Journal of King Saud University – Computer and Information Sciences 34 (2022) 9644–9653

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A survey on textual entailment based question answering

Aarthi Paramasivam *, S. Jaya Nirmala

Department of Computer Science and Engineering, National Institute of Technology Tiruchirappalli, India

ARTICLE INFO

Article history:
Received 19 June 2021
Revised 27 October 2021
Accepted 21 November 2021
Available online 27 December 2021

Keywords:
Natural Language Processing
Question Answering
Textual Entailment

ABSTRACT

Question answering, an information retrieval system that seeks knowledge, is one of the classic applications in Natural Language Processing. A question answering system comprises numerous sets of subtasks. Some of the subtasks are Passage Retrieval, Answer Ranking, Question Similarity, Question Generation, Question Classification, Answer Selection, and Answer Validation. Numerous approaches have been experimented on in the question answering system to achieve accurate results. One such approach for the question answering system is Textual Entailment. Textual Entailment is a framework that captures significant semantic inference. Textual Entailment of two text fragments can be defined as the task of deciding whether the meaning of one text fragment can be inferred from another text fragment. This survey discusses how and why Textual Entailment is applied to various subtasks in question answering.

Contents

1. Introduction

Artificial Intelligence (AI) has become an essential part of everyone's life in the current digital era. The virtual digital assistants such as Alexa and Siri has changed the way we do our daily tasks. The virtual assistant handles a considerable amount of work to complete the user's task in their day-to-day life. Question Answering Systems (QAS) and other machine learning techniques

* Corresponding author at: National Institute of Technology, Tiruchirappalli - 620015, Tamil Nadu, India.
E-mail address: paramasivamaarthi@gmail.com (A. Paramasivam).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.11.017
1319-1578/

are used by the virtual assistant. Question Answering (QA) is a type of conversational AI that focuses on generating natural language answers to the questions posed by the human user. The growth of technology has resulted in the massive volume of information we have access to the QA System. Hence, the ability to answer questions of any subject becomes a challenging task. QA is a branch of AI within the Natural Language Processing (NLP) and Information retrieval (IR) fields. NLP helps a computer understand the human language. QA systems automatically answers the question asked by the human in natural language. The two major paradigms in the QA system are IR-based QA and Knowledge-based QA. An IR-based QA system can be broken into three stages. The stages are question processing, passage retrieval and ranking, and answer extraction. Knowledge-based QA uses well-structured information stored in the database to answer the natural language questions. QA systems are either open domain or closed domain. IBM's Watson is an example of an Open-domain QA system.

QA System can also be used to derive information from the image to answer the user's question. Kodra and Mece (2017) have presented us with a review of the QA system, which gives us an overview of the QA system and discusses the development, challenges, and trends of the QA system. Fig. 1 illustrates the four different perspectives of the QA system studied by them. The four perspectives are system characteristics, research topics, research challenges, and solution approaches.

Though many surveys have been done on QA and Textual Entailment (TE) separately, it is the first survey to the best of our knowledge to analyze the role of TE in QA. Also, discovering all related research approaches for the TE based QA is a complex task. The survey briefs about the following

- Dataset available for the TE task.
- The role of TE in various subtasks of QA.
- Limitations faced while using TE in QA.
- Results achieved by using TE.
- Potential Future Research.

To better understand the information retrieved from any source, TE framework is used in QA system. Section 2 discusses about TE. Section 3 studies the related work done in TE. Section 4 gives an overview of the datasets available for TE. Section 5 outlines how TE is adopted in QA at various stages by different researchers to improve performance. Section 6 discusses the summary and the future research directions TE based QA. Section 7 concludes the survey.

2. Textual Entailment

Natural Language Inference (NLI), also known as Recognizing Textual entailment (RTE), studies whether one text fragment can be inferred from another text fragment. Given a pair of text fragments known as Text (T) and Hypothesis (H), TE recognition is the task of deciding whether the hypothesis can be inferred from the text. RTE is an NLP task used in various NLP applications such as QA, IR, text summarization, machine translation, information extraction, and paraphrasing. For instance, in a QAS, the answer obtained for one question after the IR process should be entailed by the text's supporting snippet. The classic entailment definition by Chierchia and McConnell-Ginet (2000) is "A T entails H if H is true in every circumstance in which T is true". For example,

T: Mary killed the spider.

