

Beyond top-k: knowledge reasoning for multi-answer temporal questions based on reverification framework (#81547)

1

First submission

Guidance from your Editor

Please submit by **15 Aug 2023** for the benefit of the authors (and your token reward) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).

1 Latex file(s)
4 Raw data file(s)



Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor

 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).




BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [Peerj standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [Peerj policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Conclusions are well stated, linked to original research question & limited to supporting results.



The best reviewers use these techniques

Tip

Example

Support criticisms with evidence from the text or from other sources

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 - the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

Beyond top-k: knowledge reasoning for multi-answer temporal questions based on reverification framework

Jun-ping Yao¹, Cong Yuan^{Corresp., 1}, Xiao-jun Li¹, Yi Su¹, Yi-jing Wong¹

¹ Xi'an Research Inst. of High-Tech, xi'an, ShaanXi, China

Corresponding Author: Cong Yuan

Email address: yuancong001@126.com

Answer sorting and filtering is two closely related steps for determining the answer to a question. Answer sorting is designed to produce an ordered list of scores based on Top-k and contextual criteria. Answer filtering optimizes the selection according to other criteria, such as the range of time constraints the user expects. However, the unclear number of answers and time constraints, as well as the high score of false positive results, indicate that the traditional sorting and selection methods cannot guarantee the quality of answers to multi-answer questions. Therefore, this study proposes MATQA, a component based on multi-answer temporal question reasoning, and adopts the revalidation framework to improve the existing Top-k answer expression form. First, the highly correlated subgraph is selected by calculating the scores of the boot node and the related fact node. Second, the subgraph attention inference module is introduced to determine the initial answer with the highest probability. Finally, the alternative answers are clustered at the semantic level and the time constraint level. Meanwhile, the candidate answers with similar types and high scores but do not satisfy the semantic constraints or the time constraints are eliminated to ensure the number and accuracy of final answers. Experiments on Multi-answer TimeQuestions demonstrate the advantages of MATQA over traditional question answering schemes.

Beyond Top-k: Knowledge Reasoning for Multi-Answer Temporal Questions Based on Reverification Framework

Junping Yao¹, Cong Yuan¹, Xiaojun Li¹, Yijing Wang¹, and Yi Su¹

¹Xi'an Research Inst. of High-Tech, ShaanXi Xi'an, 710025, China

Corresponding author:

Cong Yuan¹

Email address: xi_anwyj@163.com

ABSTRACT

Answer sorting and filtering is two closely related steps for determining the answer to a question. Answer sorting is designed to produce an ordered list of scores based on Top-k and contextual criteria. Answer filtering optimizes the selection according to other criteria, such as the range of time constraints the user expects. However, the unclear number of answers and time constraints, as well as the high score of false positive results, indicate that the traditional sorting and selection methods cannot guarantee the quality of answers to multi-answer questions. Therefore, this study proposes MATQA, a component based on multi-answer temporal question reasoning, and adopts the revalidation framework to improve the existing Top-k answer expression form. First, the highly correlated subgraph is selected by calculating the scores of the boot node and the related fact node. Second, the subgraph attention inference module is introduced to determine the initial answer with the highest probability. Finally, the alternative answers are clustered at the semantic level and the time constraint level. Meanwhile, the candidate answers with similar types and high scores but do not satisfy the semantic constraints or the time constraints are eliminated to ensure the number and accuracy of final answers. Experiments on Multi-answer TimeQuestions demonstrate the advantages of MATQA over traditional question answering schemes.

INTRODUCTION

A high-quality question answering model(Jia et al., 2018) is sensitive to constraints on semantic quantitative boundaries of input questions. Mainstream question answering approaches intentionally reduce the task to a “one best answer per question” scheme. But in practice, many temporal problems are open-ended and ambiguous, with multiple valid answers (or groups of answers), and often all of these answers must be captured so as to answer one question(Rubin et al., 2022). (Min et al., 2020) pointed out that over 50% of the query intent in Google search is ambiguous. In order to show strong reasoning ability, the question answering model not only needs to give the answer with high confidence, but also the exact number of answers. Nevertheless, the existing question answering systems can only obtain the Top-k list of a single answer by scoring ranking(Wang et al., 2021). When there are multiple valid answers to a temporal question, users cannot directly obtain valid solutions with high accuracy and accurate number.

Multi-answer reasoning stems from reading comprehension. Currently, multi-answer reasoning is based on unstructured text databases and aims to retrieve all answers from multiple passages that satisfy the intention of a question. Limited by the ambiguity of natural language, questions can be interpreted with multiple meanings, so multiple answers will be recalled from the text. Limitations of existing work(Rubin et al., 2022; Min et al., 2020; Shao and Huang, 2022) concern various forms of paragraph parsing and question and ambiguous answer matching. Retrieval and reading paradigm is the major method of text paragraph multi-answer reasoning, which involves the correct reasoning of long sequence of paragraphs in the computation process, the maximum number of paragraphs supported by hardware, and the mutual restriction between the two. For example, AMBIGNQ(Min et al., 2020) uses Bert dual encoding model to retrieve 100 paragraphs and reorder them, concatenating the question with the top paragraph to generate the answer in order in an end-to-end system. (Shao and Huang, 2022) used the “recall-revalidation”

46 framework to avoid the problem of multiple answers sharing a limited reading budget by separating the
47 reasoning process of each answer, and to better verify the answer with re-found evidence. (Liu et al.,
48 2021) alleviated the error propagation problem by explicitly modeling three matching granularities of
49 paragraph recognition, sentence selection and answer extraction through MGRC, an end-to-end reading
50 comprehension model.

