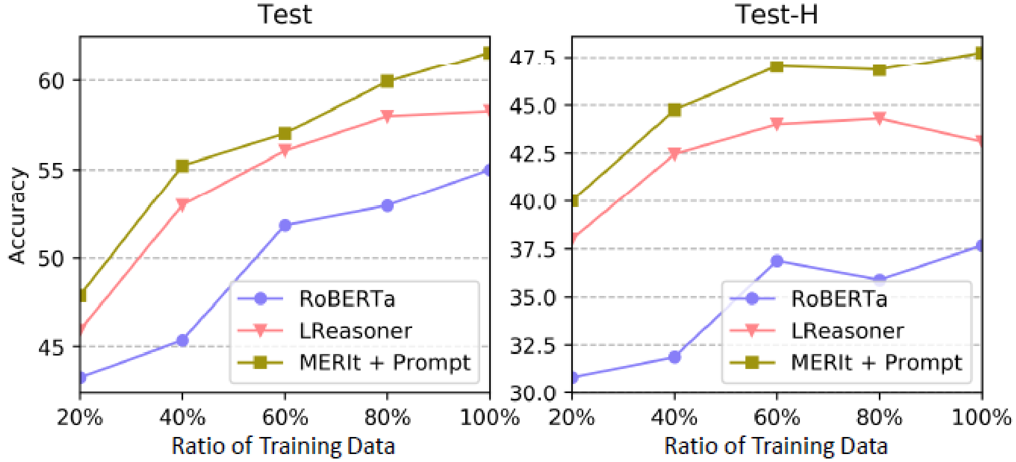


Figure 4: Results on the test set (left) and the test-H set (right) of ReClor.

Experiments

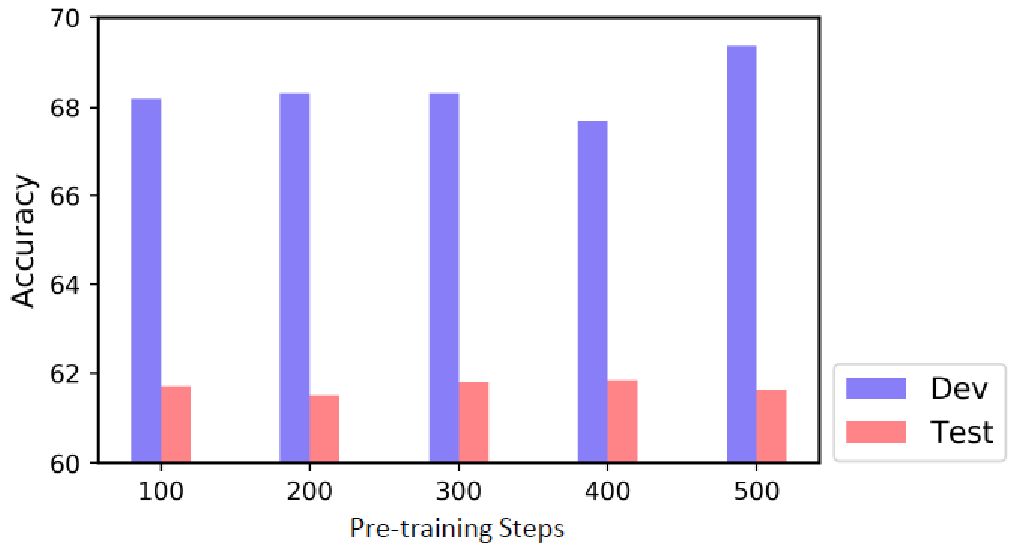


Figure 5: The prompt-tuning results on ReClor using the models pre-trained with different steps.

Model	Dev	Test	Test-E	Test-H
RoBERTa	35.8	35.7	44.5	28.8
MERIt (500 steps)	**39.0**	35.2	41.8	30.0
100 steps	37.5	**38.1**	**47.5**	30.6
200 steps	38.1	38.0	47.3	30.7
300 steps	37.4	36.4	43.6	30.7
400 steps	38.5	35.9	42.5	30.7
ALBERT	43.6	40.2	46.6	35.2
MERIt (ALBERT)	**46.3**	**44.6**	**51.8**	**38.9**

Table 4: Results of Linear Probing on ReClor.