H: Spider is Dead.

Here H can be inferred from T, so T entails H. The probabilistic interpretation of the textual entailment (Glickman et al., 2005) can be given as

$$P(\text{H is true}|T) > P(\text{H is true})$$

i.e., T increases the likelihood of H being true.

The architecture of the RTE can be viewed as the classification problem, as shown in Fig. 2. The three possible outputs from the entailment recognition are Entailment, contradiction, and unknown. If T entails H, then the pair is marked as *Entailment*. Else when T does not entail H, then the pair is marked as a *Contradiction*. The final possible output is *Unknown*, where H's truth could not be determined based on T.

3. Related work

RTE issues were identified and addressed through a series of RTE challenges. The first PASCAL RTE challenge (Dagan et al., 2005) was an initial attempt to create a generic empirical task that captures the major semantic inference. It also provided the first benchmark for the entailment task. The first PASCAL RTE challenge dataset contains over 1000 English Text-Hypothesis (T-H) pairs from the news domain. The goal of the challenge is to decide whether T entails H or not for each T-H pair. The main task of the second RTE challenge Haim et al. (2006) is to recognize whether a Hypothesis(H) is entailed by the Text(T) but with the more realistic T-H example. The data set for the second RTE challenge is mainly based on the outputs of the actual system. The third RTE challenge (Giampiccolo et al., 2007) follows the same

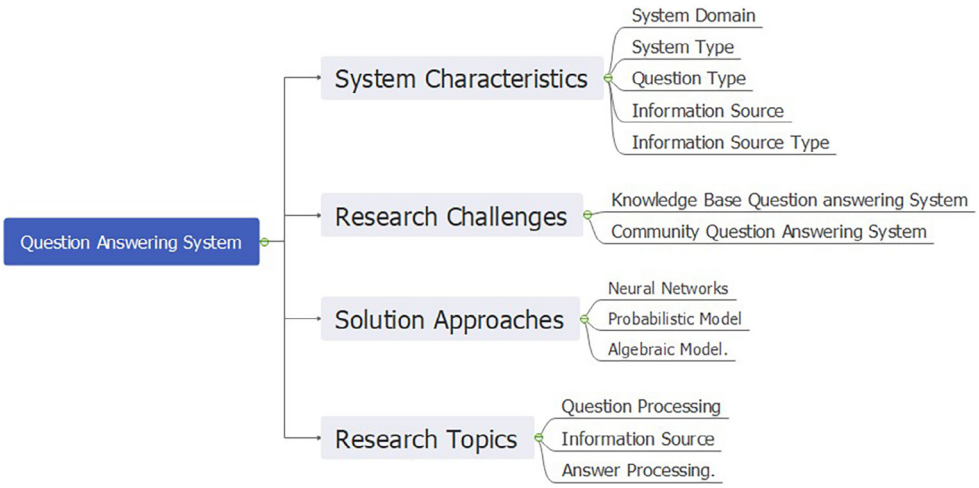

Fig. 1. Review of QAS by Kodra and Mece (2017).

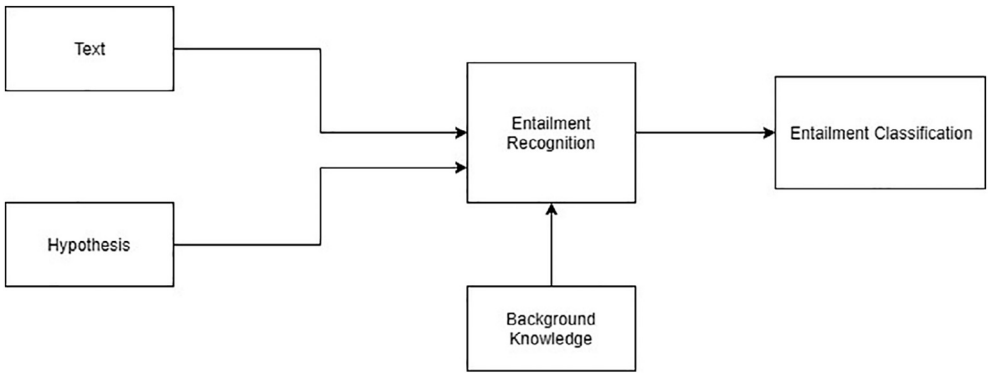

Fig. 2. Textual Entailment System.

