U3E: Unsupervised and Erasure-based Evidence Extraction for Machine Reading Comprehension

Suzhe He, Shumin Shi*, Chenghao Wu, Heyan Huang

Beijing Institute of Technology, Beijing 100081, China
{suzehe, bjssm, wch, hhy63}@bit.edu.cn

Abstract: More tasks in Machine Reading Comprehension (MRC) require predicting the answer and extracting supported evidence sentences simultaneously. However, the annotation of supporting evidence sentences is usually time-consuming and labor-intensive. In this paper, we address this issue and consider that most of the existing extraction methods are semi-supervised, we propose an unsupervised and erasure-based evidence extraction method named U3E, which takes the changes after sentence-level feature erasure in the document as input, simulating the decline in problem-solving ability caused by human memory decline. In order to make selections based on fully understanding the semantics of the context, we also propose metrics to quickly select the optimal memory model for these input changes. To compare U3E with typical evidence extraction methods and investigate its effectiveness in evidence, we conduct experiments on different datasets. Experimental results show that U3E is simple but effective, also improves model performance.

Keywords: Machine Reading Comprehension; evidence extraction; unsupervised learning; feature erasure

1 Introduction

Machine reading comprehension (MRC) has received extensive attention and can be roughly divided into two categories: extractive and non-extractive. Extractive MRC requires one or more paragraphs of text to be selected as the answer, such as SQuAD [1]. Non-extractive MRC needs reasoning ability more than the former. It includes multiple-choice MRC [2], verification MRC [3], etc. As performance improves, the requirements for interpretability are manifested. In MRC, evidence plays an important role. There is no doubt that evidence sentences can help the MRC model to predict and improve the interpretability of the model. Several works reveal this phenomenon. [4] effectively fuse the extracted evidence in reasoning to enhance the power of relation extraction. [5] iteratively performing long document segmentation enhances text generation performance.

However, it is difficult to label evidence sentences on large-scale datasets, especially for non-extractive MRC. Because a large number of these questions are not just extractive (eg 87% of questions in RACE [6]). Answering such questions may require more advanced reading comprehension skills such as single-sentence or multi-sentence reasoning skills [7].

Considering the high cost of human-labeling evidence sentences, some work [8] generates remote tags using handcrafted rules and external sources, [9] applies remote supervision to generate labels, denoising using a deep probabilistic logical learning framework. Some recent works focus on weakly supervised extraction of evidence sentences, [10] uses a small number of evidence annotations combined with a large number of document-level labels to select evidence, and [11] uses a self-training approach that uses automatically generated evidence labels.

In this paper, we propose an unsupervised method U3E for selecting evidence that is more in line with human intuition. Inspired by [12], they change the characteristics of paragraph input (called erasure) by penalizing illogical predictions to improve the model.

Instead of the loss value used in their work, we choose to use the change in the predicted value to reflect this memory-decay behavior because the predicted value is more sensitive. We innovatively apply it to sentence level and the task of evidence extraction in MRC. Besides, we propose an optimal model selection method BMC to make U3E achieve better results quickly. Experiments on two challenging MRC datasets demonstrate the effectiveness of our proposed method.

2 Related work

2.1 Attribution interpretation methods

In the field of hindsight, there are “variable importance” methods and gradient-based methods. The “variable importance” method [13] refers to the difference in the prediction performance of a model when the value of variable changes. In gradient-based methods, the magnitude of the gradient is used as the feature importance score. Gradient-based methods are suitable for differentiable models [14]. Erasure [12] as a "variable importance" method is model independent. The advantage of the erasure method is that it is conceptually simple and can be optimized for well-defined objectives [15].
2.2 Evidence extraction

Evidence extraction aims to find evidence and relevant information for downstream processes in the task, which arguably improves the interpretability of the task. It can be roughly divided into three categories: one is supervised learning, which requires a lot of resources to manually label all the evidence sentence labels, such as: HOTPOTQA [16] require evidence extraction and work on it [17] iteratively sorts the importance of sentences to select evidence. The second is semi-supervised learning. [9] use remote supervision to generate imperfect labels. [18] use reinforcement learning to obtain better evidence extraction strategies. The last is unsupervised learning. Some methods based on attention and gradient may show counter intuitive behavior [19].

