Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Temporality-enhanced knowledge memory network for factoid question answering*

Xin-yu DUAN[1], Si-liang TANG‡[1], Sheng-yu ZHANG[2], Yin ZHANG[1], Zhou ZHAO[1], Jian-ru XUE[3], Yue-ting ZHUANG[1], Fei WU‡[1]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*School of Information Management, Wuhan University, Wuhan 430000, China*

[3]*Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China*

E-mail: {duanxinyu, siliang}@zju.edu.cn; light.e.gal@gmail.com; {zhangyin98, zhaozhou}@zju.edu.cn; jrxue@mail.xjtu.edu.cn; {yzhuang, wufei}@zju.edu.cn

Received Nov. 25, 2017; Revision accepted Jan. 24, 2018; Crosschecked Jan. 25, 2018

Abstract: Question answering is an important problem that aims to deliver specific answers to questions posed by humans in natural language. How to efficiently identify the exact answer with respect to a given question has become an active line of research. Previous approaches in factoid question answering tasks typically focus on modeling the semantic relevance or syntactic relationship between a given question and its corresponding answer. Most of these models suffer when a question contains very little content that is indicative of the answer. In this paper, we devise an architecture named the temporality-enhanced knowledge memory network (TE-KMN) and apply the model to a factoid question answering dataset from a trivia competition called quiz bowl. Unlike most of the existing approaches, our model encodes not only the content of questions and answers, but also the temporal cues in a sequence of ordered sentences which gradually remark the answer. Moreover, our model collaboratively uses external knowledge for a better understanding of a given question. The experimental results demonstrate that our method achieves better performance than several state-of-the-art methods.

Key words: Question answering; Knowledge memory; Temporality interaction
https://doi.org/10.1631/FITEE.1700788 **CLC number:** TP391

1 Introduction

Question answering (QA) is an application that enables users to post their questions and to solve problems. The benefits of QA systems have been well recognized in Jurczyk and Agichtein (2007) and Li et al. (2012). Some QA sites are becoming more and more popular in the real world and have accumulated a vast number of questions with their corresponding answers over time. With a huge amount of QA data, QA has become an active line of research and attracted a lot of attention in the fields of information retrieval and natural language processing (NLP) (Bilotti et al., 2010). Among the varieties of QA research, factoid question answering (FQA) is one of the most widely studied tasks (Wang, 2006).

Given a natural language question, FQA aims to extract entity answers. The single sentence question (Bao et al., 2014; Yao and Durme, 2014) is the most common form of FQA, e.g., "which continent is the largest in the world?" There is another form of FQA which is given as a paragraph describing a certain

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2015CB352302), the National Natural Science Foundation of China (Nos. 61625107, U1611461, U1509206, and 61402403), the Key Program of Zhejiang Province, China (No. 2015C01027), the Chinese Knowledge Center for Engineering Sciences and Technology, and the Fundamental Research Funds for the Central Universities, China

ORCID: Xin-yu DUAN, http://orcid.org/0000-0002-6803-7964

entity. A typical example of such FQA is called quiz bowl question answering (Boyd-Graber et al., 2012; Iyyer et al., 2014). It is a popular trivia game of students in high school and college throughout the world. In quiz bowl question answering, users read several sentences one by one and attempt to answer the question after reading fewer sentences. A quiz bowl question is composed of a sequence of ordered sentences describing an answer from different perspectives, which can be extracted from an entity set. It has a property called pyramidality, which means that the prior sentences in a question description remark harder and more obscure cues, whereas later sentences are 'giveaways'. Table 1 shows an example of a quiz bowl question with its corresponding answer. Words with the same color share the same indicative cues to the answer (e.g., the Notre Dame du Haut and the Villa Savoye are two buildings designed by Le Corbusier using reinforced concrete framing. The long horizontal sliding windows are one highlight of the Villa Savoye). The question consists of six sentences in order.

The existing approaches tend to treat the quiz bowl problem as a text classification task, focus merely on exploiting the semantic relevance between the individual sentence and its corresponding answer, and attempt to discover a better embedding space to perform the semantic classification on quiz bowl data. However, some studies in the fields of sociology and biology (Carr, 1993; Ivry, 1996) discovered that the formation of a decision is in general influenced by the temporal cue transmittance in a flow of information. In a real quiz bowl competition, given a sequence of ordered sentences (e.g., the example in Table 1, where the sentences form a single question in quiz bowl), the indicative cues are gradually delivered sentence by sentence. Thus, it is better to collaboratively use the temporal cues in sentence order to identify the corresponding answer. Moreover, human beings will probably retrospect relevant auxiliary information when reading sentences. In most cases, it is rational that we use external knowledge that relates to each descriptive sentence to give a correct answer (Minsky, 1991; Schweppe and Rummer, 2013). As a result, we speculate that temporal cue transmittance and auxiliary external knowledge are two factors that contribute to answering quiz bowl questions, and we assume that it is probably meaningful to capture the temporal cues in a sequence of ordered sentences and leverage auxiliary knowledge that relates to each reading sentence.

Table 1 A quiz bowl example question that describes the architect 'Le Corbusier'

Number	Question description
1	This architect is one of the first designers to use reinforced concrete.
2	This man argued that architecture has historically been dominated by the influence of 'regulating lines'.
3	Pilotis, roof gardens, and long horizontal sliding windows are included in this architect's five points of architecture introduced in his book, towards a new architecture.
4	He connected one of his buildings with the wings of a dove and included irregularly spaced rectangular openings within its two-feet thick walls.
5	This city planner of Chandigarh, India used columns called pilotis to raise many of his buildings above the ground.
6	For 10 points, name this Swiss-French designer of the Notre Dame du Haut and the Villa Savove.
Answer	Le Corbusier

The descriptive sentences contain many obscure or obvious cues to the answer from different aspects. The sentences ordered ahead remark fewer cues to the answer, but the sentences in later order deliver more helpful indications to capture the answer. Specifically, we endow words that share the same indicative cues to the answer with the same color and intuitively demonstrate the temporal cues in the sequence of ordered sentences

Inspired by our assumption, we are more interested in how to effectively leverage the temporal cues in a sequence of ordered sentences (the answer is in general described from different aspects) and auxiliary external knowledge (each sentence can spark relevant information that relates to itself with respect to a given question), to improve the answering performance for the quiz bowl question. In particular, we devise a new architecture named the temporality-enhanced knowledge memory network (TE-KMN). This end-to-end architecture introduces an attention-based sequential model, which is an extension of the gated recurrent unit (GRU) (Cho et al., 2014), to capture the temporal cues in the sequence of ordered sentences (i.e., how the cues in ordered ahead sentences influence the understanding of the later sentences). Meanwhile, external memory is employed to leverage auxiliary knowledge that relates to each reading sentence and augments the understanding of each reading sentence.