51 Multi-answer reasoning based on knowledge base is in its infancy. (Moon et al., 2022) in 2022
52 proposed RxWhyQA, a clinical question answering dataset for multi-answer questions, and pointed out
53 that clinical reasoning and decision making are still constrained by multi-answer questions. In the same
54 year, (Zhong et al., 2022) proposed RoMQA, a benchmark for multi-evidence, multi-answer question
55 answering. Despite revealing the shortcomings of existing zero-sample, small-sample learning and
56 supervised learning schemes on this benchmark, they failed to propose a clear solution. In the field of
57 temporal question answering, there is no perfect method to solve the multi-answer reasoning problem.
58 This study aims to extend the multi-answer question answering to the field of temporal knowledge question
59 answering. Based on the knowledge base, the main work is to ensure the numerical quality of valid
60 answers to temporal questions. Although the existing unstructured question answering (Cao et al., 2021)
61 and knowledge-based question answering schemes have achieved good results, there are still the following
62 new challenges in the field of multi-answer temporal question reasoning:

63 **The number of answers is undetermined.** In practice, there exists a class of multi-answer problems
64 in which the answer consists of multiple entities or attributes. For example, in temporal question answering,
65 there are usually more than one candidate answer to be accepted within a given time interval. However,
66 the traditional Top-k list only shows the ranking of answer scores and cannot limit the specific number of
67 answers to the question, so the user has to determine the number of answers by guessing. As shown in
68 Figure 1, the question “who held the position of secretary of state when Andrew Jackson was president?”
69 has three accurate answers, “Martin Van Buren, Edward Livingston, and Louis McLane.” In the traditional
70 answer representation mode, users can only get a few answers with high scores according to the Top-K
71 list, but they cannot be sure about the specific number of answers that meet the semantic conditions.

72 **Answers with higher scores are not necessarily correct.** There is a special case where a specific
73 number of answers to a question has been given, but there are still wrong answers among the candidates.
74 Therefore, in general cases, there are still false positives for answers with high scores. In the list of
75 Top-5 answers in Figure 1, only the first two are standard answers, the answer with the third high score is
76 wrong, and the third accurate answer is not obtained by reasoning, so there are still errors in the answer
77 combination screened by the user’s intuition.

78 **Time constraints are not fully considered in multi-answer temporal problems.** The WikiData
79 data excerpt for the question in Figure 1 shows that Andrew Jackson was president of the United States
80 for a period of time [18290304-18370304], and three secretaries of state met this time constraint. Other
81 candidates for secretary of state should be eliminated because they do not meet the time constraint. Most
82 KGQA models however ignore the important role of timing constraints when dealing with multi-answer
83 questions, leading to incorrect results. The key to answering such multi-answer temporal questions is
84 to determine the candidates that satisfy the time constraint interval of the answer. A time fact can be
85 considered as a correct answer only if it conforms to the temporal logic of the problem, that is, the
86 temporal constraint represented by a given explicit or implicit fact needs to be satisfied.

87 This paper therefore proposes a Multi-Answers Temporal Question Answering (MATQA) component
88 for multi-answer reasoning, which can be combined with any KGQA system to improve the answering
89 effect. The time constraint on the correct fact in the knowledge graph candidates makes it possible
90 to output all the standard answers. To address the above problems, MATQA proposes the following
91 solutions. First, inspired by the multi-paragraph open-domain question answering, after introducing the
92 multi-answer question into the field of knowledge graph temporal question answering, the reverification
93 framework is used to improve the existing Top-k answer display form, and the question answering process
94 with a certain number of answers is constructed. Second, the correct initial answers among the candidate
95 answers are filtered by embedding the question and answer pairs into the graph as boot nodes. Finally,
96 considering the same type of relationship and time constraint between answers to a temporal question,
97 semantic constraints and time constraints are constructed in the answer clustering verification process,
98 and candidate answers that do not meet the time constraints but have similar types and high scores that are
99 eliminated. At the same time, the incorrect answers with high scores can be filtered again at the semantic
100 level to ensure the accuracy. Experiments using a recent temporal question answering benchmark and a

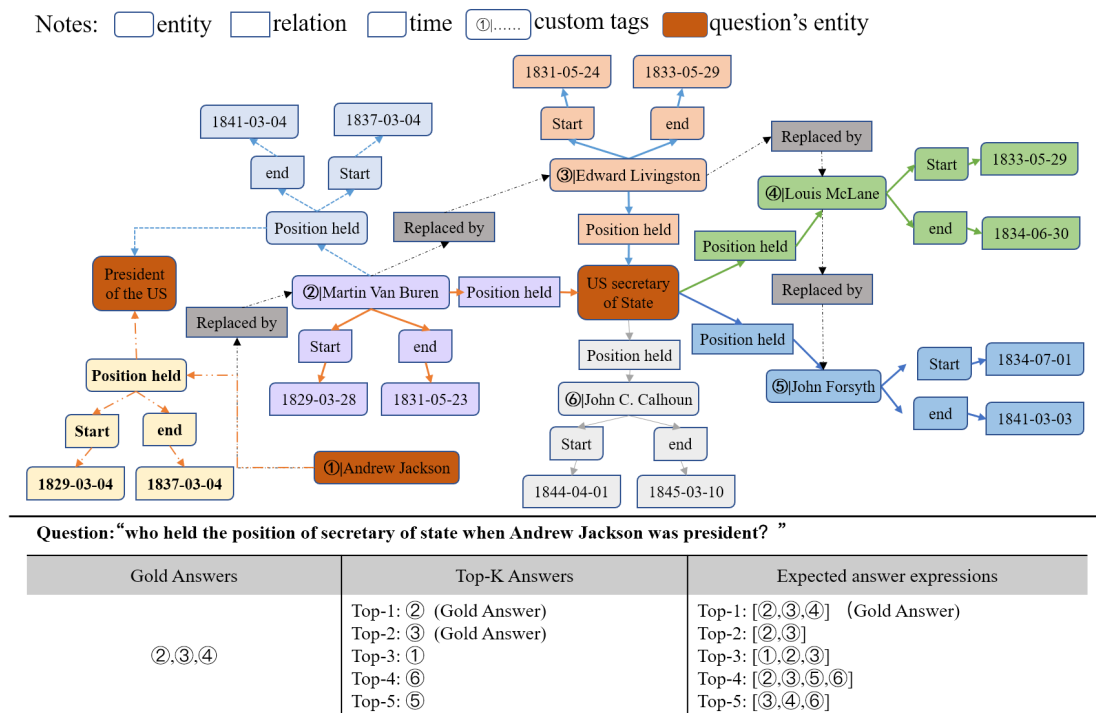


Figure 1. A expressions of answers to the question and an excerpt from the Wikimap of the question