Model	Dev	Test
RoBERTa	84.9	84.2
MERIt	**85.9**	**85.5**

Table 5: The accuracy of different models on DREAM dataset.

More Experiments

Model	Dev	Test	Test-E	Test-H
DeBERTa-v2-xlarge	76.7	71.0	83.8	60.9
MERIt (DeBERTa-v2-xlarge)	**78.0**	**73.1**	**86.2**	**64.4**
DeBERTa-v2-xxlarge	78.3	75.3	84.0	68.4
MERIt (DeBERTa-v2-xxlarge)	**80.6**	**78.1**	**84.6**	**72.9**

Table 6: Results on ReClor with DeBERTa as the backbone.

Leaderboard

Leaderboard of testing set of ReClor

Phase: Test Phase, Split: Test Split

Order by metric

B - Baseline   * - Private   V - Verified

Rank	Participant team	Test (↑)	Test-E (↑)	Test-H (↑)	NA (↑)	SA (↑)	S (↑)	W (↑)
1	Chitanda (MERIt-deberta-v2-xxlarge)	79.30	85.23	74.64	85.09	83.33	82.98	71.68
2	Knowledge Model Team (Knowledge model)	79.20	91.82	69.29	89.47	80.00	76.60	68.14
3	LReasoner Team (LReasoner ensemble)	76.10	87.05	67.50	80.70	80.00	76.60	67.26
4	RainaCUED (ELECTRA and ALBERT)	71.00	83.41	61.25	77.19	73.33	68.09	63.72
5	Logical QA NLI (xlnet-large-uncased [extended)	69.30	84.77	57.14	79.82	60.00	68.09	69.03

Conclusion & Future Works

- Reasoning?
  - Relation?
- More pretext task
- Interpretable reasoning steps
- Causality

Algorithm for Meta-Path Extraction

Algorithm 1 The DFS algorithm to obtain the meta-paths.

Input: The graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$; The sentences of the document $\mathcal{D} = \{s_1, \cdots, s_m\}$; The entity set of the i-th sentence $\mathcal{V}_i$;

Output: $\mathcal{P}$, $\mathcal{S}$, and $\mathcal{A}^+$;

for each $(e_i, e_j) \in \mathcal{V} \times \mathcal{V}$ and $i \neq j$ **do**
    $\mathcal{A}^+ = \{s_k | e_i \in \mathcal{V}_k, e_j \in \mathcal{V}_k\}$;
    $\mathcal{D}' = \mathcal{D} \setminus \mathcal{A}^+$;
    cond, $\mathcal{P}, \mathcal{S} \leftarrow$ DFS$(e_i, \{e_i\}, \varnothing, e_j, \mathcal{G}, \mathcal{D}')$;
    if cond is TRUE and $\mathcal{A}^+$ is not $\varnothing$ **then**
        return $\mathcal{A}^+, \mathcal{P}, \mathcal{S}$;
    end if
end for
return $\varnothing, \varnothing, \varnothing$;

function DFS$(e_i, \mathcal{P}', \mathcal{S}', e_d, \mathcal{G} = (\mathcal{E}, \mathcal{V}), \mathcal{D}')$
    if $e_i = e_d$ **then**
        return TRUE, $\mathcal{P}', \mathcal{S}'$;
    end if
    for each $(e_j, s_k) \in \mathcal{V} \times \mathcal{D}'$ and $(e_i, e_j) \in \mathcal{E}, e_j \in \mathcal{V}_k$ **do**
        $\mathcal{G}' = (\mathcal{E}, \mathcal{V} \setminus \{e_j\})$;
        $\mathcal{P}'' = \mathcal{P}' \cup \{e_j\}$;
        if $e_i \in \mathcal{V}_k$ **then**
            $\mathcal{D}'' = \mathcal{D}' \setminus \{s_k\}$;
            $\mathcal{S}'' = \mathcal{S}' \cup \{s_k\}$;
        else
            $\mathcal{D}'' = \mathcal{D}', \mathcal{S}'' = \mathcal{S}'$;
        end if
        return DFS$(e_j, \mathcal{P}'', \mathcal{S}'', e_d, \mathcal{G}', \mathcal{D}'')$;
    end for
    return FALSE, $\varnothing, \varnothing$;
end function