3 Method

The overall architecture of U3E consists of three stages:

1. Train and Acquire (T&A): train models according to the specific task and achieve changes.
2. Select and Reacquire (S&R): select the optimal memory model according to different selection methods and use the model to reacquire changes.
3. Apply and Retrain (A&R): extract evidence through changes and retrain according to the evidence.

Its structure is shown in Figure 1 below.

![Figure 1](image1.png)

**Figure 1** The overall structure of U3E, which includes three stages: T&A, S&R, and A&R.

In the following specific implementation, we will explain these stages in order.

3.1 Task definition

Assuming that each sample of the dataset can be formalized as follows: Given a reference document consisting of multiple sentences \( D = \{S_1, S_2, ..., S_m\} \) and a statement \( O \) (If there is a question, then \( O \) is represented as the concatenation of the question and the candidate). The model should determine whether to support this statement according to the document, the support is marked as 1, otherwise 0. It can also use to extract the evidence sentence set \( E = \{S_j, S_{j+1}, ..., S_{j+k-1}\} \), which contains \( k \) (\(< m\) sentences in \( D \).

3.2 Train and acquire

1) Task-specific Training: We first train according to the specific task (here is the classified task), and then save the model \( M = \{M_1, M_2, ..., M_x\} \) under all epochs, where \( x \) represents the largest epoch trained. The model structure during training is a pretrained model and linear layer. The input is in “[CLS] + Option + [SEP] + Document + [SEP]” format. The hidden representation of the [CLS] token goes through a linear layer for binary classification to predict:

\[
\hat{y} = \text{softmax}(W_p h_{cls})
\]

where \( W_p \) is the model parameter and \( h_{cls} \) is the hidden representation of [CLS] token. The loss function is the cross entropy loss:

\[
L = -\sum_{i=1}^{n} (y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i))
\]

*Where* \( i \in [1,N] \)

2) Change Acquisition: This operation is shown in Figure 2 below.

![Figure 2](image2.png)

**Figure 2** Implementation details for achieving changes.

We use the leave-one-out method [13] to simulate memory decay. As shown in Figure 2, \( M^I \) is the model after the \( I \)-th epoch training is completed, \( R \) represents the binary classification result predicted by model \( M^I \), where \( R_0 \) means the probability value of that document \( D \) does not support statement \( O \), otherwise, \( R_1 \) means support, and \( R_2 \) marked in red means that the correct answer for this example is support.

\( D^{-j} \) represents the new document obtained after erasing the \( j \)-th sentence. Then predict in the same format to get \( R^{-j} \). The importance of the sentence \( C = \{c_1, c_2, ..., c_m\} \) is calculated as follows:

\[
c_j = \text{abs}(R_j - R_j^{'})
\]

*Where* \( j \in [1,m] \)

Where \( y \) is the sample label, \( \text{abs} \) is the absolute value.
function and $c_j$ indicates the importance of sentence $j$.

We choose this method because the predicted value changes can be more sensitive to the input changes.

3.3 Select and reacquire

1) Optimal Model Selection: We have experimented with two optimal model selection methods as follows:
   
a) MTEST: This method considers the generalization of the model, and specifically selects the model according to the maximum accuracy on the test set.
   
b) BMC: Considering the data difference between the two datasets, using MTEST method cannot make the model fully understand the training set, we propose an optimal model $M_0$ selection expression BMC (balance model and changes).

The core of BMC method is to select the model that is most sensitive to the evidence sentence and prevent the model from biasing evidence sentences selection due to excessive pursuit of specific tasks because spurious correlations [20]. The BMC expression is as follows:

$$M_0 = \arg\max_{M} (-\lambda \cdot \text{Acc}(M^l) + \text{SC}^l)$$

$$\text{Where SC}^l = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{k=1}^{k} C_k(M^i)}{\sum C(M^i)}$$

Where $M^i$ is the model after the $i$-th epoch is trained, Acc represents the accuracy on the testing set and $\lambda$ is a hyperparameter. k is the number of evidence sentences to be selected.