basic structure of the previous challenges but with relatively longer text. Also, the third RTE challenge introduces a repository to share the resources used by researchers. The fourth RTE challenge (Giampiccolo et al., 2008) was proposed at the Text Analysis Conference (TAC). The goal of the fourth RTE challenge is to make a 3-way classification of the given T-H pair. Entailment, Contradiction, and Unknown are the three classes.

The fifth RTE challenge (Bentivogli et al., 2009) involved using a real text corpus with a longer average text length. The sixth RTE challenge (Bentivogli et al., 2010) on recognizing TE in Summarization and Knowledge Base Population(KBP). In summarization given a corpus, a set of candidate sentences are retrieved from the corpus. Then the task is to identify all sentences which entails the given Hypothesis H among the candidate sentence. In the KBP validation task, the RTE system has to validate the output of the system participating in the KBP slot filling task. The seventh RTE challenge (Bentivogli et al., 2011) also focused on Summarization and KBP. Here in summarization, the corpus is relatively larger up to a paragraph long. The eighth RTE challenge (Dzikovska et al., 2013) is organized as the joint challenge at SemEval-2013. The task was to label students answers with categories to help generate effective feedback on errors. And the community was offered a 5-way student response labeling task and 3- way and 2- way RTE style tasks on educational data.

4. Datasets

Machine learning is only as good as the training data that we use for the model to learn. For RTE, various datasets are available for the researchers to use. The FraCaS test suite (Cooper et al., 1996) comprises 346 NLI problems manually constructed by experts in the mid-1990's. Each NLI problem in FraCaS consists of one or more premise sentences followed by the question sentence and answer. Most FraCaS problems are labeled with three answers. YES indicates the hypothesis can be inferred from the premise. NO indicates the hypothesis contradicts the premise. UNK indicates the hypothesis is compatible with the premise. Eight RTE challenges have provided the dataset for researchers to evaluate their approach. The experts constructed the dataset for the eight RTE challenges. RTE-1 to RTE-5 (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Giampiccolo et al., 2008; Bentivogli et al., 2009) compromises the traditional RTE task dataset. In contrast, the other three datasets (Bentivogli et al., 2010; Bentivogli et al., 2011; Dzikovska et al., 2013) are created for a more realistic scenario and in a specific application setting of the TE.

Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014) is a corpus for the 2014 SemEval shared competition task. About 10K examples were created by using the crowdsourcing technique for entailment and semantic similarity. Each pair is annotated for the relatedness (a 5 point scale rating) and an entailment relation (with three possible labels: entailment, contradiction, and neutral). Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) was the first corpus created which encouraged neural network models for NLI. The SNLI corpus contains 570K of sentence pairs labeled for entailment, contradiction, and semantic independence. In SNLI, premises are derived from the image caption, whereas crowd workers created the hypotheses.

Multi-Genre NLI (MNLI) (Williams et al., 2017) is a follow-up to SNLI. MNLI contains about 433K examples. SNLI contains examples only from the single text genre. To overcome this limitation, premises in MNLI are collected from ten different written and spoken languages. The crowd wrokers had written the hypotheses. In Multi-Premise Entailment (MPE) (Lai et al., 2017), each hypothesis is paired in an unordered set of independently written premises describing the same event. MPE contains about 10K examples where the entailment relation is categorized using three labels: entailment, contradiction, and neutral. Cross-lingual NLI (XNLI) (Conneau et al., 2018) is the development and test set of MNLI translated into 15 different languages. XNLI consists of 7500 human-annotated examples in each language. Thus a total of 112,500 annotated pairs are available.

