Supporting Evidence Retrieval with Cross-Sentence Coupling

1st Che-Wei, Lin*
Halıcıoğlu Data Science Institute
University of California, San Diego
California, USA
chl820@ucsd.edu

2nd Meng-Tse Wu
Institute of Information Science
Academia Sinica
Taipei City, Taiwan
moju@iis.sinica.edu.tw

3rd Keh-Yih Su
Institute of Information Science
Academia Sinica
Taipei City, Taiwan
kysu@iis.sinica.edu.tw

***Abstract*—This paper proposes a novel approach to identify the supporting evidence (SE) sentence (within a related document) that has low lexicon recall rate with the given question passage. Previous approaches such as BERT mainly identify SE sentences via implicitly measuring the text similarity between the question and each sentence in the related document. However, some SE sentences possess low recall rates because they have low word-overlapping with the question text. This situation frequently occurs when the complete information crosses more than one sentence. We thus propose a novel model that incorporates cross-sentence coupling between adjacent sentences. The experiments conducted on a Chinese QA data-set show the proposed model has made 2.1% EM improvement comparing with BERT baseline.**

***Index Terms*—Explainable AI, Image Classification, Grad-CAM, Integrated Gradient**

I. Introduction

Supporting evidence (SE) is a passage which contains barely sufficient information to answer a particular question or to infer a statement to be correct. Figure 1 shows an example of the question and answering (QA) task. In this example, the question is paired with the corresponding document as its background knowledge. Retrieving the associated SE according to the given question is an important task within a QA framework [1] even if a related document is given, as filtering out irrelevant content in advance not only can save computation resource but also can improve QA performance (because the searching space is reduced). In addition, SE retrieval can be applied in the task of claim verification [2]. In order for a model to verify a claim correctly, it has to rely on appropriate reasoning statements to prove the claim to be right. Therefore, the process of retrieving correct reasoning statement is also essential to ensuring the model’s high performance.

Question: 《蘇文忠公全集》是由何人編撰？

Document:
S_0: 有《東坡先生大全集》及《東坡樂府》詞集傳世
S_1: 宋人王宗稷收其作品
S_2: 編有《蘇文忠公全集》
S_3: 其散文、詩、詞、賦均有成就

SE: S_1, S_2

Fig. 1. An example that demonstrates the necessity of considering coupling to the anchor sentence while retrieving supporting evidence.

Previous approaches can be divided into three main categories: (1) Keyword matching technique [1], [3] (2) Translation-based technique [4], and (3) Neural-semantic technique [5]. Keyword matching approaches are mainly based on the intuition that the question and the supporting evidence should share considerable words. However, keyword matching technique is unable to recognize a synonym with a different string. For example, if a question asks for the name of the current president of Taiwan, keyword matching technique might ignore some synonyms of the word 「總統」, such as 「元首」、「領袖」、「領導人」in a potential SE. Translation-based technique adopts statistical translation model to measure the similarity. However, it requires human-specified features, but feature engineering is time-consuming. Current neural-semantic approaches solve this issue by using word embeddings; nonetheless, they usually ignore the cross-sentence coupling to the reference anchor sentences (which possess high word-overlapping rates with the question text). Details will be explained as follows.

While recent neural-semantic technique (such as BERT [6]) successfully leverages contextualized word embeddings and attention mechanism for measuring the similarity between the question and each document-sentence, it mainly utilizes surface features and is only based on each individual sentence. However, some sentences that should be extracted have low overlapping rates with the question text, which is mainly due to that they might not be complete sentences. This kind of SE sentence thus needs its neighboring sentences to act as reference anchors, which are SE sentences that could be reliably identified with the word-overlapping rate. For example, Figure 1 shows that to answer the given question, both S_1 and S_2 are required. Although S_2 can be easily identified due to its

[*] This work was done during his 2020 summer internship in IIS, Academia Sinica.

high similarity with the question, S_1, however, is missed by BERT due to its low similarity value with the question. After observing these two supporting sentences, we discover that there is a tight coupling between these two sentences: S_2 contains the predicate (編有), S_1 contains its subject (王宗稷). Humans identify S_1, which contains the actual answer, mainly based on the coupling between S_1 and S_2 (that is, S_1 provides the subject/agent to S_2, which could be reliably identified). Due to ignoring the coupling to the anchor sentences, current neural-semantic techniques thus result in limited performance.