$\sum C_k(M^i)$ is the sum of the largest k values in C, $\sum C(M^i)$ is the sum of all the values in C, the division of both represents the sensitivity of the model to the evidence sentence, we call it salient changes (SC).

Note that our model selection can be done simultaneously with training, so when training is over, the optimal model has already been selected.

2) Changes Reacquisition: The steps to reacquire the changes are similar to Change Acquisition, except that the model $M^l$ is replaced with the optimal model $M_0$. The final optimal changes are $C_0$.

3.4 Apply and retrain

We choose the k sentences with the largest $c_j$ in $C_0$ as the evidence sentence set E of statement O. Then under the same model structure, the documents are replaced with ordered evidence sentences (consistent with the relative order of the original sentences) for retraining.

Retrain used to study whether the extracted evidence can improve model performance 4.3, and is not needed in the evidence validity 4.4.

4 Experiments and analysis

4.1 Dataset

VGaokao [3] is a validation dataset, which comes from the Chinese native speaker’s Gaokao Chinese test and is a public dataset. It includes documents and statements, each statement requires at most two evidence sentences, and the answer yes/no directly indicates whether the document supports the statement.

C³ [21] is a multiple-choice reading comprehension dataset. Here we use C³ dataset (https://ymcui.com/expmrc) proposed in [22]. It requires pseudo-evidence annotation for the training set first, then training on the fake-evidence annotated training set, and prediction on the human-annotated test set.

We use VGaokao to study the effect of generalization and the improvement of model performance, and C³ to illustrate the accuracy of extracting evidence.

4.2 Baseline

1) Model Structure: Both use the method of “pre-training model + linear layer”. The loss functions are all cross entropy loss. The former all-process pretrained model uses chinese-roberta-wwm-ext, and the latter uses chinese-macbert-base (https://huggingface.co/). Other differences in the C³ are introduced in 4.4.

2) Implementation: For dataset VGaokao, in order to obtain a more accurate experimental conclusion, during training, we select multiple most relevant sentences for each statement according to the static word vector (https://github.com/Embedding/Chinese-Word-Vectors), ensuring that context does not exceed the length limit of the pretraining model (average sentence length after filtering is 7.793). Then the new document is formed after sorting according to the original position. During testing, we use block prediction (step 128) and max-pooling for classification. Later, the subsequent experiments were conducted on the new document, and the tests were the same.

We use the following baselines: (1) Word Vector: select the most relevant k sentences from document D as evidence sentences for retraining based on word vectors. Here we use average pooling. (2) Beam Search: beam search using the Hard Masking method proposed by [3], they delete the previously selected sentence information in the query, and iteratively select the evidence sentence. (3) Full Context: in order to study the improvement of model performance, we use the original document D for training.

For dataset C³, in order to exclude interference, we delete the data that is too long in the training set provided by [22], and block prediction is also used. The model structure used is consistent with the baseline proposed in [22]. Since the test set is not public, we evaluate it on the validation set. Here, only
one sentence is required to be extracted as evidence.

4.3 Performance improvement
The improvement results are shown in Table II. Bold indicates the maximum accuracy.

| Method                      | Respective Acc | Acc.   |
|-----------------------------|----------------|--------|
| Full context                |                | 64.46  |
| Top 2 sentences by word vector |              | 63.79  |
| Top 2 sentences by beam search |             | 63.63  |
| Top 2 sentences by U3E_{max} |   69.91 / 64.46 | 63.93  |
| Top 2 sentences by U3E_{BMC} | 99.64 / 64.18  | 64.54  |
| Top 2 sentences by U3E_{MAX} | 99.51 / 62.32  | 65.96  |

1) Implementation: In Table II, Acc. represents the test set accuracy. Respective Acc. for all experiments is obtained under the same experimental settings. The Respective Acc. represents the accuracy of different methods of selecting the optimal model on the two data sets. Results of method MTEST is subscribed by 'mtest'. For BMC, the hyperparameter λ is set to 0.1, and the result is subscribed by 'BMC'. Then we also recorded the optimal result in 10 epochs, which is subscribed by 'MAX'.