101 set of competitors based on unstructured text sources show the advantages of MATQA: The model can
 102 give the number of correct answers based on the knowledge graph, and can use the time information of
 103 the temporal question to filter the answers. Given a new answer expression, it can better guarantee the
 104 quantity and quality of the answers.

105 In summary, the key contributions are 3-fold:

- 106 • Multi-answer reasoning is introduced into temporal knowledge graph question answering to improve
 107 Top-k, and a new answer expression is proposed, which gives the user the exact number of answers.
- 108 • Based on the reverification framework, a component that contains time information is designed to
 109 guarantee the quantity and quality of answers.
- 110 • A series of experiments show that MATQA can not only infer the number of answers to temporal
 111 questions, but also take into account the accuracy of knowledge question answering.

112 RELATED WORK

113 **Top-k algorithm.** The traditional Top-k method aims to return the top k answers that are closest to the
 114 expected value. The main idea is to filter a series of candidate matches constructed according to the
 115 similarity criterion so as to obtain the answer that matches the target value. Each step of KGQA, such as
 116 named entity recognition, entity disambiguation, and entity linking, results in a ranked Top-k list. The
 117 whole question answering process is the Top-k retrieval of multi-link ranking mechanism fusion. The
 118 main methods are Fagin algorithm and threshold algorithm, and the core task is to sort the candidates of
 119 multiple dimensions, and then calculate according to a specific pruning strategy(Auer et al., 2008). For
 120 example, (Christmann et al., 2021) fused the quantitative scores such as semantic coherence of candidate
 121 items, connectivity of knowledge graph, relevance to the question, etc., to reduce the candidate domain in
 122 knowledge question answering, and then used the threshold algorithm to filter the score list of multiple
 123 indicators to obtain the most relevant candidate neighborhood to the question. (Wang et al., 2021) filtered
 124 the semantically weighted scores of edges using upper and lower bound filtering and defined a star Top-k
 125 query scheme with early termination of matching. Top-k query is related to the quality of answers.
 126 However, the traditional Top-k query is presented in the form of a single answer list, which cannot reflect
 127 the standard answers of multi-answer questions, including the number and answer of answers. MATQA

128 extends the single-answer display form to a multi-answer one, which can better ensure the quality in
129 multi-answer question answering.

130 **Multi-answer Question Retrieval based on Unstructured Text Sources.** Unstructured text sources
131 often organize knowledge in the form of articles or paragraphs and are crucial in the field of question
132 answering. Open-domain question answering based on multi-paragraph multi-answer reasoning challenges
133 the ability to comprehensively utilize evidence from large-scale corpora. Due to the ambiguity and
134 openness of questions, a question often has multiple correct answers. Predicting the answer contained
135 in each paragraph in turn after retrieving the reordered paragraphs has become the mainstream question
136 answering paradigm in this field. AMBIGNQ(Min et al., 2020) uses Bert model to sort paragraphs and
137 generate answers in turn. (Shao and Huang, 2022) proposed the “recall and reverification” framework to
138 separate the reasoning process of each answer and used the new evidence obtained from recall to verify
139 the answer. Although unstructured multi-answer question answering has received extensive attention, the
140 multi-answer question answering based on structured data cannot meet the needs of obtaining all correct
141 answers to the question. Therefore, it is of great practical significance to extend multi-answer question to
142 knowledge graph question answering.

143 **Multi-answer Reasoning based on Temporal Knowledge Questions.** Good progress has been
144 made in the question answering of temporal questions. A series of advanced schemes(Jia et al., 2021;
145 Saxena et al., 2021; Mavromatis et al., 2022; Jiao et al., 2022; Chen et al., 2021) have proved that the
146 processing of time information in the question is helpful to guarantee the quality of complex knowledge
147 question answering. The time information contained in the question limits the time interval of the answer.
148 When the semantic constraints are satisfied, the number and accuracy of the answers to the multi-answer
149 question are measured by the time interval. The facts beyond the time interval do not satisfy the user
150 intention and should be excluded from the answer output. As a special branch of temporal questions, the
151 multi-answer question faces great challenges. The single answer list and false positive answers make it
152 difficult for users to determine the number and accuracy of answers to a question. This paper therefore
153 aims to expand the answer expression form of multi-answer temporal question, and investigate the factors
154 that ensure the quality of temporal question answering based on the complete question answering process.

155 RESEARCH METHOD

156 **Task description:**The objective of this paper is to answer multi-answer temporal questions with question
157 answering pair information and structured knowledge. Given a problem q_i and m sets of candidate answers
158 a_m^i , MATQA needs to obtain from the candidate answer set a_m^i the number of valid answers to question q_i
159 and correct entities or attributes.