To overcome the weakness above mentioned, we propose to enhance the neural-semantic approach via further incorporating the coupling to the anchor sentence. That is, while judging if the targeted sentence is a SE, the proposed approach also takes its adjacent sentences (i.e., the potential reference anchor sentences) into account. In this paper, we evaluate various approaches based on the FgcQA dataset, which is a question-answering dataset created and used by the 2020 Formosa Grand Challenge[1] . We manually annotate the SE sentences of each question as our benchmark. The FgcQA dataset is well suited for our task, as it has long documents in which the sentences are segmented by punctuation, providing great challenges to our approach.

In summary, our contributions in this paper are threefold:

- We propose a novel approach to utilize the cross-sentence coupling with those reference anchor sentences for SE retrieval.
- We construct a benchmark dataset for evaluating the performance of SE retrieval, which could help other people conduct related research.
- We conduct experiments to demonstrate the effectiveness of our proposed approach.

II. Proposed Model

A. Task Formulation

For every question q with a paired source document D, the SE retrieval model aims at extracting some SE sentences from D. For each document-sentence S_i in document D, the model outputs its corresponding decision $d_i \in \{support, not\,support\}$ indicating whether S_i is a SE sentence. This problem could be formulated as:

$$\hat{d_1^n} = \underset{d_1^n}{\operatorname{argmax}}\; P(d_1^n \mid q, D) \tag{1}$$

where n is the number of sentences in D, and d_1^n denote the corresponding decision sequence, $\{d_1, ..., d_n\}$.

B. BERT + Neighbor Sentence Attention

Our model is composed of two stages: question-sentence similarity generation and sentence similarity aggregation.

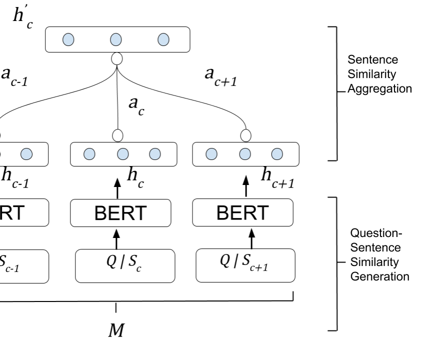

Fig. 2. The proposed model structure, in which Q/S denotes question-sentence input with the format specified in Equation (2), and h_{c-1}, h_c, h_{c+1} are sentence similarity vectors from three adjacent sentences

Each stage is explained in greater details below.

1) Question-sentence similarity generation: First, we adopt BERT to evaluate the semantic similarity between the question and a specific sentence S_i. Specifically, we first concatenate the given question Q and S_i, and then pass them to a BERT encoder. Afterwards, we use the output vector $h_i \in \mathbb{R}^{d_i}$ of BERT's [CLS] token as the similarity vector between Q and S_i (as shown in Eq. (2)). For each targeted sentence S_c, which is the sentence we are going to predict if it's a SE, we will generate its semantic similarity h_c with the question, and also the semantic similarities (h_{c-1}, h_{c+1}) from its two adjacent sentences (one on each side).

$$h_i = BERT_{[CLS]}\left([CLS]Q[SEP]S_i[CLS]\right) \tag{2}$$

2) Sentence similarity aggregation: Afterwards, we aggregate each $h_i \in \{h_{c-1}, h_c, h_{c+1}\}$ within the window M (width = 3). Inspired by [7], we first transform each individual similarity vector h_i into $k_i \in \mathbb{R}^{d_k}$, as shown in equation (3). We also separately transform the similarity of the targeted sentence, denoted as h_c, into $q_c \in \mathbb{R}^{d_k}$, as shown in equation (4).

$$k_i = W^k h_i, h_i \in h_{c-1}, h_c, h_{c+1} \tag{3}$$

$$q_c = W^Q h_c \tag{4}$$

where $W^Q \in \mathbb{R}^{d_q \times d_h}$and $W^K \in \mathbb{R}^{d_k \times d_h}$ are learned during training (we set $d_q = d_k = d_h = 768$ in all our experiments). To aggregate similarity measures of all sentences within the window, we need to obtain the attention weight a_i for each h_i. The attention weight a_i is computed as follows:

$$e_i = \frac{q_c\, k_i^T}{\sqrt{d_k}} \tag{5}$$

$$a_i = Softmax\left(e_i\right) = \frac{exp(e_i)}{\sum_{j \in M} exp(e_j)} \tag{6}$$

[1] The Grand Challenge Competition is designed to assess current AI capabilities via a QA task and is organized by MOST, Taiwan: https://fgc.stpi.narl.org.tw.

where e_i is the weight for h_i and a_i is the weight after a softmax function is applied. Afterwards, the aggregated sentence similarity $h_c^{'}$ is obtained by the weighted sum of all similarity vectors:

$$h_c^{'} = \sum\nolimits_i^M a_i h_i \tag{7}$$

Finally, the probability that the targeted sentence is a part of SE is calculated by:

$$P = Sigmoid\,(MLP(h_c^{'})) \tag{8}$$

where MLP is a multi-layer feed forward network.

