Winnowing Knowledge for Multi-choice Question Answering

Yeqiu Li[1], **Bowei Zou**[2], **Zhifeng Li**[1], **Ai Ti Aw**[2], **Yu Hong**[*1], **Qiaoming Zhu**[1]
[1] School of Computer Science and Technology, Soochow University
[2] Institute for Infocomm Research, A*STAR, Singapore
{unlimitedaki, tianxianer}@gmail.com, li_zaaachary@163.com
{zou_bowei,aaiti}@i2r.a-star.edu.sg, qmzhu@suda.edu.cn

Abstract

We tackle multi-choice question answering. Acquiring related commonsense knowledge to the question and options facilitates the recognition of the correct answer. However, the current reasoning models suffer from the noises in the retrieved knowledge. In this paper, we propose a novel encoding method which is able to conduct interception and soft filtering. This contributes to the harvesting and absorption of representative information with less interference from noises. We experiment on CommonsenseQA. Experimental results illustrate that our method yields substantial and consistent improvements compared to the strong Bert, RoBERTa and Albert-based baselines. [1]

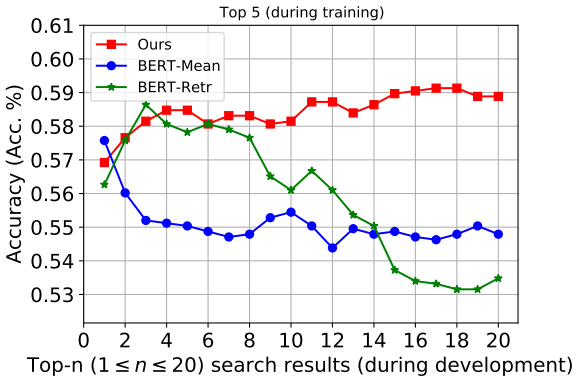

Figure 1: MQA performance changes when the number of search results is increased (during development).

1 Introduction

Multi-choice question answering (MQA for short) is required to select an answer from a set of candidate options when given a question (Rajani et al., 2019). The task is slightly different from multi-choice reading comprehension which provides the passage containing background knowledge for reasoning (Richardson et al., 2013). Frankly, due to the lack of commonsense knowledge, MQA is more challenging. For example, it appears to be difficult for MQA to determine the true answer in the following case, where the commonsense knowledge regarding "*island country*" deterministically contributes to reasoning, though such knowledge is not offered in any form by default:

(1) **Question**: *What island country is ferret popular*?
Options: [*own home*] [*hutch*] [*outdoors*] [*north Carolina*] [*Great Britain*]
Answer: [***Great Britain***]

Therefore, actively acquiring the closely related commonsense knowledge from external sources is crucial for MQA. Previous studies either retrieve the commonsense knowledge from Wikipedia (Lv et al., 2020) and ConceptNet (Lin et al., 2019; Wang et al., 2020; Bian et al., 2021), or hunt supportive evidence in the unstructured Internet data (Emami et al., 2018). Bringing the retrieved knowledge into the encoding process of questions and options has been proven effective in strengthening MQA.

We follow the previous work to perform MQA using external knowledge bases. Information retrieval is utilized for knowledge acquisition as usual. The difference is that we intend to enhance the joint encoding of question, option and knowledge by soft filtering and interception.

The filter is used to shield the encoder from the negative influence of the mistakenly-retrieved irrelevant or unrepresentative knowledge (called noise hereafter). It is motivated by our findings that purposefully retrieving a larger number of knowledge items actually results in performance degradation. As illustrated in Figure 1, the performance curve of the retrieval-based MQA model (green curve) shows a trend of fluctuating downward when the number of the adopted highly-ranked search results

*Corresponding author
[1] https://github.com/unlimitedaki/HeadHunter

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1157–1165
November 7–11, 2021.

is increased. It is most likely caused by the noises that sneak into the list of retrieval knowledge items.

Instead of thoroughly filtering noises out from the retrieval knowledge, we perform soft filtering which retains noises but merely assigns negligible attention weights to them. On the basis, we develop an interceptor to "eavesdrop" on the encoding channel of knowledge items, and salvage the recyclable latent information hidden in them. It is motivated by the fact that part of the content of a certain less-relevant knowledge item is probably informative. See the knowledge item in (2), in which the constituent "*a large island*" is informative and recyclable (as it even bridges the key words "*Great Britain*" in the relevant knowledge and "*island country*" in the question). To recycle available evidence in retrieval knowledge, the interceptor conducts information fusion among them, conditioned on the assignment of interactive attention to them.