Several contributions of our work are highlighted as follows:

1. Different from traditional content-based methods, a novel architecture TE-KMN is proposed to leverage auxiliary knowledge that relates to each reading sentence and the temporal cues in a sequence of descriptive sentences with respect to a given question.

2. Based on traditional GRU, we introduce a sequential model with attention mechanism in our proposed architecture, namely temporality-enhanced gated recurrent unit (T-GRU), to capture the temporal cues in a sequence of ordered sentences. This attention mechanism transfers more effectively the cues from ordered ahead sentences than those from the later sentences.

3. The auxiliary knowledge that relates to each reading sentence in an external memory is adopted to understand each reading sentence in our knowledge memory network.

2 Related work

Question answering is a popular research line in NLP. Previous studies on FQA can be categorized into two tasks: answer classification and answer ranking. Answer classification aims at classifying the quality of answers while answer ranking focuses on figuring out the best answer among numerous candidate answers concerning a certain question. These studies commonly rely on exploring semantic or syntactic features of QA data. Sun et al. (2005) tried to detect the semantic and syntactic relationship between questions and answers. Navigli and Velardi (2010) presented a lattice-based approach to definition and hypernym extraction. Huang et al. (2007) integrated textual features to represent the candidate QA pairs and used a support vector machine (SVM) to classify QA pairs. Despite the typical features, Shah and Pomerantz (2010) trained a logistic regression (LR) classifier and predicted the quality of answers in CQA. Ding et al. (2008) proposed a model on the conditional random field (CRF) to capture contextual features from the answer sequence for semantic matching between QA pairs. Figueroa and Atkinson (2011) proposed maximum entropy context models for ranking biographical answers to definition queries on the Internet. Some methods use various features at the thread level that allow more consistent global decisions and exploit the relationship between pairs of comments at any distance in the answer list with respect to a question (Barrón-Cedeño et al., 2015; Joty et al., 2015). Some translation models were used to match QA by transferring the answers to the corresponding question (Jeon et al., 2005; Xue et al., 2008; Zhou et al., 2011). However, the performance of translation-based approaches deteriorates when there are many informal words or phrases in QA archives.

Recently, some studies reported the applications of deep neural networks to QA tasks. Wang et al. (2010) proposed a deep belief network (DBN) based semantic relevance model to learn the distributed representation of QA pairs. Shen et al. (2015) calculated a similarity matrix for each QA pair containing the lexical and sequential information and then used a deep convolutional neural network (CNN) to estimate the suitable answer probability. Different from the classical CNN used in Shen et al. (2015), Qiu and Huang (2015) introduced a dynamic CNN (Kalchbrenner et al., 2014) to encode the variable-length sentences of questions and answers in the semantic space and model their interactions with a tensorial top layer. Besides CNN, another kind of neural network has been successfully applied in textual content analysis. In Le and Mikolov (2014), a recurrent neural network (RNN) was employed to represent each sentence or one document by a dense vector. In Sutskever et al. (2014), a multi-layer RNN was used to map the input sentence to a vector of fixed dimensionality.

The success of RNN on NLP (Mikolov et al., 2010) shows that it is capable of handling long-term dependencies by adaptively memorizing values for either long or short durations. Long short-term memory (LSTM) neural network (Mikolov et al., 2010), GRU (Cho et al., 2014), and their bidirectional variants BiLSTM (Graves et al., 2005) also show promising results in many NLP fields such as machine translation (Bahdanau et al., 2014), named entity recognition (Ma and Hovy, 2016), and reading comprehension (Chen et al., 2016). Meanwhile, many studies (Chorowski et al., 2015; Luong et al., 2015; Rush et al., 2015) applied attention mechanism to RNNs to improve the performance of NLP tasks. The idea is to allow the model to attend over past output vectors, thereby mitigating the RNN's cell state bottleneck.

Quiz bowl question answering is a typical example of FQA. In recent years, many approaches

have been proposed to solve the quiz bowl questions using machine learning methods. Boyd-Graber et al. (2012) introduced a naive-Bayes model to identify the answer based on manually defined sequence matching rules and bag of words (BOW) representations. To enrich the semantics in sentence representation, Iyyer et al. (2014) used a dependency-tree recursive neural network (DT-RNN) to exploit dependency information in sentence encoding. Zheng et al. (2015) introduced a multi-channel CNN to represent quiz bowl questions at the sentence level and paragraph level. Iyyer et al. (2015) proposed a deep averaging network (DAN) to map a descriptive paragraph to its answer.

Reasoning via cross-media (e.g., textual and visual modalities) is very important for human beings. There are many applications of cross-media (Yang et al., 2008; Zhuang et al., 2008). In the NLP field, many researchers have tried to endow their models with the ability of reasoning by leveraging auxiliary knowledge since the importance of auxiliary knowledge has long been recognized (Fillmore, 1976; Minsky, 1991). Earlier NLP systems mostly exploit restricted linguistic knowledge such as manually-encoded morphological and syntactic patterns. With the advanced development of the knowledge base, large amounts of semantic knowledge has become available. Many studies have exploited the use of these knowledge bases (e.g., DBPedia and FreeBase) to improve the performance of their models in many NLP tasks (Nakashole and Mitchell, 2015; Wei et al., 2017; Yang and Mitchell, 2017). Recently, the memory network (Sukhbaatar et al., 2015) has achieved influential results in machine reading comprehension (Pan et al., 2017), which proves that the memory network is able to better understand the input documents. Meanwhile, Ghazvininejad et al. (2017) used a memory network to memorize and understand the external knowledge in a conversation generation task and has shown certain promise.