However, the above-mentioned attention mechanism ignores the order of the sentences within the window. Therefore, we further encode the relative distance between the targeted sentence and its adjacent sentences as the position embedding p_i, which is also learned during training. Equation (3) thus becomes:

$$k_i = W^k h_i + p_i \tag{9}$$

III. Evaluation

A. Data Set Adopted

Our proposed model is evaluated on FgcQA dataset, which consists of documents extracted from Wikipedia, news and government websites. It has questions types such as single-span extraction, multi-span extraction, etc. Within the dataset, each question is paired with a document, along with the corresponding indexes of the sentences that serve as the supporting evidence(s) within the document. In our experiments, this dataset is split into training, development, and test three non-overlapping subsets. Table I shows the statistics of these three sub-sets.

	Train	Development	Test
Questions	871	239	182
Documents	104	29	25
Sentences per Document	35.1	34.6	31.3
Sentences per SE	2.14	2.25	2.28

TABLE I: Statistics of FgcQA Benchmark Dataset

B. Baseline Model Adopted

We adopt BERT as our baseline model, because it not only provides SOTA performance, but also is widely adopted for comparing various approaches. Within BERT, we concatenate the question and the targeted sentence (to be judged if it is a SE) with a special "[SEP]" token between them and use the output of another special "[CLS]" token to make the judgment. Specifically, we input the pair of the question and the targeted sentence to the BERT encoder as shown in Equation (2). We then pass BERT's "[CLS]" output embedding h_i to a feed-forward layer, which will export a single value v. Then $P(\mathrm{s}_i|\mathrm{q}) = Sigmoid(v)$ would be the probability that s_i supports the question q. If this probability exceeds a pre-specified threshold value (which is 0.5 in all our experiments), then the given sentence is regarded as a part of SE.

C. Experiment Settings

We adopt the same experiment setting for various approaches. We select the following best configuration based on the development-set: the learning rate is $2e^{-5}$, and the batch size is 64. Furthermore, due to memory constraint, we set the maximum length of the input token sequence to be 300. We conduct early stopping by picking the model that has the best development set performance among 20 epochs.

IV. Performance

	EM	Precision	Recall	F1
BERT-baseline	16.8	59.3	59.1	54.2
Our model	16.3	60.1	**63.2**	**56.3**
Our model+Sentence Position	**18.9**	**60.3**	56.3	53.5

TABLE II: Performances of various models on the test set

Table II shows the performance of each model on the test set. Our model obtains the best F1 score without sentence position information; on contrast, it outperforms the baseline model 2.1% in EM but has 0.7% drop in F1 score if sentence position information is involved.

Question: 先生的奶奶要怎麼稱呼她？

Document:
S_1: 我們要稱呼他為祖翁（爺爺）；
$S_2(targeted)$: 當對象是丈夫的祖母，
S_3: 我們要稱呼他為祖姑（奶奶）。

Gold SE: S_2, S_3

Fig. 3. An example of a successful case (S_2 is the targeted sentence)

		Our Model+
	Baseline	Sentence Position
Attention Weight (a_i)	X	S_2: [0.01, 0.21, 0.78]
Possibility (P)	S_2: 0.00, S_3: 0.99	S_2: 0.84, S_3: 0.99
Prediction	S_3	S_2, S_3

TABLE III: Attention Weights and Scores of a successful case (S_2 is the targeted sentence)

Figure 3 and Table III show a successful case in which the attention weight of S_3 is dominant and hence enhances the possibility of the targeted sentence S_2 being a supporting evidence. From this example, we can see that S_2 has the phrase "當…" which implies that it is not a complete sentence. Therefore, the attention on the next sentence S_3, which is our reference anchor SE sentence, should be high.

V. Error Analysis and Related Work

While our model does perform as what we expect on some test cases, there are some others that our model fails to make the correct prediction.