(2) **Question**: *What island country is ferret popular*?
Relevant knowledge: *You are likely to find a ferret in Great Britain*.
Knowledge item: *Great Britain is* ***a large island***.

We implement the interceptor and soft filter by self-attention network and attention pooling layer, which are collectively referred to as "Headhunter". We couple a certain pre-trained model with Headhunter for encoding, and deploy them along with ElasticSearch in the commonly-used two-stage MQA architecture. We experiment over the CommonsenseQA dataset (Talmor et al., 2019). Experimental results show that Headhunter yields significant performance gains all along when coupled with different pre-trained models, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and Albert (Lan et al., 2020). Besides, the case of combining Albert and Headhunter achieves better performance than most state-of-the-art models, and it is ranked second on the CommonsenseQA leaderboard for track 1. The developed results show that the performance advantages can be safely attributed to the constraint on the use of less-relevant knowledge in Headhunter. To some extent, it successfully avoids the severe performance degradation when a larger set of qualified and less-qualified commonsense knowledge is taken (see the red curve in Figure 1).

Question	*What island country is ferret popular?*
Answer	*Great Britain*
Attention	**Knowledge**
$\breve{\alpha}_1 = 0.949$	*You are likely to find a ferret in Great Britain*
$\breve{\alpha}_2 = 0.017$	*Great Britain is a country*
$\breve{\alpha}_3 = 0.017$	*Great Britain is a large island*
$\breve{\alpha}_4 = 0.017$	*A ferret is an animal.*

Table 1: An example that attention pooling helps to highlight the representative commonsense knowledge.

2 Approach

2.1 Headhunter's Interceptor

We utilize the self-attention model (Vaswani et al., 2017) as the interceptor. Assume H is an $n \times m$ matrix, in which each row corresponds to a hidden state vector h_i. Thus we compute the attention weights $\mathcal{A}$ at the matrix level for all the hidden states ($\forall h_i$) in H: $\mathcal{A}$=softmax($QK^\top$), where Q and K serve as the matrices transformed from H, and they are computed with nonlinear activation functions using different parameters.

In terms of this computation algorithm, the i-th row in $\mathcal{A}$ forms the attention vector α_i of the hidden state h_i, recording the attention weights of h_i upon all the other hidden states and itself. Thus if the attention weights can be imagined as the measures of relevance degrees, we are able to intercept the relevant information from other hidden states and bring that into h_i. We do so by accumulating the attentively-weighted hidden states, as that has been accomplished in the self-attention model: $\breve{h}_i$=$\alpha_i V$, where V is transformed from H by nonlinear activation function. This operation is carried out for all hidden states in H by the calculation of $\breve{H}$=$\mathcal{A}V$.

From here on, we specify that each h_i in H has been encoded as the hidden state vector that contains the latent information of a piece of commonsense knowledge (see Section 2.3). Thus, by the attention modeling mentioned above, each hidden state $\breve{h}_i$ in $\breve{H}$ intercepts and absorbs the relevant information from other commonsense knowledge, regardless of whether the knowledge is relevant or less-relevant.

2.2 Headhunter's Soft Filter

Attention pooling layer is used as the filter. It only comes into play when positioned behind the self-attention network. Given the attention matrix $\mathcal{A}$, we pool the attention for each column of $\mathcal{A}$. Softmax normalization is used among all columns. Specially, the pooled attention for the j-th column is

calculated as: $\check{\alpha}_j$=softmax($\sum_{i=1}^{n} \alpha_{i,j}$), where n denotes the number of rows in $\mathcal{A}$ (which is equivalent to the number of hidden states in H).

Theoretically speaking, the resultant $\check{\alpha}_j$ pools the attention of all hidden states in H upon a certain hidden state h_j. Therefore, it is able to reflect the global representativeness of h_j. Table 1 shows an example regarding the attention pooling, where the most representative commonsense knowledge is assigned with a significantly higher value of $\check{\alpha}_j$. Using the attention pooling layer, we can perform soft filtering on H, highlighting the representative hidden states (with higher $\check{\alpha}_j$) and eclipsing the unrepresentative (with lower $\check{\alpha}_j$). Soft filtering can be carried out by multiplying H by $\check{\alpha}$.