160 **Approach Introduction:** Figure 2 presents the overall structure of MATQA. It uses four modules to
161 perform the process of answering multi-answer temporal questions, corresponding to the construction of
162 the boot node, the scoring of the KG node related to the question, the judgement of the initial answer,
163 and the clustering of the same type of answers under time constraints. First, the Q&A pair is associated
164 with the knowledge graph as a special node, which can bridge the information gap between Q&A pair
165 and subgraph in the subsequent reasoning process, and guide the model to approach the standard Q&A.
166 Second, the degree of correlation between the key entities in the resolved triplet facts in the question and
167 the particular node in the Q&A pair is measured, and only the KG nodes associated with the question
168 are retained. Subsequently, the information of Q&A pairs and subgraphs is aggregated and updated
169 on the graph by the graph neural network of attention mechanism, and the possible solution with the
170 highest score is deduced. Finally, the time constraints after problem analysis are used to cluster the other
171 candidate answers in the Q&A pairs, and all the answers satisfying both the semantic and time constraints
172 are selected as the solution set of the problem.

173 Boot node representation

174 In order to use the answer information to guide the problem reasoning, the question q_i and the candidate
175 answer set a_m^i provided by other question answering schemes are together inserted into the knowledge
176 graph as a special node, known as boot node (*Boot*), denoted as $[q_i; a_m^i]$, as shown in Figure 3. Herein,
177 a_m^i can be a traditional form of *Top-k* solution to question q_i given by any question answering scheme,
178 and the standard answer in the candidate solution set a_m^i is clearly marked. In the special nodes formed
179 by Q&A pairs, the question is taken as the starting point of the reasoning model, and the answer as the
180 end point, implicitly expressing the information of the question and answer context. The boot node is

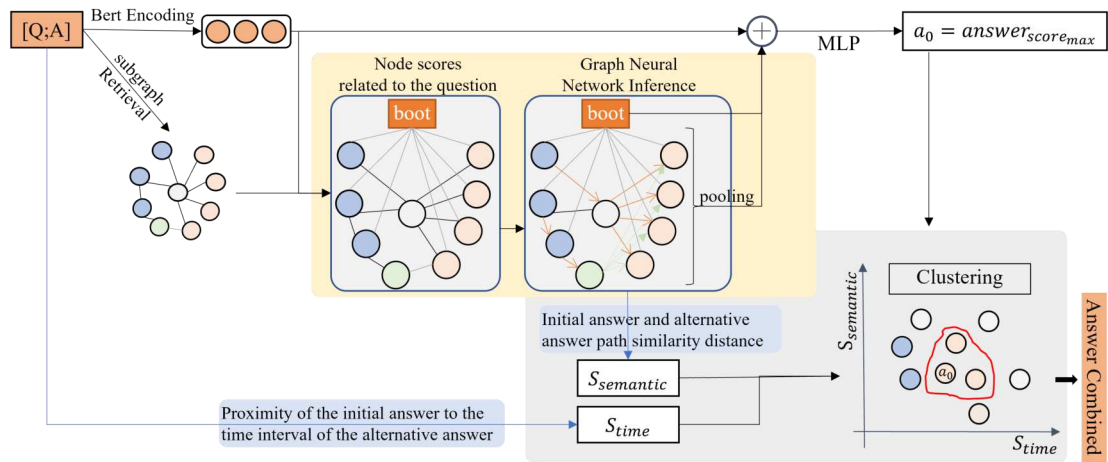


Figure 2. The structure of MATQA. The component can be attached to the question answering system. Based on the **revalidation** framework, it uses the boot node (another form of Q&A pair) representation, as well as the KG node score related to the question, to determine the initial answer, and finally obtains the answers through the time and semantic dimension of the alternative answer clustering.

181 associated with entities contained in the question, and the mapping item of the boot node and the marked
 182 standard answer node in the knowledge graph is linked, and the new relation “gold answer” is given,
 183 which is shown by the orange dotted line in Figure 3. Therefore, a new answer-guided knowledge graph
 184 is constructed between the boot node and the knowledge graph, and between the answer node and the
 185 corresponding boot node, known as inference graph G_R herein.

The Boot node is regarded as a long sequence text and encoded by Bert, where f_e is the encoding function.

$$Boot^{bert} = f_e(\text{text}(Boot)) \quad (1)$$

186 After the Boot node is given, the subgraph $G_{sub}^{boot} = (v_{sub}^{boot}, e_{sub}^{boot})$ after entity link is extracted from
 187 knowledge graph $G = (V, E)$, where V is the entity node of the knowledge graph, E is the relationship
 188 between two entities, v_{sub}^{boot} is the entity nodes in all boot nodes extracted from the graph, e_{sub}^{boot} is the
 189 relationship nodes in all the boot nodes extracted from the graph, and G_{sub}^{boot} is the subgraph associated
 190 with the boot node extracted from the knowledge graph.

191 Scoring of KG nodes associated with the question

There are many paths unrelated to the question in the subgraph after entity link disambiguation. As shown in Figure 1, Martin Van Buren’s path as president is unrelated to his path as Secretary of State. These unrelated paths cause the model to waste a lot of time in the inference process to exclude invalid paths. To address this problem, this paper uses the question correlation fact determination module to calculate the similarity score between the boot node and KG fact node.

$$S_{sub}^{boot} = f_h(f_e[\text{text}(boot); \text{text}(v_{sub}^{boot})]) \quad (2)$$

192 Where $f_h \circ f_e$ is the probability that the boot node is connected to the subgraph node; S_{sub}^{boot} is the score of
 193 correlation between the boot node and the subgraph node, which describes the importance of each node to
 194 the boot node, and is used to prune the inference graph G_R .

195 Initial result determination

196 The answer with the highest score in the question answering system has the greatest probability of being
 197 the standard answer. This paper therefore finds out the most likely answer to the multi-answer question
 198 through subgraph reasoning, and regards it as the correct answer. MATQA’s reasoning process is based
 199 on the graph attention GAT framework.