SciTail (Khot et al., 2018) is the first entailment dataset created from existing text. In SciTail, hypotheses are derived from science questions and their corresponding candidate answers. The premises are from relevant web sentences retrieved from a large corpus. SciTail contains about 27K pairs of examples. TabFact (Chen et al., 2019) is the dataset constructed for verifying whether a textual hypothesis holds based on the given evidence from a semi-structured dataset. TabFact consists of 118K manually annotated statements, which are labeled as either entailed or refuted statements.

5. Textual Entailment in Question Answering

RTE can be performed in four ways. The four such approaches include the lexical approach, semantic approach, logical representation, and using the AI model. The paper discusses the different techniques used in the TE-based QA. This section discusses how the different kinds of entailment methods could be used in a QAS.

5.1. Lexical Approach

Lexical Approach for RTE derives the linguistic information from the text to classify whether the text is entailed or not. It does not consider the syntactic and semantic properties of the text. Instead, it only works with the input surface string. TE is used in various subtasks of QA, and those subtasks which use lexical based TE approach are discussed below. Fig. 3 shows a graphical representation of the RTE using the lexical approach.

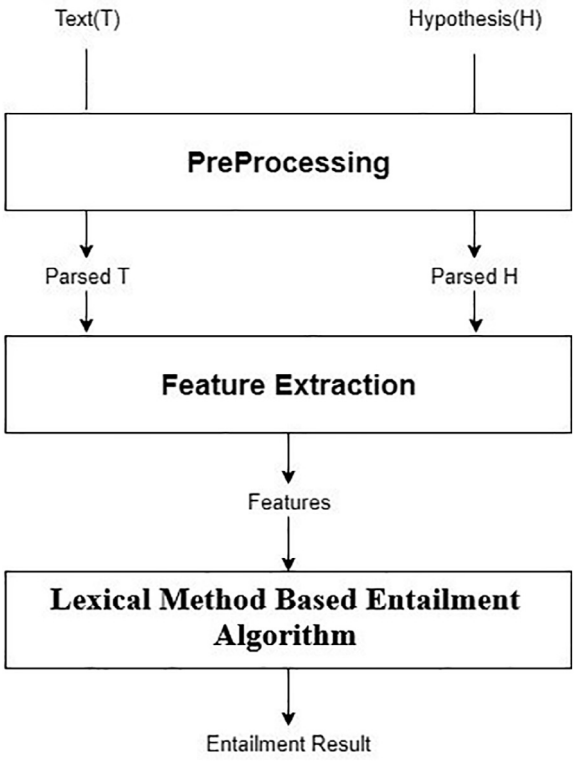

Fig. 3. Lexical Approach.

Re-ranking the candidate answer, Limiting the number of passages retrieved, Identifying the appropriate passage with the answer (Harabagiu and Hickl, 2006): Three different QA ideas using TE have been proposed to achieve better results. The first one uses TE to remove the candidate answers, which does not meet the minimum entailment confidence. The remaining answers are re-ranked accordingly. The second architecture limits the passage retrieved for answering either by ranking or filtering the passage using the entailment information. The third idea is to establish an entailment relation between the question and the question generated from the retrieved passage. It helps in identifying the passage having accurate results. The TE system consists of a pre-processing module that derives linguistic knowledge from the text pair. It is followed by the Alignment module, which uses the maximum entropy classifier to identify the corresponding entities, predicates, or phrases found in the text's pair to inform the entailment judgment. Finally, a decision tree-based classifier module is used to check whether an entailment relation exists between the pair of text. By using TE, the accuracy has improved by 20% overall. But it fails to capture deeper semantic relations.

Answer Validation (Kouylekov et al., 2006): TE has been adopted for answer validation due to its ability to address the language variability issue in cross-language QA. An edit distance algorithm is applied to discover the TE Relation between the hypotheses question and the text answer. An entailment relation exists when the edit distance between the pair is less than the set threshold. If the edit distance is greater than the threshold, then no entailment exists. Due to TE recognition, more generic questions which were not previously considered were handled effectively. Here a module is missing to convert document fragments into valid TS. Due to the TE recognition, more generic questions were handled.