3 Methodology

In this section, we introduce the proposed temporality-enhanced knowledge memory network (TE-KMN) in detail (Fig. 1). The proposed model consists mainly of the following steps: individual encoding of each sentence, temporal cue transmittance in a sequence of ordered sentences using T-GRU, and utilization of auxiliary knowledge that relates to each reading sentence.

3.1 Sentence encoder

The first step in our architecture is to represent the textual contents of descriptive sentences and knowledge with proper semantic embeddings. In our model, an RNN-based method (Mikolov et al., 2013) is employed to model the distributed representation of each word in the dataset $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_m\}$ and the knowledge memory $K = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_n\}$, where $\boldsymbol{d}_i$ $(i = 1, 2, \ldots, m)$ is a question (a descriptive paragraph consisting of a sequence of ordered sentences), $\boldsymbol{k}_j$ $(j = 1, 2, \ldots, n)$ is one fact (i.e., in the form of one sentence) in the knowledge memory, and m and n the numbers of questions in the dataset and the facts in the memory, respectively. After training, each word is associated with a unique distributed vector. Each sentence in the dataset and knowledge memory is encoded as a vector of a fixed number of dimensions, which is the average summation of the distributed representation of words with respect to the sentence.

3.2 Temporality-enhanced GRU

A significant characteristic of quiz bowl is that users are given a sequence of ordered sentences describing different aspects of an answer. These sentences have a property called pyramidality, which means that sentences ordered ahead deliver harder and more obscure cues, while the later ones are 'giveaway'. When users read these sentences one by one, they are likely to recall cues in the previously read sentences and collaboratively use all of comprehended sentences to give the answer. As a result, we here introduce T-GRU, which is a variation of the standard GRU in Cho et al. (2014), to encode the temporal cues in the ordered sequence sentences.

The lower part of Fig. 1b illustrates the details of T-GRU. T-GRU can enhance the temporal cues from previous T-GRUs by introducing attention mechanism. $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{t-1}$ are the hidden output of previous T-GRUs at time $1, 2, \ldots, t-1$, respectively. $\boldsymbol{x}_t$ is the distributed representation of the input sentence and $\boldsymbol{h}(t)$ is the hidden output of the current T-GRU at time t. σ and tanh are the sigmoid activation function and tanh activation function, respectively.

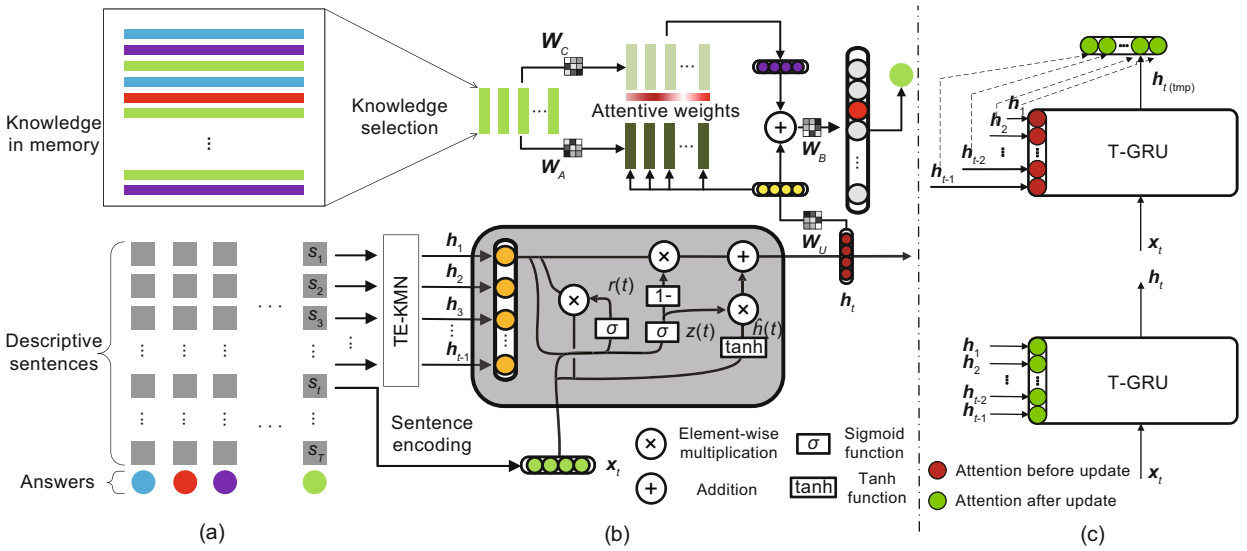

Fig. 1 Overview of our proposed temporality-enhanced knowledge memory network (TE-KMN)

(a) The lower left is the training data. For each answer, there are several ordered descriptive sentences which gradually describe the different aspects of one answer (i.e, one entity). The ordered ahead sentences remark fewer and harder cues with respect to a given answer, and the later sentences are 'giveaways'. The upper left is a knowledge memory consisting of auxiliary external knowledge that will be used during reading each sentence. In this study, the knowledge is from Wikipedia. (b) Given an answer, assume there are T ordered sentences in total describing different aspects of this answer (i.e., $s_1, s_2, \ldots, s_T$). Each sentence s_i is encoded to a distributed vector $\boldsymbol{x_i}$. The hidden representation $\boldsymbol{h_i}$ $(1 \leq i \leq t-1)$ of each sentence will influence the learning of hidden representation $\boldsymbol{h_t}$ of $\boldsymbol{x_t}$. The hidden representation $\boldsymbol{h_t}$ is then transmitted to the knowledge memory network to be further handled by the relevant auxiliary knowledge that relates to sentence s_t. (c) The update of the attention mechanism in T-GRU. The upper right is the T-GRU with attention before update. The lower right is the same T-GRU with attention after update. Parameters other than α_i remain intact during the update. $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{t-1}$ are the hidden outputs from the T-GRUs that read the previous sentences. The red and green solid circles represent the values of α_i before and after update, respectively. References to color refer to the online version of this figure

To strengthen the cues triggered by the previous sentences, we add another attention gate $\boldsymbol{m}_t$, which merges $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{t-1}$ into a uniform vector. $\boldsymbol{m}_t$ is updated by the following equation:

$$\boldsymbol{m}_t = \sum_{i=1}^{t-1} \alpha_i \cdot \boldsymbol{h}_i, \tag{1}$$

where $\alpha_i \in [0,1]$ and $\sum_{i=1}^{t-1} \alpha_i = 1$. α_i are initialized to $1/(t-1)$ at the beginning of the training. Because α_i can directly decide the way in which T-GRU accepts the previous cues, the values of α_i influence the $\boldsymbol{m}_t$ of T-GRU at each time step. Therefore, we argue that α_i contribute to the performance of our proposed model in the quiz bowl task.