Question: 目前有紀錄的鯨鯊最久可以活到幾歲？

Document:
S_7: 其他體型更大的個體報告並未確認。
$S_8(targeted)$: 鯨鯊生活在熱帶和溫帶海域中，
S_9: 壽命可達 *70* 年至 *100* 年。

Gold SE: S_8, S_9

Fig. 4. An example of a failed case (S_8 is the targeted sentence)

	Baseline	Our Model+ Sentence Position
Attention Weight (a_i)	X	S_8: [0.06, 0.93, 0.01]
Possibility (P)	S_8: 0.00, S_9: 0.96	S_8: 0.00, S_9: 0.95
Prediction	S_9	S_9

TABLE IV: Attention Weights and Scores of a failed case (S_8 is the targeted sentence)

Figure 4 and Table IV show a failed case where the attention weight of S_9 is close to zero and hence fails to enhance the score of S_8. The reasons for this are conjectured as follows: first of all, the part-of-speech and punctuation information of S_8 and S_9, which are essential to indicate that S_8 and S_9 are closely coupled, are not adopted in the current model. Secondly, the dataset might not be large enough for the model to correctly learn the coupling among adjacent sentences. Out of 239 questions in the development set, roughly only 6.2% of them have SE sentences that are coupled to their adjacent sentences. The proposed model is expected to perform better if more training data of this kind could be provided.

VI. Related Work

The previous approaches for extracting SE could be grouped into three categories (1) Keyword matching technique [1], [3], (2) Translation-based technique [4], and (3) Neural-semantic technique [5]. Keyword matching technique relies on exact-matched terms for its similarity scoring metric. Among them, Chen et al. [1] used bigram hashing and TF-IDF matching to retrieve the related documents (from all the Wikipedia documents) for the given question. Wu et al. [3] adopted n-gram BLEU score and F-measure to measure the similarity and used Particle Swarm Optimization [8] algorithm to find the optimal weights for the n-gram metrics. On the other hand, translation-based technique utilizes translation model to get the probability that a specific sentence is translated from the given question/query. For example, Cui et al. [4] utilizes relation paths extracted from the dependency trees of both sentence and question and train a translation model that estimates the translation probability of the relations. Recently, neural-semantic approaches have dominated this field, which adopt word embeddings as the input features and devise some neural models for measuring the similarity between the question and each document sentences. Furthermore, contextualized word embeddings could be generated from either ELMO [9] or BERT via taking the context words into account. Besides, Karpukhin et al. [5] encoded the passages and the question with a passage encoder and a question encoder, respectively. Those encoders use the output of BERT’ s [CLS] token as the encoded passage/question vectors (as our Equation 2). And the similarity is the dot product of the passage and question vectors. Compared to previous approaches, our proposed model aggregates the neighbor sentences with an attention mechanism in order to take the adjacent reference anchor sentence into consideration.

VII. Conclusion

We propose an approach for referring to the adjacent reliable anchor sentence while judging if the targeted sentence is a part of SE. Different from the BERT baseline model which only independently evaluates the similarity between each SE sentence candidate with the question, our model can recover those low overlapping (with the question) SE sentences via considering the coupling between the targeted sentence and its adjacent high overlapping anchor sentences. Our experimental results show that our proposed model could effectively take the adjacent reference anchor sentence into account while making SE judgment.

References

[1] D. Chen, A. Fisch, J. Weston, and A. Bordes, “Reading wikipedia to answer open-domain questions,” *arXiv preprint arXiv:1704.00051*, 2017.

[2] A. Soleimani, C. Monz, and M. Worring, “Bert for evidence retrieval and claim verification,” in *European Conference on Information Retrieval*. Springer, Cham, 2020, pp. 359–366.

[3] M.-T. Wu, Y.-C. Lin, and K.-Y. Su, “Supporting evidence retrieval for answering yes/no questions,” 中文計算語言學期刊, vol. 23, no. 2, pp. 47–65, 2018.

[4] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, “Question answering passage retrieval using dependency relations,” in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 400–407.

[5] V. Karpukhin, B. Oğuz, S. Min, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, “Dense passage retrieval for open-domain question answering,” *arXiv preprint arXiv:2004.04906*, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95-International Conference on Neural Networks*, vol. 4. IEEE, 1995, pp. 1942–1948.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” *arXiv preprint arXiv:1802.05365*, 2018.