In Headhunter, interception and soft filtering are performed successively: interception first, then filtering. Thus, the objects need to be dealt with during soft filtering are the mutually-intercepted hidden states $\check{H}$ instead of the original H. Thus, the final output of Headhunter is computed as: $\check{\mathbb{H}}$=$\check{\alpha}\check{H}$.

2.3 Two-stage MQA Using Headhunter

We build a two-stage MQA system which comprises knowledge acquisition and reasoning modules. Headhunter is used in the reasoning stage.

For **knowledge acquisition**, we extract 705,647 sentence-level knowledge items from the Open Mind Common Sense corpus (Singh et al., 2002). On the basis, we index all the knowledge items by Elastic Search engine. Given an MQA example (i.e., one question plus five options), we formulate a query by concatenating the question and one of the options. As a result, we obtain 5 queries in total for each MQA example. For every query, we apply Elastic Search engine to retrieve knowledge, and rank the search results in descending order of relevance. Top-n highly-ranked search results are retained, and they will be considered as the available commonsense knowledge for reasoning (i.e., n knowledge items per pair of question and option).

During **reasoning**, we use the pre-trained model (e.g., Albert) and Headhunter for encoding. Besides, a fully-connected layer with softmax normalization is used for predicting the answer.

Given a group of question q, option o and knowledge k, we feed them into the pre-trained language model in terms of the following structure:

$$[CLS]\ q\ [SEP]\ o\ [SEP]\ k\ [SEP]$$

The transformers deployed in the pre-trained model (Vaswani et al., 2017) facilitates the interaction and fusion of the input q, o and k, and integrates their information all into the real-valued m-dimensional vector $[CLS]$. We employ the vector $[CLS]$ as the knowledge-aware representation of q and o. In this way, we obtain n $[CLS]$s for each pair of question and option, conditioned on the top-n retrieved knowledge items. Using these $[CLS]$s, we form the $n \times m$ input matrix H of Headhunter, where each $[CLS]$ acts as a row in H. On the basis, we transform H into the mutually-intercepted representation $\check{H}$ by Headhunter's interceptor, and further transform $\check{H}$ into the final representation $\check{\mathbb{H}}$ by Headhunter's filter (Section 2.2).

We feed $\check{\mathbb{H}}$ into the fully-connected layer, so as to estimate the probability that the corresponding option may be the answer: y=$w\check{\mathbb{H}}$+b, where $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$ stand for trainable parameters. Note that given a question, we perform retrieval, encoding and headhunting for the five options respectively. This causes five unique prediction processes, yielding five prediction results. Thus, we use softmax normalization over the predictions, so as to select the most probable option as the answer.

3 Experimentation

3.1 Dataset, Hyperparameter and Evaluation

We experiment on CommonsenseQA (Talmor et al., 2019), a dataset containing 12,102 MQA examples. We use 9,741 examples in it for training, 1,221 for development and 1,140 for test. The knowledge base we use is taken from Open Mind Common Sense (Singh et al., 2002) which comprises a large number of sentence-level commonsense knowledge items obtained by crowdsourcing.

Our best model employ Albert-xxlarge as the basic encoder. During training, the maximum length of the input sequence is set to 80. The batch size is set to 1 and the gradient accumulation step is set to 20. The learning rate is set to 1e-5. The dropout rate is set to 0.1. All the considered models are trained for 5 epochs. The number n of knowledge-oriented search results is an additional hyperparameter. We set n to 8 during training and 7 during development in our best model. Accuracy (*Acc.*) is used as the evaluation metric. The loss function during training is Cross-Entropy.

3.2 Baselines and Comparisons

We consider three baselines, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and Albert (Lan et al., 2020). They are connected with

Baselines	Acc.	+Headhunter	Acc.
BERT-base	55.4	BERT-base+Headhunter	60.0
BERT-large	60.2	BERT-large+Headhunter	61.4
RoBERTa-base	58.6	RoBERTa-base+Headhunter	67.3
RoBERTa-large	74.0	RoBERTa-large+Headhunter	77.6
Albert-xxlarge	79.4	Albert-xxlarge+Headhunter	83.3

Table 2: Comparison to baseline models.

fully-connected layer though without headhunting. All of them are retrained and fine-tuned.