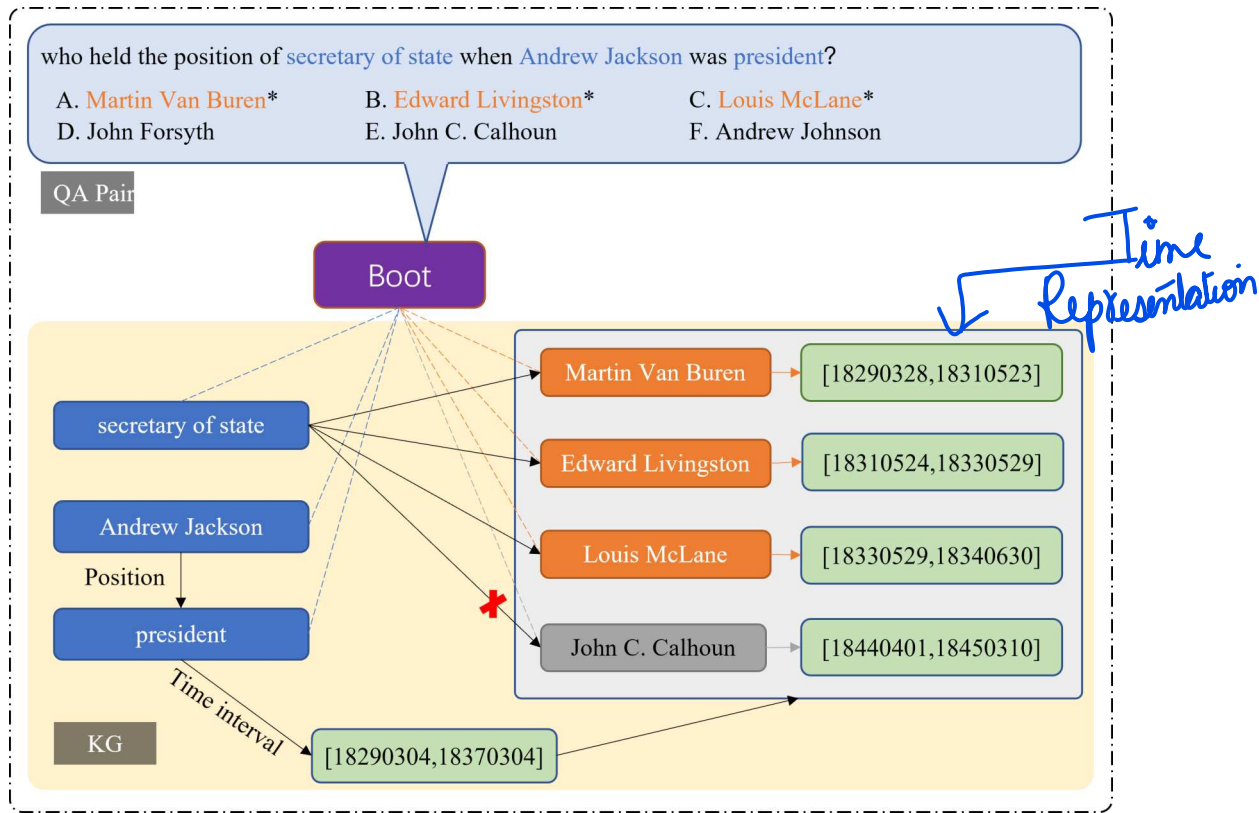


Figure 3. Diagram of “inference graph”

In a l layer graph network model, for a node $v \in V_{sub}$ in any subgraph, vector initialization is performed by Bert encoding, i.e., $h_v^0 = f_e(\text{text}(v))$. Then the updating model can be expressed as:

$$h_v^{l+1} = \left(\sum_{n \in N_v \cup \{v\}} \delta_{nv} m_{nv} \right) + h_v^l \quad (3)$$

Where N_v is the neighbor node of node v , m_{nv} is the message from each neighbor node n to node v , and δ_{nv} is the weight of the message from node n to node v . The calculation of message m_{nv} should take into account the characteristic h_n^l , type u_n , and time attribute t_n of the node, as well as the embedded relation r_{nv} . The calculation formula is as follows:

$$m_{nv} = \text{linear}(h_n^l, u_n, t_n, r_{nv}) \quad (4)$$

Where u_n is the type's one-hot code embedded of the neighbor node n of node v , t_n is the embedded time attribute of neighbor node n , and r_{nv} is the embedded relation between nodes n to v .

To calculate the attention weight vector of nodes n to v , two key query vectors are constructed according to node types:

$$\begin{cases} q_n = \text{Linear}(h_n^l, u_n, S_{n \in sub}^{boot}) \\ k_v = \text{Linear}(h_v^l, u_v, S_{v \in sub}^{boot}, r_{nv}) \end{cases} \quad (5)$$

The final attention weight vector can be obtained by formula (6) below.

$$\delta_{nv} = \frac{\exp(\gamma_{nv})}{\sum_{n' \in N_v \cup \{v\}} \exp(\gamma_{n'v})}, \quad \gamma_{nv} = \frac{q_n^T k_v}{\sqrt{D}} \quad (6)$$

Then the reasoning process of the initial answer $p(a_0^i | q_i)$ is given by:

$$p(a_0^i | q_i) = \exp(\text{MLP}(\text{Boot}^{bert}, h_{boot}^l, G_{sub}^{pooling})) \quad (7)$$

Where Boot^{bert} is the vector representation of boot node, h_{boot}^l is the updating representation of the boot node at l layer, and $G_{sub}^{pooling}$ is the pooling representation of subgraph.