2) Result Analysis: In general, U3E_{BMC} and U3E_{MAX} are higher than Full context, especially U3E_{MAX} is more than 1 point higher, which means that our U3E method helps to improve the performance of the model. And U3E_{BMC} higher than U3E_{mtest} by 0.61 indicates that the BMC model selection method is more effective than simply using the accuracy of the test set. This also suggests that excessive pursuit of task-specific accuracy may reduce the probability that the model predicts the answer from the evidence sentence. This is consistent with the study of [22]. However, U3E_{BMC} is still more than 1 point lower than U3E_{MAX}, indicating that the BMC method still has room for improvement.

4.4 Evidence validity
To further demonstrate the accuracy of our method for extracting evidence, we conduct experiments on dataset C³.

1) Implementation: Since the candidate answers of the c dataset are not long, we do not need to perform evidence extraction for each candidate, so we do the splicing of “document, question and candidate” for each candidate, and then enter the model together.

Considering that there are two tasks, we design the total loss function as follows:

\[
\text{Loss}_{\text{total}} = \text{Loss}_{\text{ans}} + \alpha \cdot \text{Loss}_{\text{evi}}
\]

Where \( \alpha \in [0,1] \) (5)

Where \( \text{Loss}_{\text{ans}} \) is the loss of multiple answers, and \( \text{Loss}_{\text{evi}} \) is the loss of evidence span. We record the maximum experimental results of the parameter \( \alpha \) every 0.1 from 0 to 1.

Here, we use U3E’s MTEST method for evidence extraction, which is to extract evidence according to the best model on the validation set. Note that here we only extract evidence for training set. The evidence validity results are shown in Table III.

| Type of PrM | Method      | ANS_F1 | EVI_F1 | ALL_F1 |
|-------------|-------------|--------|--------|--------|
| RAW         | 72.245      |        |        |        |
| Base        | WV          | 71.881 | 65.533 | 49.634 |
|             | U3E_{mtest} | 72.475 | 68.840 | 53.385 |
| RAW         | 77.425      |        |        |        |
| Large       | WV          | 76.832 | 68.049 | 55.355 |
|             | U3E_{mtest} | 78.812 | 76.425 | 62.720 |

In Table III, RAW represents the answer prediction accuracy without evidence extraction task. WV means using word vectors to extract evidence, both of which select evidence sentences according to the correct answers and questions. ANS_F1 is the accuracy of multiple-choice answers, EVI_F1 is the accuracy of evidence, and ALL_F1 is the accuracy of the combination of both. The evaluation method is provided by [22].

2) Result Analysis: As shown in Table III, the evidence obtained using our method greatly improves the upper limit of the evidence accuracy rate, from 65.533 to 68.840, and simultaneous training can help improve the accuracy of multiple-choice answers, from 72.245 to 72.475. Moreover, according to the 4.3 research, method MTEST does not use the optimal model, and a greater improvement is expected.

Moreover, when the pre-training model is larger (https://huggingface.co/hfl/chinese-macbert-large), observe more significant changes. U3E not only improves the evidence prediction by more than 8 points, but also improves the multi choice answer prediction by more than 1 point compared with the original training data.

We also compare the specific effects of using word vectors and U3E. The results are shown in Table IV. WV is the static word vector. The following examples are translated from Chinese.

| Evidence extraction results |
|-----------------------------|
| **Query:** What might their relationship be? |
| **WV:** Woman: Yes, it’s been more than ten years since we saw each other. |
| **U3E_{mtest}:** Man: But you haven’t changed at all, just like when you were in college. |
| **Query:** What kind of teacher did Zhang Liyong become? |
| **WV:** Zhang Liyong did not become a teaching assistant in the English Department of Tsinghua University, … |
| **U3E_{mtest}:** He is an English tutor for a third-year student. |

The former tends to select more similar sentences, which misleads the model. The latter makes full use of the favorable conditions of the model “understanding
the semantics of context”. It’s more intuitive and more accurate.