The update of α_i is shown in Fig. 1c and the process of update can be divided into two parts:

1. T-GRU generates $\boldsymbol{h}_{t(\mathrm{tmp})}$ using $\boldsymbol{x}_t$ and the hidden outputs from previous T-GRUs that read in the sentences in a sequence order with their corresponding α_i before the update. Then all α_i's are updated using the following equation:

$$\alpha_i = \frac{\exp(\boldsymbol{h}_{t(\mathrm{tmp})}^{\mathrm{T}} \cdot \boldsymbol{h}_i)}{\sum_{i=1}^{t-1} \exp(\boldsymbol{h}_{t(\mathrm{tmp})}^{\mathrm{T}} \cdot \boldsymbol{h}_i)}. \tag{2}$$

2. The updated values of α are the new weights for all of the previous hidden outputs. The current T-GRU generates its new hidden output $\boldsymbol{h}_t$ using $\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_{t-1}$ with their corresponding new α_i.

The gates in T-GRU can modulate the interactions between T-GRU and its environment. The update gate $\boldsymbol{z}_t$ decides how much the unit updates its activation or content and is computed as follows:

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{m}_t). \tag{3}$$

The reset gate effectively makes the unit act as if it were reading the first symbol of an input sequence, allowing it to forget the previously computed state. The reset gate $\boldsymbol{r}_t$ is computed in a manner similar to the update gate:

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{m}_t). \tag{4}$$

The candidate activation $\hat{\boldsymbol{h}}_t$ is computed as

$$\hat{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{\hat{h}}\boldsymbol{x}_t + \boldsymbol{U}_{\hat{h}}(\boldsymbol{r}_t \odot \boldsymbol{m}_t)), \tag{5}$$

where '$\odot$' is the element-wise multiplication operation. The activation $\boldsymbol{h}_t$ of T-GRU at time t is a linear interpolation between attention gate $\boldsymbol{m}_t$ (i.e., all of the previously hidden outputs) and candidate activation $\hat{\boldsymbol{h}}_t$:

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t)\boldsymbol{m}_t + \boldsymbol{z}_t\hat{\boldsymbol{h}}_t. \tag{6}$$

Note that $\boldsymbol{W}_*$ and $\boldsymbol{U}_*$ are weight matrices that are almost the same as their counterparts in the standard GRU. The architecture and implementation of T-GRU are based on a standard GRU.

Based on the distributed representation of input sentences, T-GRU in our method enhances the temporal cues in a sequence of ordered sentences in quiz bowl and therefore constructs the semantic links between sentences in terms of specific aspects.

3.3 Knowledge memory network

In general, when reading a sentence, human beings will recall relevant knowledge that relates to that sentence to better understand the sentence. To endow our architecture with the ability to employ the already memorized external knowledge, we introduce the knowledge memory network to assist T-GRU and boost the understanding of the sentence currently being read. Given a sequence of ordered sentences $s_1, s_2, \ldots, s_T$, where T is the number of sentences in a question, and $\boldsymbol{x}_i$ the distributed representation of the corresponding sentence, the memory network will select the relevant fact $\boldsymbol{k}_j$ to sentence $\boldsymbol{x}_i$ using a score function:

$$\text{score}(\boldsymbol{x}_i, \boldsymbol{k}_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{k}_j}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{k}_j\|}. \tag{7}$$

The network selects the top K facts of each sentence and puts them into the memory (i.e., set Q). The knowledge memory network learns to understand the stored knowledge combining the hidden output $\boldsymbol{h}_t$ of T-GRU and gives an answer to the current input sentence. The stored knowledge set Q is converted into memory vector set A and output vector set C using the following equations:

$$\boldsymbol{a}_i = \boldsymbol{W}_A\boldsymbol{q}_i, \quad \boldsymbol{c}_i = \boldsymbol{W}_C\boldsymbol{q}_i. \tag{8}$$

The hidden output $\boldsymbol{h}_t$ of T-GRU is also embedded to obtain an internal state $\boldsymbol{u}_t$:

$$\boldsymbol{u}_t = \boldsymbol{W}_U\boldsymbol{h}_t. \tag{9}$$

In the embedding space, we compute the match between $\boldsymbol{u}_t$ and each memory $\boldsymbol{a}_i$ by taking the inner product followed by a softmax function:

$$p_i = \text{softmax}(\boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{a}_i), \tag{10}$$

where $\text{softmax}(\boldsymbol{z}) = \mathrm{e}^{\boldsymbol{z}}/\sum_j \mathrm{e}^{z_j}$ and $\boldsymbol{p} = [p_1, p_2, \ldots]^{\mathrm{T}}$ is a probability vector over the inputs. Then the response vector from memory $\boldsymbol{o}_t$ is a sum over the transformed inputs $\boldsymbol{c}_i$, weighted by the probability vector from the input:

$$\boldsymbol{o}_t = \sum_i p_i\boldsymbol{c}_i. \tag{11}$$

The input embedding $\boldsymbol{u}_t$ together with $\boldsymbol{o}_t$ is then passed through a final weight matrix $\boldsymbol{W}_B$ and a softmax function to produce the predicted label $\boldsymbol{b}_t$:

$$\boldsymbol{b}_t = \text{softmax}(\boldsymbol{W}_B(\boldsymbol{o}_t + \boldsymbol{u}_t)). \tag{12}$$