Case Study for Data Construction

Example 1 (Option-based CL)

Context:

Napoleon appointed his brother Louis Bonaparte to the Kingdom of Holland in May 1806. The Dutch rebellion first broke out in Amsterdam on 14–15 November.

Negative Candidates:

- Since their trade was badly damaged by Napoleon's Continental System, the French people were ready to throw off the Dutch yoke.
- However, on 9 July 1810, the French emperor extinguished the kingdom and annexed the Dutch to the Napoleon.
- Depressed by the loss of his son in Napoleon, the French civil leader Dutch responded ineffectively to the crisis.

Answer:

The Dutch contributed only 17,300 soldiers to Napoleon's armies in 1811–1813, but their severe casualties in the French invasion of Russia shocked the population.

A Counterfactual Sample of Example 1

Context:

The Din rebellion first broke out in Amsterdam on 14–15 November. Bihar appointed his brother Louis Bonaparte to the Kingdom of Holland in May 1806 .

Negative Candidates:

- Since their trade was badly damaged by French's Continental System , the Din people were ready to throw off the Bihar yoke .
- In early November, Din corps commander Ferdinand von Wintzingerode sent a 3,500-man "Streifkorps" led by Alexander Khristoforovich Benckendorff into Bihar.
- In early November, Bihar corps commander Ferdinand von Wintzingerode sent a 3,500-man "Streifkorps" led by Alexander Khristoforovich Benckendorff into Din.

Answer:

The Dutch contributed only 17,300 soldiers to Napoleon's armies in 1811–1813, but their severe casualties in the French invasion of Russia shocked the population.

Case Study for Data Construction

Example 2 (Context-oriented CL)

Context:

Napoleon appointed his brother Louis Bonaparte to the Kingdom of Holland in May 1806. The Dutch rebellion first broke out in Amsterdam on 14–15 November.

Negative Contexts:

- Depressed by the loss of his son in Napoleon, the French civil leader Kingdom of Holland responded ineffectively to the crisis. The Dutch rebellion first broke out in Amsterdam on 14–15 November.
- Since their trade was badly damaged by Kingdom of Holland's Napoleon, the Dutch people were ready to throw off the French yoke. The Dutch rebellion first broke out in Amsterdam on 14–15 November.
- Depressed by the loss of his son in Russia, the Napoleon civil leader Kingdom of Holland responded ineffectively to the crisis. The Dutch rebellion first broke out in Amsterdam on 14–15 November ..

Answer:

The Dutch contributed only 17,300 soldiers to Napoleon's armies in 1811–1813, but their severe casualties in the French invasion of Russia shocked the population.

A Counterfactual Sample of Example 2

Context:

Bihar appointed his brother Louis Bonaparte to the Kingdom of Holland in May 1806. The Din rebellion first broke out in Amsterdam on 14–15 November.

Negative Contexts:

- The Din rebellion first broke out in Amsterdam on 14–15 November. Since their trade was badly damaged by Kingdom of Holland's Continental System, the Din people were ready to throw off the Bihar yoke.
- The Din rebellion first broke out in Amsterdam on 14–15 November. Depressed by the loss of his son in Kingdom of Holland, the French civil leader Bihar responded ineffectively to the crisis.
- Since their trade was badly damaged by Bihar's Continental System , the Kingdom of Holland people were ready to throw off the French yoke. The Din rebellion first broke out in Amsterdam on 14–15 November.

Answer:

The Din contributed only 17,300 soldiers to Bihar’s armies in 1811–1813, but their severe casualties in the French invasion of Russia shocked the population.