In addition, we compare with two groups of state-of-the-art MQA models.

Group 1 includes RoBERTa and Albert which operate without using commonsense knowledge. In addition, we take the enlarged version of RoBERTa, as well as the optimized Albert by ensemble learning. Moreover, Zhu et al. (2020)'s FreeLB is considered which enhances RoBERTa-large by adversarial training. None of the MQA models in this group had used commonsense knowledge.

Group 2 comprises KE, KEDGN and DREAM, all of which use RoBERTa for encoding. In particular, KE conducts transfer learning on Open Mind Common Sense and fine-tunes RoBERTa on CommonsenseQA. It additionally retrieves supportive evidence from Wikipedia for reasoning. KEDGN embeds RoBERTa into the Dual Graph Network. DREAM is similar to KE but uses ElasticSearch for knowledge acquisition.

The models in Groups 1&2 have made their mark on the official CommonsenseQA leaderboard[2], settling in track 1 where ConceptNet (Speer et al., 2017) is unavailable by default. We list the reported performance for comparison. Although performing better, the highly-ranked MQA models(Lv et al., 2020; Xu et al., 2020) in the other track are not considered for comparison. It is because ConceptNet is available there and, more importantly, the 5-way MQA instances in CommonsenseQA (experimental dataset) are created using 4-node subgraphs in ConceptNet and a manually generated distractor answer for each. This potentially reduces the problem to the 4-way MQA.

3.3 Results and Analysis

We evaluate the performance of baselines and when Headhunter is connected with them. Table 2 shows the performance on the development set when convergence is persistent. It can be observed that Headhunter yields substantial improvements all the time.

[2] https://www.tau-nlp.org/csqa-leaderboard

Group	Model	Dev	Test
Group 1	RoBERTa-large (single)	78.5	72.1
	RoBERTa+FreeLB (single)	78.8	72.2
	RoBERTa+FreeLB (ensemble)	-	73.1
	Albert (single)	81.2	73.5
	Albert (ensemble)	**83.7**	76.5
Group 2	RoBERTa+DREAM (single)	81.6	73.3
	RoBERTa+KE (single)	78.7	73.3
	RoBERTa+KEDGN (single)	80.4	74.4
Ours	Albert+Headhunter (single)	83.3	**78.4**

Table 3: Comparison to the state-of-the-art models.

Compared to the previous work, we achieve competitive performance on both development and test sets, as shown in Table 3. More importantly, our method obtains relatively robust performance, yielding less performance degradation when the development process is switched to the test. Frankly, our best performance (78.4%) is slightly lower than that (79.1%) of Khashabi et al. (2020)'s UNIFIEDQA, the top-ranked model on the leaderboard of track 1, which is sophisticated (11B parameters) and trained on eight QA datasets. Nevertheless, our model is vest-pocket (283M parameters) due to the ease of reproduction and training with less data.

Ablation Study We study the contribution of Headhunter's interceptor and soft filter in Figure 1. Bert-Retr (green curve) refers to the traditional retrieval-based approach, which concatenates all retrieved results after the question and options. Bert-Mean (blue curve) applies Headhunter's interceptor but connected with a mean pooling layer. As shown in Figure 1, continuous improvement has been achieved by Headhunter as the number of retrieved results increases, demonstrating that Headhunter can effectively shield the noises from retrieved results. We can also observe that the soft filter plays a crucial role in recycling information from all retrieved results, which achieves much better performance than mean pooling.

Another finding is that, during training, the setting of the number n of the retrieved knowledge items significantly affects the performance when other hyperparameters remain unchanged. In Appendix A., we exhibit a variety of performance curves corresponding to different numbers of knowledge items. Figure 1 is a diagram taken from the appendix.

Cost-effectiveness Analysis The utilization of a large number of external knowledge items (i.e., the

ones retrieved by ElasticSearch) for encoding unavoidably results in a time-consuming training process. For example, we spent about 26min[3] on training the BERT-based baseline using 5 knowledge items for each MQA case but, by contrast, 1h41min when 20 knowledge items (per MQA case) are used. Similarly, we necessarily spend much more time on development. It may be acceptable if MQA performance appears to be better. However, the opposite is true.