204 Answer clustering of the same type under time constraints

After the initial answer a_0 is obtained, the rest of the answers to the question should be deduced. Since all answers to the question should meet the same constraints, including the semantic and time constraints, MATQA processes the other answers through clustering. In order to correctly measure the gap between the alternative answer and the initial answer a_0 , the subgraph path $(V_{sub}, E_{sub}, a_{other})$ of the alternative answer is extracted to calculate the semantic similarity score between it and the path (V_{sub}, E_{sub}, a_0) of the initial answer.

$$S_{semantic} = \cos[(V_{sub}, E_{sub}, a_{other}), (V_{sub}, E_{sub}, a_0)] \quad (8)$$

The final answer to each question is constrained by the time interval. Therefore, the matching between the time interval of the fact and the real time interval of the question can exclude the answer that does not satisfy the condition. KG retrieval and TimeML (Pustejovsky et al., 2003) are used to calculate the time constraint interval of the question, which is $[T_s, T_e]$ (T_s and T_e are the start time and end time, respectively). At the same time, the time interval $[T_s^{other}, T_e^{other}]$ of the fact corresponding to the alternative answer is extracted. The final predicted score of time similarity S_{time} can be obtained by:

$$S_{time} = \text{Relu} \left\{ \begin{array}{l} 1, T_s < T_s^{other} \text{ and } T_e^{other} < T_e \\ -1, T_s^{other} < T_s \text{ or } T_e^{other} > T_e \end{array} \right\} \quad (9)$$

Why 2?

205 The number of clusters is set to 2, and the K-means algorithm is used for clustering. The answers that are
 206 close to the initial answer consists of the final answer combination. And the rest answers are excluded
 207 and then activated by Relu function so as to make zero the scores of answers that do not conform to the
 208 time constraints. Answers that do not have a final time score of 0 are the answers that satisfy the time
 209 constraint.

210 The answers that satisfy the semantic and time constraints after clustering are regarded as the true
 211 predicted answers a_m^i to the question q_i . Each row of Top-k is a combination of answers, as shown in the
 212 expected answer expressions in Figure 1.

213 EXPERIMENT

214 Datasets

215 TimeQuestions [12] is a wikidata-based question-answering data set consisting of 16,181 Q&A pairs,
 216 among which 9708 questions are used for training, 3236 for verification and 3237 for testing. The type of
 217 each question (explicit, implicit, time, and order) is indicated in the Q&A pairs. At the same time, the
 218 signal words for time interaction in the question are specified, such as before/after, start/end, etc. In order
 219 to process the multi-answer questions, all question pairs with more than one answer are extracted from
 220 the TimeQuestions data set to construct the multi-answer TimeQuestions data set. The new multi-answer
 221 question dataset contains 2264 training sets, 778 verification sets and 801 test sets, and the labels of the
 222 question type and time signal.

223 Evaluation metrics

224 Two measures are used to evaluate the quality of answers to the multi-answer question.

- 225 • $P@1^m$ (the precision of multi-answers): For a new answer form given in a question, the highest-
 226 ranked combination of answers has a precision of 1 when the combination is exactly the same as
 227 the standard answers (both in the quantity and the label), which is denoted as $P@1_{hard}^m$. When the
 228 highest-ranked answer combination contains all the standard answers, that is, the first result of
 229 the prediction includes other results besides the standard answers, it is denoted as $P@1_{soft}^m$ with
 230 broader constraints.
- 231 • $Hits@5^m$ (the hits of multi-answers): The combination of answers depends on the number and
 232 label of answers. The label needs to satisfy the semantic matching relation of the question, and the
 233 number is all possible solutions that satisfy the semantic constraints. Because of the complexity
 234 of language questions, semantic constraints cannot be fully satisfied, and there are many possible
 235 combinations of answers. Under the new answer expression form, the first five groups of answers are
 236 ranked in descending order of the proportion of the standard answers on the list. If a list containing
 237 any subset of the standard answer appears in the first five positions, it is set to 1, otherwise to 0.

Table 1. Comparison of results of MATQA

Model	$P@1^m_{hard}$	$P@1^m_{soft}$	Hits@5 ^m
TransE+MATQA	0.402	0.439	0.513
EXAQT+MATQA	0.431	0.453	0.546
TERQA+MATQA	0.459	0.472	0.538

← more justification

238 Baselines

239 MATQA proposed in this paper is a component attached to the traditional QA model and can be used with
 240 any model. In addition, this paper introduces the multi-answer question into knowledge temporal question
 241 answering for the first time. In order to measure the model effect, a relatively simple link prediction
 242 scheme and two advanced temporal question answering schemes are selected as the providers of candidate
 243 answers, which are then combined with MATQA to judge whether the multi-answer prediction model is
 244 effective.

- 245 • TransE: it is the most classical vector embedding method ^{which} completes the missing answers according
 246 to the translational semantic invariance law.
- 247 • EXAQT[12]: it is an end-to-end temporal question answering scheme, which for the first time
 248 builds the temporal question answering system on wikidata, a large-scale open-domain knowledge
 249 graph. It does not require the process of constructing a temporal knowledge graph. The final answer
 250 prediction and accuracy is performed using R-GCN by augmenting the embedding of subgraphs
 251 and questions, performing temporal augmentation of subgraphs, or reconstructing subgraphs to
 252 augment recall in three ways.
- 253 • TERQA(Yao et al., 2022): On the basis of EXAQT, inspired by capsule network, improved the
 254 fusion of time features and triplet features and learned the exact dependence between time features
 255 and triplet facts, which enhanced the accuracy of the model to predict the answer.

256 Experimental settings

257 MATQA uses PyTorch for implementation, and sets the vector embedding dimension after Bert initial-
 258 ization to 200. It has five layers of GNN, each of which with a dropout of 0.2. Moreover, it uses Adam
 259 for initial answer inference optimization and Relu for time constraint score optimization. Furthermore,
 260 batch_size is set to 32, learning rate to 2e-3, and cluster number to 2.

261 RESULTS

262 Key findings

263 Table 1 shows the effects of multi-answer judgment on the multi-answer question data set. The index
 264 $P@1^m_{hard}$ demonstrates that MATQA can improve the traditional Top-k expression form to make each line
 265 a new form of a list of answers, which is consistent with the expected human expression form in Figure 1.
 266 Therefore, MATQA can better meet user's requirements on the number and accuracy of questions with
 267 multiple answers. At the same time, MATQA has proved that its effectiveness is largely related to the
 268 alternative answers provided. That is, the more accurate the candidate answers, the more accurate the
 269 initial answer, and the better the final result after clustering.