3.4 Training

Before training TE-KMN, we pre-train the embedding of each word in the dataset and knowledge memory. After that, we model the distributed representation of each sentence and each fact with the trained word vectors. Next, we select the top K most relative facts that relate to each reading sentence in a quiz bowl question using Eq. (7) and store them into the knowledge memory. Then we feed the sentences of each question in order into our model one by one. TE-KMN produces a probability vector $\boldsymbol{b}_t$ over the inputs at each time step. According to the predicted label $\boldsymbol{b}_t$, the loss function for quiz bowl can be written as

$$l_t = -\sum_{i=1}^{n} y_i \cdot \log(b_{ti}), \tag{13}$$

where $\boldsymbol{y}$ is the true label of the answer. With the loss function (13), by denoting all the parameters in our model as $\boldsymbol{\theta}$, the optimization problem in the training process is

$$\min_{\boldsymbol{\theta}} \; L(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \sum_D \sum_S \; l_t + \lambda \left\|\boldsymbol{\theta}\right\|_2^2, \tag{14}$$

where D represents the training data, S all the sentences of a descriptive paragraph, and $\lambda > 0$

a hyper-parameter to trade-off training loss and regularization.

To minimize the objective function, we use the stochastic gradient descent (SGD) method in the back propagation algorithm. At time step t, parameter $\boldsymbol{\theta}$ is updated as follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\rho}{\sqrt{\sum_{i=1}^{t} \boldsymbol{g}_i^2}} \boldsymbol{g}_t, \tag{15}$$

where ρ is the initial learning rate and $\boldsymbol{g}_t$ the subgradient at time step t. The whole training process is summarized in Algorithm 1.

Algorithm 1 TE-KMN for quiz bowl

Input: Quiz bowl question set Q, answer set A, knowledge set K, selected knowledge C, and the number of iterations m.

Output: Given each sentence of a quiz bowl question, predict the answer.

```
Initialize all parameters of TE-KMN
for i = 1 to m do
  for q ∈ Q do
    C = ∅
    for s ∈ q and s is a sentence of q do
      C += knowledgeselection(s)
    end for
    for s ∈ q and s is a sentence of q do
      h_t = T-GRU(h_1, h_2 ..., h_{t-1}, x_i)
      b_t = KMN(h_t, C)
      Calculate the loss l_t with b_t
    end for
    Accumulate the training loss
    Update parameters by the stochastic gradient descent method
  end for
end for
```

4 Experiments

4.1 Experimental setup

To empirically evaluate and validate our proposed architecture TE-KMN, we conduct the experiments on a dataset of quiz bowl question answering that involves fields of history, literature, places, etc. This dataset was originally published by National Academic Quiz Tournaments (NAQT, https://sites.google.com/view/hcqa/data). Each question in the dataset consists of a sequence of ordered sentences and a corresponding answer. We filter the questions with respect to the answers that appear substantially less often than others. The statistics of the data is summarized in Table 2.

Table 2 Statistics of the quiz bowl dataset

Dataset	Number of questions	Number of sentences	Average number of sentences in a question
Training	9059	43 271	4.78
Development	1013	4921	4.86
Test	4319	20 901	4.84
All	14 391	69 093	4.80

The knowledge we use in the experiments is mainly from two resources: (1) We crawl the Wikipedia webpages that relate to all the answers in the dataset. The summation part of each page is taken as the auxiliary knowledge. (2) We maintain the knowledge in the training set. We replace unclear subjects in each sentence with its associated answer. Therefore, we form a knowledge memory that includes more than 49 747 facts. A total of 6476 facts are from Wikipedia and the remaining facts are from the training set.

4.2 Evaluation metric

Considering that the quiz bowl question answering problem is similar to a classification task, we evaluate the performance of our proposed TE-KMN based on a widely used classification evaluation metric, which is defined as follows:

$$\text{Accuracy} = \frac{|\{d \in D \mid r_d = r_{\text{true}}\}|}{|D|}, \tag{16}$$

where $|D|$ is the number of questions in the test dataset, d indicates a single quiz bowl question, r_d is the answer conjectured by the evaluated method, and r_{true} is the correct answer of question d.

4.3 Baselines and parameter setting

To demonstrate the efficiency and effectiveness of our proposed TE-KMN, we compare it with six popular content-based algorithms and its two degenerated versions, which are listed as follows:

1. BOW (Boyd-Graber et al., 2012): a BOW method is used to represent the textual content, and a naive-Bayes model is used to identify the answer.

2. QANTA (Iyyer et al., 2014): a DT-RNN is developed and extended to combine predictions across

sentences to produce a question answering neural network with trans-sentential averaging. QANTA takes word embedding and the dependency tree as the input, and predicts the corresponding answer.

3. DAN (Iyyer et al., 2015): DAN feeds an unweighted average of word vectors through multiple hidden layers before classification.

4. HCNN (Zheng et al., 2015): HCNN is a stacked CNN to learn both sentence representation and paragraph representation of the original text and map these features to their corresponding answers.

5. GRU (Cho et al., 2014): A standard GRU receives a sentence at each time step and is used to model the interaction in a sequence of ordered sentences without attention mechanism.

6. BiLSTM (Graves et al., 2005): BiLSTM takes the word vector at each time step. The information propagates directionally through the whole network.

7. TE-KMN$^{-\mathrm{K}}$: the knowledge memory network is removed from TE-KMN and the influence of auxiliary knowledge is ignored. TE-KMN$^{-\mathrm{K}}$ maintains only the connections between a sentence and its previous sentences.

8. TE-KMN$^{-\mathrm{T}}$: Similar to TE-KMN$^{-\mathrm{K}}$, TE-KMN$^{-\mathrm{T}}$ is another degenerated version of TE-KMN which neglects the enhancement of the temporal cues from the previous sentences but maintains the knowledge memory network.

The words in the dataset and knowledge memory are represented by a 300-D vector pre-trained by the word2vec tool (Mikolov et al., 2013) filtering the words appearing only once or over 400 times. The input sentences and facts are represented with a 300-D vector with the trained word vectors. We select five most relevant facts of each reading sentence from the knowledge memory. The weights of neural networks are randomly initialized by a Gaussian distribution with zero mean. The hyper-parameters and parameters that achieve the best performance on the development set are chosen to conduct the test evaluation.