Error Analysis We investigate the errors occurring during development. Our best model (Albert plus Headhunter) is considered in the investigation. We randomly select 100 errors from those occurring in the development process. In terms of the distinctive properties, we divide the errors into 5 categories:

- **Indistinguishable** error refers to the MQA case in which some candidate options are less distinguishable from each other. It is observed that the common errors are caused by the difficulty of making a distinction between indistinguishable options, such as "*happiness*" and "*satisfaction*".
- **Out-of-vocabulary** emerges when a candidate option is not included in the commonsense knowledge base or there is lack of intrinsically relevant knowledge to the question.
- **Unreasonable** error occurs when the encoder fails to predict the correct answer, even though the top-priority knowledge does serve as the most reasonable evidence for reasoning.
- **Less grounded** problem happens when the reliable knowledge items fail to be included in the top-n search results, even if they do exist in the commonsense knowledge base.
- Within the sampled data, we find 5 cases which were obviously **mislabeled**.

Appendix B. will show the details of examples which correspond to five types of errors.

4 Conclusion

We develop a vest-pocket model to squeeze reliable information out of commonsense knowledge as completely as possible. It is proven beneficial to MQA performance when cooperating with BERT, RoBERTa and Albert-based encoders. Error analysis demonstrates that a critical bottleneck lies in the disambiguation towards indistinguishable options. In the future, we will study the dictionary-based disambiguation approach, detecting and representing the most distinct aspects of words in terms of definitions. Moreover, a multi-task learning architecture will be developed, where knowledge acquisition, word sense disambiguation and MQA share the encoding channels of both general and distinctive word senses (named entities are not included).

[3]We run the models on an NVIDIA Tesla V100 SXM2 16GB GPU (Volta microarchitecture).

Acknowledgement

We thank all reviewers for their insightful comments, as well as the great efforts our colleagues have made so far. This work is supported by the national Natural Science Foundation of China (NSFC) and Major National Science and Technology project of China, via Grant Nos.62076174, 61836007, 2020YBF1313601.

References

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12574–12582. AAAI Press.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1949–1958. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format

boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8449–8456. AAAI Press.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4932–4942. Association for Computational Linguistics.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 193–203. ACL.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings*, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro A. Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4129–4140. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing context into knowledge graph for commonsense reasoning. *CoRR*, abs/2012.04808.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Problem	#Examples
Indistinguishable	54
Out-of-vocabulary	24
Unreasonable	12
Less grounded	5
Mislabeled	5

Table 4: Statistics for five categories of MQA errors.

Appendix.

A Knowledge Amount Analysis

The number n of knowledge-oriented search results is an additional hyperparameter. If an enormous amount of knowledge items are retrieved, there will be more noises brought into the encoding and reasoning processes. This definitely imposes unbearable pressure upon Headhunter, resulting in a time-consuming training process or even system breakdown. On the contrary, if a few cases are considered, some potentially valuable knowledge may be missed. We set n to 8 during training and 7 during development. Undoubtedly, the performance changes when n is set to different values. In appendix B., we exhibit the changing trends.

The proposed auxiliary model, Headhunter, is able to shield the encoder from the misleading of noises, and it facilitates the salvage of "recyclable" knowledge in noises. As a result, Headhunter helps to improve the performance of knowledge-based MQA. Besides, it obtains a relatively rapid convergence rate with the change of hyperparameter n (i.e, number of the retrieved knowledge items).

Figure 2 shows the changing trend of performance ($Acc.$) when different n are used during development. Each diagram in Figure 2 is obtained when a fixed number of knowledge items are used in the 5-epoch training process. It can be observed that, in most cases, Headhunter causes substantial performance improvement when it cooperates with the BERT-based baseline. Meanwhile, by Headhunter, the changing trend of performance comes to be plain at an earlier time.

B Error Analysis

We analyze the errors made by our best joint model (i.e., Albert coupling with Headhunter), so as to reveal the challenges we will meet in the future. We randomly select 100 errors from those occurring in the development process. In terms of the distinctive properties, we divide the errors into 5 categories, including 1) indistinguishable, 2) out-of-vocabulary, 3) unreasonable, 4) less grounded and 5) mislabeled. Table 4 shows the statistics of the sampled errors for each category.