270 Through the revalidation framework of "initial answer → clustering", MATQA can provide a solution
 271 to the multi-answer temporal reasoning question. The primary shortcoming of MATQA is that its final
 272 output is largely affected by the initial result. In other words, in the case of an incorrect initial answer, the
 273 subsequent clustering module cannot correct ~~it~~ it and can only make invalid prediction ^S on a wrong basis.
 274

275 Disambiguation experiment

276 Table 2 shows the results of MATQA after removing each module. It can be seen that the introduction of
 277 the boot node enables the question and the candidate answers to inspire the inference model. In addition,
 278 the boot node and the interference graph it consists have positive feedback to P@1. In the case of no boot

justification? →

↓ activation?

output

Table 2. Results of TransE + MATQA after removal of module

Model		$P@1_{hard}^m$	$P@1_{soft}^m$
No boot nodes		0.382	0.391
GNN	No node types	0.398	0.401
	No score of nodes related to question	0.386	0.394
	No pooling layer	0.382	0.389
Clustering	No semantic constraints	0.254	0.287
	No time constraints	0.305	0.348

Table 3. Top-1 results of improved questions

Question	Gold answers	Predicted answers
In which year, did the Steelers win the super bowl, the latest occasion?	Super Bowl 'IX', 'X', 'XIII', 'XIV', 'XL', 'XLIII'	Super Bowl 'IX', 'X', 'XIII', 'XIV', 'XL', 'XLIII'
Who ran against Lincoln in the 1864 presidential election?	"John C. Breckinridge" and "Stephen A. Douglas"	"John C. Breckinridge" and "Stephen A. Douglas"
When did owner Fred Wilson's sports team win the pennant?	"1969 World Series" and "1986 World Series"	"1969 World Series" and "1986 World Series"

279 nodes, the QA model cannot get the information guidance of hidden answer, and the Q&A context cannot
 280 be updated with KG, which cannot bridge the information gap between question and knowledge graph
 281 and thus damages the system performance $P@1_{hard}^m:40.2\% \rightarrow 38.2\%$, $P@1_{soft}^m:43.9\% \rightarrow 39.1\%$).

282 When semantic constraints are removed during clustering, the model effect declines most seriously,
 283 because the clustering of answers mainly measures the degree of fact similarity. Additionally, among
 284 temporal questions, a large proportion have answers within a specific time constraint interval. When time
 285 constraint is removed, the entities of the answers cannot be measured by time constraint, which will easily
 286 lead to incorrect answers. Finally, the addition of the boot node makes up the information gap between
 287 the question context and the knowledge graph, and has a great influence on the determination of the initial
 288 answer. Removing modules from GNN also has an effect on the prediction of the final initial answer

289 Typical questions

290 The effectiveness of MATQA is fully demonstrated by three typical questions. In Table 3, the question
 291 "in which year, did the Steelers win the super bowl, the latest occasion?" has the standard answers of
 292 "Super Bowl 'IX', 'X', 'XIII', 'XIV', 'XL', 'XLIII'". The model has accurately predicted the number of
 293 answers and the correct answer. It is proved that MATQA framework has a good effect on the processing
 294 of multi-answer temporal questions, and makes up the defects of traditional top-k which cannot show the
 295 number of answers and has false positive results.

Table 4. Incorrect results obtained by MATQA

Question	Gold Answers	Predicted Answers
What is inflation rate of Dominica that is point in time is 1983-1-1?	"2.7"	"ACM Software System Award" and "Turing Award"
When did Anne Hathaway begin attending New York University and when did she graduate?	"1995" and "1998"	History of art

What is the difference

296 Error types

297 As shown in Table 4, MATQA is likely to predict multiple unrelated entities as the answers when the
298 question expects a numerical answer. This shows that MATQA cannot determine the number of answers
299 to some single-answer questions through semantic and time constraints, which leaves room for further
300 improvement. In addition, the question expects to output multiple valid times as the answers, but MATQA
301 predicts a single entity as the answer. This indicates that when there are errors in the initial answer
302 predicted by MATQA, there will be errors as well in subsequent clustering. Therefore, the framework
303 needs to be further updated in future studies to mitigate its impact.

304 DISCUSSION

305 As a special type of questions, multi-answer questions occupy an important position in the field of
306 intelligent Q&A. At present, multi-answer questions are widely used in the field of multi-paragraph
307 unstructured text sources, but have not been paid attention by researchers on structured knowledge graphs.
308 This study introduces multi-answer questions into the temporal knowledge Q&A scenario, aiming to
309 update the shortcomings of the traditional Top-k answer representation form.

310 In this study, MATQA defines the true number of answers and eliminates false positives through a
311 "revalidation" framework. The combined use of initial answer establishment and semantic time based dual
312 factor clustering ideas was shown to have a positive effect on the number of answers and correctness of
313 questions. Previous research [2] has shown that the revalidation framework is able to take full advantage
314 of the information collected to further filter the answers. This is consistent with the study in this paper.
315 Further, the "revalidation" framework was shown to be able to determine not only the correctness of
316 answers but also the number of answers, with only the addition of semantic and temporal constraints
317 on clustering. Based on this, this paper shows that the "revalidation" framework in the form of "initial
318 answer → clustering" can provide a solution to the multiple answer reasoning problem in the context of
319 temporal knowledge quiz. However, MATQA suffers from severe upstream error-dependent transmission.
320 When the initial answer is wrong, the subsequent clustering module cannot correct the result, but only
321 makes invalid predictions based on the original one.