4.4 Performance comparison

To evaluate the performance of our proposed architecture, we conduct several experiments on the quiz bowl dataset.

As mentioned previously, we argue that the appropriate capturing of temporal cues in a sequence of descriptive sentences and the effective leveraging of auxiliary knowledge that relates to each sentence are two essential factors in quiz bowl questions. To validate that these two factors could improve the performance of our proposed architecture, we degenerate our proposed approach into two simplified versions: TE-KMN$^{-\mathrm{T}}$ and TE-KMN$^{-\mathrm{K}}$.

Table 3 shows the average accuracy after giving the first i ($i = 1, 2, \ldots, 5$) sentence(s) for each quiz bowl question. Table 4 presents the average accuracy when different ratios of training data are used.

With these experimental results, we can have several interesting observations:

1. DAN, QANTA, HCNN, and BiLSTM have better performances than GRU and TE-KMN$^{-\mathrm{K}}$, because GRU and TE-KMN$^{-\mathrm{K}}$ take sentence vectors as the input, losing a lot of semantics when averaging the summation of each word vector with respect to the corresponding sentence.

2. TE-KMN and TE-KMN$^{-\mathrm{T}}$ have promising experimental results when the training data is incomplete, which shows the power of leveraging auxiliary external knowledge in answering quiz bowl questions. The knowledge memory network memorizes the external knowledge and helps the model give a more appropriate answer using the auxiliary knowledge that relates to each reading sentence.

3. TE-KMN$^{-\mathrm{K}}$ is more competitive in the later positions compared with the standard GRU, which

Table 3 Accuracy of obtaining the right answer given different numbers of first sentences of each quiz bowl question

Number of first sentences given	Accuracy (%)								
	BOW	QANTA	DAN	HCNN	GRU	BiLSTM	TE-KMN$^{-\mathrm{K}}$	TE-KMN$^{-\mathrm{T}}$	TE-KMN
1	34.24	43.21	43.83	44.93	40.76	50.53	41.33	63.52	**66.07**
2	50.31	56.93	57.51	56.42	51.56	62.04	52.50	67.61	**70.77**
3	51.12	57.26	57.93	59.19	53.11	63.27	54.74	69.44	**72.91**
4	51.40	57.45	58.02	60.25	54.21	63.61	56.32	69.93	**73.94**
5	51.72	57.59	58.11	61.44	55.46	64.02	57.70	70.22	**74.46**

Table 4 Accuracy with different ratios of training data

Ratio of	Accuracy (%)								
training data	BOW	QANTA	DAN	HCNN	GRU	BiLSTM	TE-KMN^{-K}	TE-KMN^{-T}	TE-KMN
25%	22.38	30.75	32.50	31.94	26.17	33.31	26.40	46.22	**47.70**
50%	31.42	42.88	43.59	46.09	39.00	46.85	39.94	57.03	**60.41**
75%	40.37	49.93	51.78	53.62	46.23	55.20	47.04	63.59	**67.00**
100%	51.91	57.64	58.22	61.51	55.53	64.19	57.85	70.35	**74.80**

proves that the temporal cues in a sequence of ordered sentences have a positive effect on answering performance of quiz bowl questions. The sentences ordered ahead contain cues that inspire the model to better understand the sentences in later order. Thus, the temporal cue is able to improve the performance of TE-KMN.

4. TE-KMN^{-T} outperforms the standard GRU and achieves better performance when there is less training data, which indicates that our knowledge memory network can greatly boost TE-KMN with auxiliary knowledge.

5. TE-KMN and TE-KMN^{-T} have promising leads over other content-based baselines at the immediately prior positions (i.e., the 1st and 2nd sentences). One possible reason is that the prior sentences contain few cues that directly indicate the correct answer. It is hard for content-based methods to capture the cues of sentences ordered ahead and build semantic links to their answers. TE-KMN and TE-KMN^{-T} are more sensitive to the latent cues because the knowledge memory network can reason using the auxiliary external knowledge and the obscure cues.

6. TE-KMN^{-K} has a better performance than TE-KMN^{-T}, which shows that the knowledge memory network contributes more than T-GRU to our proposed model in answering quiz bowl questions.

7. TE-KMN achieves the best performance compared with other models. The results demonstrate that the combined use of temporal cues in a sequence of ordered sentences and auxiliary external knowledge can improve the performance of our model in answering quiz bowl questions.

4.5 Analysis of auxiliary external knowledge

The knowledge memory contains auxiliary external knowledge from Wikipedia and the training set. Wikipedia accounts for 13.02% and the remaining 86.98% of the knowledge is from the training set. To validate which part of the auxiliary knowledge is more sensitive to the cues in quiz bowl questions, we train TE-KMN^{wiki} with knowledge only from Wikipedia and TE-KMN^{train} with knowledge only from the training set. Given different numbers of the first sentences of each quiz bowl question, Table 5 illustrates the accuracy of TE-KMN^{wiki}, TE-KMN^{train}, TE-KMN^{-K}, and TE-KMN.

As shown in Table 5, given the 1st sentence of quiz bowl questions, TE-KMN^{wiki} is slightly better than TE-KMN^{-K}, and TE-KMN^{train} represents a promising improvement (53.47%) over TE-KMN^{wiki}. However, the disparity decreases to 17.44% after the first five sentences are given. Compared with the performance given the 1st sentence, the performance of TE-KMN^{wiki} has an improvement of 45.70% at the 5th sentence while the performance of TE-KMN^{train} has an improvement of 11.49%. The results reveal that the knowledge from the training set remarks hard and obscure cues that are instrumental in understanding sentences ordered ahead in quiz bowl questions (i.e., ordered ahead sentences remark fewer

Table 5 Accuracy of obtaining the right answer when using different knowledge memories

Number of	Accuracy (%)			
first sentences given	TE-KMN^{-K}	TE-KMN^{wiki}	TE-KMN^{train}	TE-KMN
1	41.33	42.08	64.58	**66.07**
2	52.50	53.28	68.71	**70.77**
3	54.74	55.66	70.58	**72.91**
4	56.32	58.60	71.45	**73.94**
5	57.70	61.31	72.00	**74.46**

cues with respect to the answer), while the knowledge from Wikipedia contains more obvious and direct cues which can clearly match the cues in later ordered sentences (i.e., later sentences deliver more helpful indications to capture the answer).