Question: She was always helping at the senior center, it brought her what? **Ground truth**: happiness (Knowledge: Sometimes helping someone causes happiness.) **Prediction**: satisfaction (Knowledge: Sometimes helping someone causes satisfaction.)
Question: Crabs live in what sort of environment? **Ground truth**: saltwater (Knowledge: You are likely to find a crab in saltwater.) **Prediction**: bodies of water (Knowledge: You are likely to find a crab in bodies of water.)

Table 5: Examples of indistinguishable MQA errors.

Question: What is someone who isn't clever, bright, or competent called? **Ground truth**: stupid **Retrieved results** include "Situation: I am clever."| "Clever people are unpredictable."| "horses are clever animals."
Question: Where can you put a picture frame when it's not hung vertically? **Ground truth**: table **Retrieved results** include "picture description: Dining table."| "picture description: Table tennis paddle."| "picture description: A table tennis paddle."

Table 6: MQA errors emerge when out-of-vocabulary knowledge is encountered.

Question: Stabbing to death of a person is what sort of way to die? **Ground truth**: gruesome (Knowledge: The effect of stabbing to death is gruesome.) **Prediction**: killing (Knowledge: stabbing to death is for killing.)
Question: Where could you find hundreds of thousands of home? **Ground truth**: city or town (Knowledge: You are likely to find a home in a city or town.) **Prediction**: apartment building (Knowledge: You are likely to find a home in an apartment building.)
Question: Where would you find a basement that can be accessed with an elevator? **Ground truth**: office building (Knowledge: You are likely to find a basement in an office building.) **Prediction**: own house (Knowledge: You are likely to find a basement in your own house.)

Table 7: Examples of unreasonable errors.

Question: A beaver is known for building prowess, their supplies come from where? **Ground truth**: wooded area **Relevant knowledge** : You are likely to find a beaver in a wooded area. **Selected knowledge**: Trees create a wooded area.

Table 8: Examples of less-grounded errors.

Indistinguishable error refers to the MQA case in which some candidate options are less distin-