322 Despite its drawbacks, this study provides a solution to multi-answer questions in a structured temporal
323 knowledge Q&A scenario and points out that the key to multi-answer questions lies in the number of
324 answers and false positive result filtering. Meanwhile, the introduction of bootstrap nodes enables
325 questions and candidate answers to shed light on the inference model, and subsequent updates jointly
326 utilize bootstrap nodes and subgraph domains to bridge the information gap between questions and
327 knowledge graphs. Based on the existing research, the establishment of initial answers and the refinement
328 of clustering factors will be the next step of research to be considered.

329 CONCLUSION

330 Although temporal question answering is crucial to knowledge workers, its multi-answer reasoning has
331 not received much attention. The traditional Top-k answer expression cannot meet user's demand for
332 the quantity and quality of answers. To address this problem, this paper proposes the MATQA model
333 based on the initial answer and semantic and time clustering. The model is able to define the true number
334 of answers and eliminate false positives by the "revalidation" framework. The number and accuracy of
335 answers can be improved by combining initial answers with clustering of semantic and temporal factors.
336 Experimental results on a large number of complex multi-answer temporal questions show that MATQA
337 can improve the most advanced general Top-k question answering scheme. In future research, the model
338 will be combined with the learned knowledge to ask questions, so as to gradually guide users to clarify
339 their intentions and output correct and realistic standard answers.

340 ACKNOWLEDGMENTS

341 We thank Zhihong Shao, Zhen Jia, and Michihiro Yasunaga for their achievements that inspired this work,
342 which was done with the support of the School of Computer Science, Xi'an Institute of High Technology,
343 and we would like to thank the anonymous reviewers for their helpful remarks.

Can
combine
to remove
redundant
material

344 **REFERENCES**

- 345 Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., Ilyas, I. F., Beskales, G., and
346 Soliman, M. A. (2008). A survey of top- k query processing techniques in relational database systems.
347 *ACM Computing Surveys*, 40(4):1–58.
- 348 Cao, Q., Liang, X., Li, B., and Lin, L. (2021). Interpretable Visual Question Answering by Reasoning on
349 Dependency Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):887–901.
- 350 Chen, X., Jia, S., Ding, L., and Xiang, Y. (2021). Reasoning over temporal knowledge graph with temporal
351 consistency constraints. *Journal of Intelligent & Fuzzy Systems*, 40(6):11941–11950.
- 352 Christmann, P., Roy, R. S., and Weikum, G. (2021). Beyond NED: Fast and Effective Search Space
353 Reduction for Complex Question Answering over Knowledge Bases. *Applied Network Science*, 6(1).
- 354 Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. (2018). TempQuestions: A Benchmark
355 for Temporal Question Answering. In *Companion of the The Web Conference 2018 on The Web
356 Conference 2018 - WWW '18*, pages 1057–1062, Lyon, France. ACM Press.
- 357 Jia, Z., Pramanik, S., Saha Roy, R., and Weikum, G. (2021). Complex Temporal Question Answering
358 on Knowledge Graphs. In *Proceedings of the 30th ACM International Conference on Information &
359 Knowledge Management, CIKM '21*, pages 792–802, New York, NY, USA. Association for Computing
360 Machinery.
- 361 Jiao, S., Zhu, Z., Wu, W., Zuo, Z., Qi, J., Wang, W., Zhang, G., and Liu, P. (2022). An improving reasoning
362 network for complex question answering over temporal knowledge graphs. *Applied Intelligence*.
- 363 Liu, Q., Geng, X., Huang, H., Qin, T., Lu, J., and Jiang, D. (2021). MGRC: An End-to-End Multigranular-
364 ity Reading Comprehension Model for Question Answering. *IEEE Transactions on Neural Networks
365 and Learning Systems*, 1(3):1–12.
- 366 Mavromatis, C., Subramanyam, P. L., Ioannidis, V. N., Adeshina, A., Howard, P. R., Grinberg, T., Hakim,
367 N., and Karypis, G. (2022). Tempoqr: temporal question reasoning over knowledge graphs. In
368 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5825–5833.
- 369 Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. (2020). AmbigQA: Answering Ambiguous
370 Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural
371 Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- 372 Moon, S., He, H., Liu, H., and Fan, J. W. (2022). Rxwhyqa: a clinical question-answering dataset with
373 the challenge of multi-answer questions. *arXiv preprint arXiv:2201.02517*.
- 374 Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev,
375 D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. *New directions
376 in question answering*, 3:28–34.
- 377 Rubin, S. J. A. O., Yoran, O., Wolfson, T., Herzig, J., and Berant, J. (2022). Qampari:: An open-domain
378 question answering benchmark for questions with many answers from multiple paragraphs. *arXiv
379 preprint arXiv:2205.12665*.
- 380 Saxena, A., Chakrabarti, S., and Talukdar, P. (2021). Question Answering Over Temporal Knowledge
381 Graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
382 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,
383 volume 1, pages 6663–6676, Online. Association for Computational Linguistics.
- 384 Shao, Z. and Huang, M. (2022). Answering open-domain multi-answer questions via a recall-then-verify
385 framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
386 (Volume 1: Long Papers)*, pages 1825–1838.
- 387 Wang, Y., Xu, X., Hong, Q., Jin, J., and Wu, T. (2021). Top- k star queries on knowledge graphs through
388 semantic-aware bounding match scores. *Knowledge-Based Systems*, 213:106655.
- 389 Yao, J., Wang, Y., Li, X., Yuan, C., and Cheng, K. (2022). TERQA: Question answering over knowledge
390 graph considering precise dependencies of temporal information on vectors. *Displays*, 74:102269.
- 391 Zhong, V., Shi, W., Yih, W.-t., and Zettlemoyer, L. (2022). Romqa: A benchmark for robust, multi-
392 evidence, multi-answer question answering. *arXiv preprint arXiv:2210.14353*.