4.6 Attention mechanism in T-GRU

To verify that the attention mechanism in T-GRU can capture the temporal cues in a sequence of ordered sentences, we take the question in Table 1 as an example and illustrate the attention before and after training respectively in Fig. 2. Note that the question in Table 1 is not used to train the model.

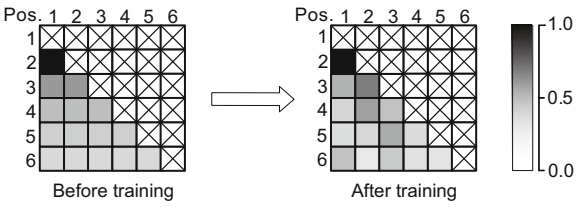

Fig. 2 Update of attention in a sequence of ordered sentences shown in Table 1

Pos. represents the position of each sentence in the question. Each row exhibits the attention changes of one sentence with each of its previous sentences. The darker color indicates a greater ratio of the corresponding attention

As shown in Fig. 2, considering the temporal cues labeled in Table 1, sentences ordered ahead have a positive influence on the later ones when they share the same indicative temporal cues to the answer. T-GRU is able to capture these temporal cues after being well trained.

5 Conclusions and future work

We have introduced a neural network method that incorporates the temporal cues in a sequence of ordered sentences with respect to a given question and auxiliary external knowledge to improve the performance in answering quiz bowl questions. Empirical results on the NAQT quiz bowl dataset show that TE-KMN outperforms other state-of-the-art methods, which proves that temporal cues and auxiliary external knowledge contribute to the better performance of our proposed model.

In the future, we hope to enhance the reasoning ability (Pan, 2016; Zhuang et al., 2017) of our proposed model and find a more effective way of dealing with temporality in quiz bowl question answering. We are also looking forward to applying our model to other deep learning tasks to prove its effectiveness.

References

Bahdanau D, Cho K, Bengio Y, 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. https://arxiv.org/abs/1409.0473

Bao JW, Duan N, Zhou M, et al., 2014. Knowledge-based question answering as machine translation. 52nd Annual Meeting of the Association for Computational Linguistics, p.967-976. https://doi.org/10.3115/v1/p14-1091

Barrón-Cedeño A, Filice S, Martino GDS, et al., 2015. Thread-level information for comment classification in community question answering. 53rd Annual Meeting of the Association for Computational Linguistics, p.687-693. https://doi.org/10.3115/v1/p15-2113

Bilotti MW, Elsas JL, Carbonell JG, et al., 2010. Rank learning for factoid question answering with linguistic and semantic constraints. 19th ACM Conf on Information and Knowledge Management, p.459-468. https://doi.org/10.1145/1871437.1871498

Boyd-Graber JL, Satinoff B, He H, et al., 2012. Besting the quiz master: crowdsourcing incremental classification games. Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, p.1290-1301.

Carr CE, 1993. Processing of temporal information in the brain. *Ann Rev Neurosci*, 16(1):223-243. https://doi.org/10.1016/S0166-4115(96)80051-3

Chen D, Bolton J, Manning CD, 2016. A thorough examination of the CNN/daily mail reading comprehension task. 54th Annual Meeting of the Association for Computational Linguistics, p.2358-2367. https://doi.org/10.18653/v1/p16-1223

Cho K, van Merrienboer B, Gülçehre Ç, et al., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Conf on Empirical Methods in Natural Language Processing, p.1724-1734. https://doi.org/10.3115/v1/d14-1179

Chorowski J, Bahdanau D, Serdyuk D, et al., 2015. Attention-based models for speech recognition. Advances in Neural Information Processing Systems, p.577-585.

Ding S, Cong G, Lin C, et al., 2008. Using conditional random fields to extract contexts and answers of questions from online forums. 46th Annual Meeting of the Association for Computational Linguistics, p.710-718.

Figueroa A, Atkinson J, 2011. Maximum entropy context models for ranking biographical answers to open-domain definition questions. 25th AAAI Conf on Artificial Intelligence, p.1173-1179.

Fillmore CJ, 1976. Frame semantics and the nature of language. *Ann New York Acad Sci*, 280(1):20-32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

Ghazvininejad M, Brockett C, Chang M, et al., 2017. A knowledge-grounded neural conversation model. arXiv:1702.01932. https://arxiv.org/abs/1702.01932

Graves A, Fernández S, Schmidhuber J, 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. 15th Int Conf on Artificial Neural Networks, p.799-804.

Huang J, Zhou M, Yang D, 2007. Extracting chatbot knowledge from online discussion forums. 20th Int Joint Conf on Artifical Intelligence, p.423-428.

Ivry RB, 1996. The representation of temporal information in perception and motor control. *Curr Opin Neurobiol*, 6(6):851-857. https://doi.org/10.1016/s0959-4388(96)80037-7

Iyyer M, Boyd-Graber JL, Claudino LMB, et al., 2014. A neural network for factoid question answering over paragraphs. Conf on Empirical Methods in Natural Language Processing, p.633-644. https://doi.org/10.3115/v1/d14-1070

Iyyer M, Manjunatha V, Boyd-Graber JL, et al., 2015. Deep unordered composition rivals syntactic methods for text classification. 53rd Annual Meeting of the Association for Computational Linguistics, p.1681-1691. https://doi.org/10.3115/v1/p15-1162

Jeon J, Croft WB, Lee JH, 2005. Finding similar questions in large question and answer archives. 14th ACM Int Conf on Information and Knowledge Management, p.84-90. https://doi.org/10.1145/1099554.1099572

Joty SR, Barrón-Cedeño A, Martino GDS, et al., 2015. Global thread-level inference for comment classification in community question answering. Conf on Empirical Methods in Natural Language Processing, p.573-578. https://doi.org/10.18653/v1/d15-1068

Jurczyk P, Agichtein E, 2007. Discovering authorities in question answer communities by using link analysis. 16th ACM Conf on Information and Knowledge Management, p.919-922. https://doi.org/10.1145/1321440.1321575

Kalchbrenner N, Grefenstette E, Blunsom P, 2014. A convolutional neural network for modeling sentences. 52nd Annual Meeting of the Association for Computational Linguistics, p.655-665. https://doi.org/10.3115/v1/p14-1062

Le QV, Mikolov T, 2014. Distributed representations of sentences and documents. 31st Int Conf on Machine Learning, p.1188-1196.