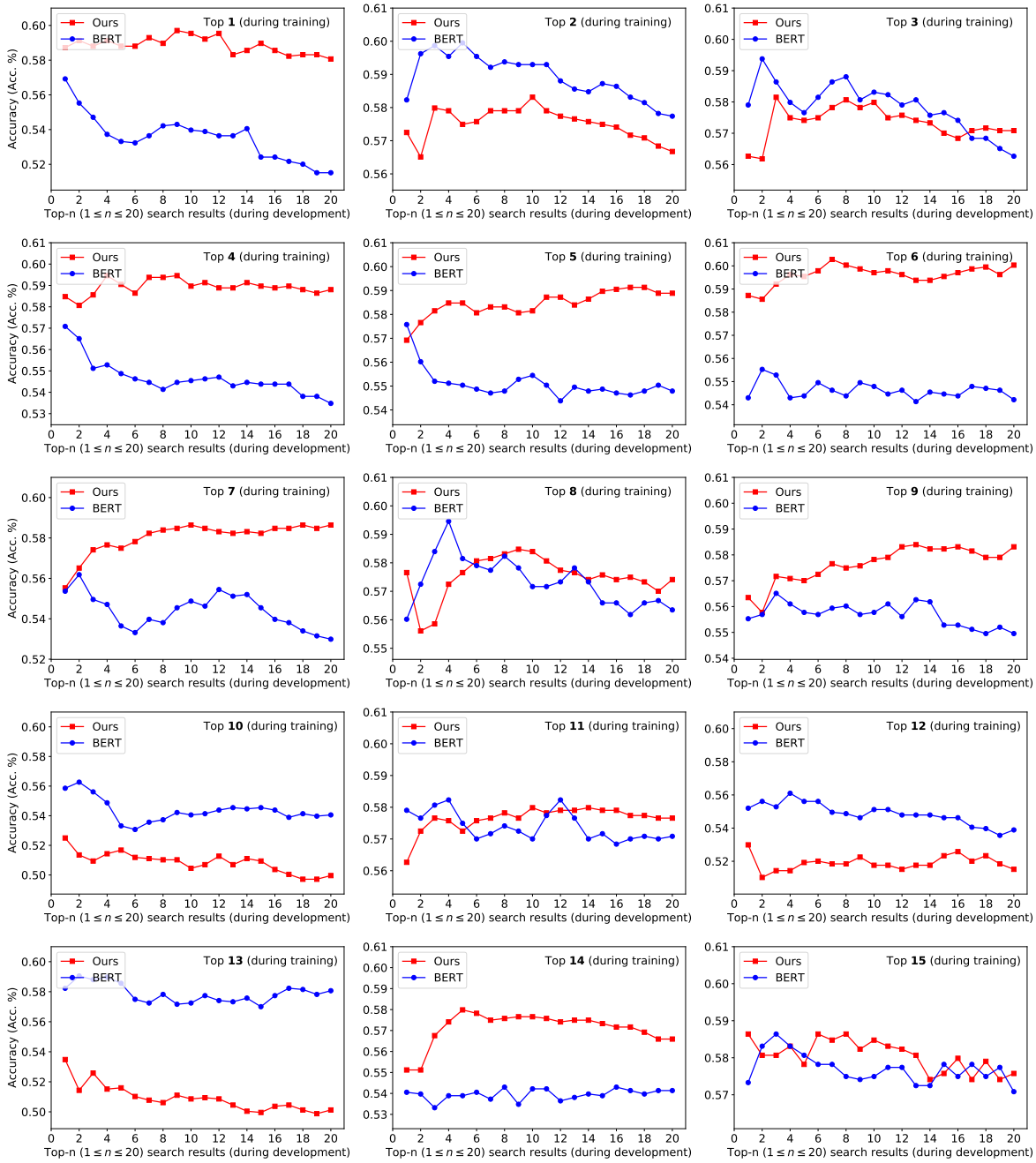

Figure 2: The changing trends of performance when different numbers of highly-ranked search results (knowledge items) are used to support encoding and reasoning. The performance curves are obtained during development.

Question: Though the thin film seemed fragile, for it's intended purpose it was actually nearly what? **Options**: indestructible (prediction), durable, undestroyable, indestructible (ground truth), unbreakable.
Question: What is a person called who doesn't have immortality? **Options**: mortal (prediction), dying, death, dead, mortal (ground truth).

Table 9: Mislabeled MQA examples.

guishable from each other. The model fails to make a distinction between them when the retrieved knowledge items are similar. Table 5 shows two examples. There are 54 cases occurring in the sampled data, constituting 0.54% of the total errors.

Out-of-vocabulary problem causes 24 errors. The problem emerges when a candidate option is not included in the commonsense knowledge base or there is lack of intrinsically relevant knowledge to the question. Table 6 shows two examples regarding the out-of-vocabulary problem. Under this situation, if the search engine toughly operates, the retrieved knowledge is definitely incorrect and

unavoidably drives the reasoning process in a completely wrong direction.

Unreasonable error occurs when the encoder fails to predict the correct answer, even though the top-priority knowledge (i.e., highest-ranked knowledge in the search result list) does serve as the most reasonable evidence for reasoning. There are 12 unreasonable errors found from the sampled data. Table 7 shows a couple of examples. Such errors may be caused by the simple reasoning process, which is merely based on a fully-connected layer.

Less grounded problem happens when the reliable knowledge items fail to be included in the top-n search results, even if they do exist in the commonsense knowledge base. There are 5 less-grounded cases found. Table 8 shows an example. It is difficult to overcome this problem when semantic-level matching has been left out of consideration in a high-speed search engine.

Within the sampled data, we find 5 cases which were obviously **mislabeled**. Table 9 shows two examples. Let us consider the second one, where the first candidate option is the same with the last, and one of them is labeled as the true answer while the other incorrect. Thus, even if the model successfully detects the duplication of the true answer, the case will still be regarded as a negative example during evaluation. Avoiding such kind of "friendly fire", an MQA system may obtain a considerable performance improvement. The preconditions include 1) endorsing the duplications and 2) double-checking all the test data. Actually, if "friendly fire" occurs frequently during training, the existing MQA models have encountered distractors at the very beginning. For a fair comparison, in our experiments, we use the mislabeled MQA examples as the canonical examples.