5 Conclusion

We propose an erasure-based method called U3E for MRC. U3E builds a model based on the reduced ability to process problems due to memory loss in humans. In order to select the optimal memory model quickly, a calculation method called BMC is proposed. The experimental results show that U3E is effective. In the future work, we will try to apply U3E to other NLP tasks such as natural language inference.

Acknowledgements

This work is supported by the National Key Research & Development Program (2018YFC0831700) and National Natural Science Foundation of China (61671064, 61732005).

References

[1] Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).
[2] Zhu, Pengfei, et al. "DUMA: Reading comprehension with transposition thinking." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 269-279.
[3] Zhang, Chen, et al. "Extract, Integrate, Compete: Towards Verification Style Reading Comprehension." arXiv preprint arXiv:2109.05149 (2021).
[4] Xie, Yiqing, et al. "Eider: Empowering Document-level Relation Extraction with Efficient Evidence Extraction and Inference-stage Fusion." Findings of the Association for Computational Linguistics: ACL 2022. 2022.
[5] Zhang, Yusen, et al. "Summ’N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents." arXiv preprint arXiv:2110.10150 (2021).
[6] Lai, Guokun, et al. "Race: Large-scale reading comprehension dataset from examinations." arXiv preprint arXiv:1704.04683 (2017).
[7] Tu, Ming, et al. "Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents." Proceedings of the AAAI conference on artificial intelligence. Vol. 34. No. 05. 2020.
[8] Lin, Yankai, et al. "Denoising distantly supervised open-domain question answering." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.
[9] Wang, Hai, et al. "Evidence sentence extraction for machine reading comprehension." arXiv preprint arXiv:1902.08852 (2019).
[10] Pruthi, Danish, et al. "Weakly-and semi-supervised evidence extraction." arXiv preprint arXiv:2011.01459 (2020).
[11] Niu, Yilin, et al. "A self-training method for machine reading comprehension with soft evidence extraction." arXiv preprint arXiv:2005.05189 (2020).
[12] Ju, Yiming, et al. "Enhancing Multiple-choice Machine Reading Comprehension by Punishing Illogical Interpretations." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

2021.
[13] Li, Jiwei, et al. "Understanding neural networks through representation erasure." arXiv preprint arXiv:1612.08220 (2016).
[14] Sundararajan, Mukund, et al. "Axiomatic attribution for deep networks." International conference on machine learning. PMLR, 2017.
[15] De Cao, Nicola, et al. "How do decisions emerge across layers in neural models? interpretation with differentiable masking." arXiv preprint arXiv:2004.14992 (2020).
[16] Yang, Zhilin, et al. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." arXiv preprint arXiv:1809.09600 (2018).
[17] Nishida, Kosuke, et al. "Answering while summarizing: Multi-task learning for multi-hop QA with evidence extraction." arXiv preprint arXiv:1905.08511 (2019).
[18] Wang, Shuohang, et al. "R3 3: Reinforced ranker-reader for open-domain question answering." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[19] Feng, Shi, et al. "Pathologies of neural models make interpretations difficult." arXiv preprint arXiv:1804.07781 (2018).
[20] Jia, Robin, et al. "Adversarial examples for evaluating reading comprehension systems." arXiv preprint arXiv:1707.07328 (2017).
[21] Sun, Kai, et al. "Investigating prior knowledge for challenging chinese machine reading comprehension." Transactions of the Association for Computational Linguistics 8 (2020): 141-155.
[22] Cui, Yiming, et al. "ExpMR: Explainability evaluation for machine reading comprehension." Heliyon 8.4 (2022): e09290.