Li B, Lyu MR, King I, 2012. Communities of Yahoo! Answers and Baidu Zhidao: complementing or competing? Int Joint Conf on Neural Networks, p.1-8. https://doi.org/10.1109/ijcnn.2012.6252435

Luong T, Pham H, Manning CD, 2015. Effective approaches to attention-based neural machine translation. Conf on Empirical Methods in Natural Language Processing, p.1412-1421. https://doi.org/10.18653/v1/d15-1166

Ma X, Hovy EH, 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. 54th Annual Meeting of the Association for Computational Linguistics, p.1064-1074. https://doi.org/10.18653/v1/p16-1101

Mikolov T, Karafiát M, Burget L, et al., 2010. Recurrent neural network based language model. 11th Annual Conf of the Int Speech Communication Association, p.1045-1048.

Mikolov T, Chen K, Corrado G, et al., 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Minsky M, 1991. Society of mind: a response to four reviews. *Artif Intell*, 48(3):371-396. https://doi.org/10.1016/0004-3702(91)90036-J

Nakashole N, Mitchell TM, 2015. A knowledge-intensive model for prepositional phrase attachment. 53rd Annual Meeting of the Association for Computational Linguistics, p.365-375. https://doi.org/10.3115/v1/p15-1036

Navigli R, Velardi P, 2010. Learning word-class lattices for definition and hypernym extraction. 48th Annual Meeting of the Association for Computational Linguistics, p.1318-1327.

Pan B, Li H, Zhao Z, et al., 2017. MEMEN: multi-layer embedding with memory networks for machine comprehension. arXiv:1707.09098. https://arxiv.org/abs/1707.09098

Pan Y, 2016. Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409-413. https://doi.org/10.1016/J.ENG.2016.04.018

Qiu X, Huang X, 2015. Convolutional neural tensor network architecture for community-based question answering. 24th Int Joint Conf on Artificial Intelligence, p.1305-1311.

Rush AM, Chopra S, Weston J, 2015. A neural attention model for abstractive sentence summarization. Conf on Empirical Methods in Natural Language Processing, p.379-389. https://doi.org/10.18653/v1/d15-1044

Schweppe J, Rummer R, 2013. Attention, working memory, and long-term memory in multimedia learning: an integrated perspective based on process models of working memory. *Ed Psychol Rev*, 26(2):285-306. https://doi.org/10.1007/s10648-013-9242-2

Shah C, Pomerantz J, 2010. Evaluating and predicting answer quality in community QA. 33rd Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.411-418. https://doi.org/10.1145/1835449.1835518

Shen Y, Rong W, Sun Z, et al., 2015. Question/Answer matching for CQA system via combining lexical and sequential information. 29th AAAI Conf on Artificial Intelligence, p.275-281.

Sukhbaatar S, Szlam A, Weston J, et al., 2015. End-to-end memory networks. Advances in Neural Information Processing Systems, p.2440-2448.

Sun R, Jiang J, Tan YF, et al., 2005. Using syntactic and semantic relation analysis in question answering. 14th Text REtrieval Conf.

Sutskever I, Vinyals O, Le QV, 2014. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, p.3104-3112.

Wang B, Wang X, Sun C, et al., 2010. Modeling semantic relevance for question-answer pairs in web social communities. 48th Annual Meeting of the Association for Computational Linguistics, p.1230-1238.

Wang M, 2006. A Survey of Answer Extraction Techniques in Factoid Question Answering. https://cs.stanford.edu/people/mengqiu/publication/LSII-LitReview.pdf

Wei X, Huang H, Nie L, et al., 2017. I know what you want to express: sentence element inference by incorporating external knowledge base. *IEEE Trans Knowl Data Eng*, 29(2):344-358. https://doi.org/10.1109/TKDE.2016.2622705

Xue X, Jeon J, Croft WB, 2008. Retrieval models for question and answer archives. 31st Int ACM SIGIR Conf on Research and Development in Information Retrieval, p.475-482. https://doi.org/10.1145/1390334.1390416

Yang B, Mitchell TM, 2017. Leveraging knowledge bases in LSTMs for improving machine reading. 55th Annual Meeting of the Association for Computational Linguistics, p.1436-1446. https://doi.org/10.18653/v1/p17-1132

Yang Y, Zhuang Y, Wu F, et al., 2008. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. *IEEE Trans Multimed*, 10(3):437-446. https://doi.org/10.1109/TMM.2008.917359

Yao X, Durme BV, 2014. Information extraction over structured data: question answering with freebase. 52nd Annual Meeting of the Association for Computational Linguistics, p.956-966. https://doi.org/10.3115/v1/p14-1090

Zheng S, Bao H, Zhao J, et al., 2015. A novel hierarchical convolutional neural network for question answering over paragraphs. IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Technology, p.60-66. https://doi.org/10.1109/WI-IAT.2015.20

Zhou G, Cai L, Zhao J, et al., 2011. Phrase-based translation model for question retrieval in community question answer archives. 49th Annual Meeting of the Association for Computational Linguistics, p.653-662.

Zhuang Y, Yang Y, Wu F, 2008. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Trans Multimed*, 10(2):221-229. https://doi.org/10.1109/TMM.2007.911822

Zhuang Y, Wu F, Chen C, et al., 2017. Challenges and opportunities: from big data to knowledge in AI 2.0. *Front Inform Technol Electron Eng*, 18(1):3-14. https://doi.org/10.1631/FITEE.1601883