Automatic Generation of Question pattern (Ou et al., 2008): Predictive questions are a set of question patterns predicted to be asked by the user in the specified domain. Along with the predictive question, their corresponding query template is also generated. A question template is used for retrieving the answer. So, when a given user question entails the generated predictive question, then the answer to the predictive question is expected to be the subset of the answer to the given user question. Here the bag-of-words method is adopted as an entailment engine. The question pattern is automatically generated by using the query template of the entailed predictive questions. To improve the performance of the system the predictive question should be able to cover all types of question patterns that are expected to be asked from the domain ontology.

Question analysis (Negri and Kouylekov, 2009): The relation extraction task is adopted for question analysis in the framework of restricted domain QA. Here all the relations of interest have to be extracted from the natural language question. The model consists of a repository containing all relational textual patterns. Here, TE-based Relation extraction is defined as a classification problem. If an entailment relation holds between the question and at least one pattern associated with the relation R_i, then the question expresses the relation Ri. Here for recognizing TE edit distance algorithms such as linear distance and tree edit distance were used. An entailment relation holds between the text and hypothesis only if the overall cost of transformation is below a certain threshold. Although the author has achieved positive results by adopting the TE in the question analysis framework on structured data, it fails to capture semantic relation.

In the Lexical Approach for question entailment, four subtasks have been discussed. Table 1 analyzes the linguistic methods adopted to approach the identified subtask in QA, along with the results achieved and their limitations. The table shows that a significant drawback of this approach is that it does not capture the semantic relations.

5.2. Logical Representation

Processing TE where one text fragment can be inferred from the other fragment with the help of some representation language.

Table 1
Lexical Approach – Analysis.

Task	Entailment Method	Results Achieved	Limitations
Re-ranking the candidate answer, Limiting the number of passages retrieved, Identifying the appropriate passage with the answer	A classification problem from the linguistic information derived from the pair of text.	52% (Answer type detected)	Does not approximate the more profound semantic phenomena
Answer Validation	Edit Distance Algorithm	22.87% (Italian/ Italian)	Missing module to convert document fragment into valid TS
Automatic Generation of Question Pattern	Bag of Words	71.76% (Questions outside the domain)	Can not cover all kind of questions. Some generated question patterns does not make sense.
Question Analysis	Edit Distance Algorithm	72.9% (F1)	Fails to capture the semantic relation

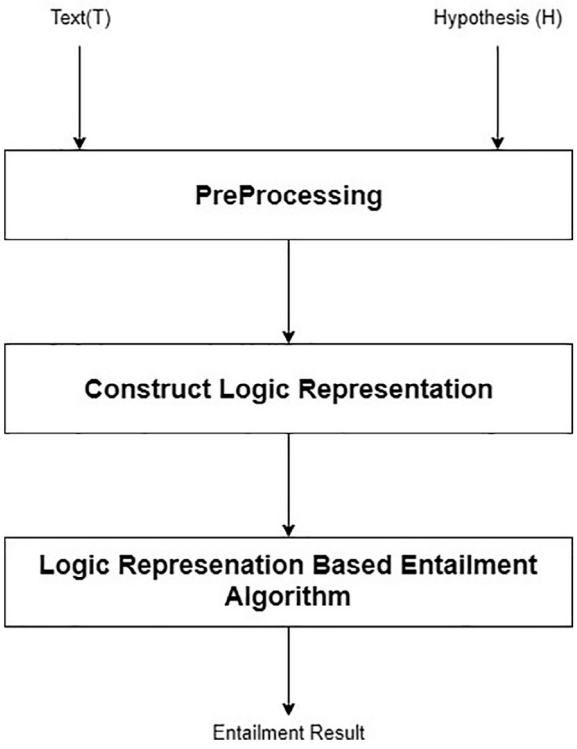

Fig. 4. Logical Representation Approach.

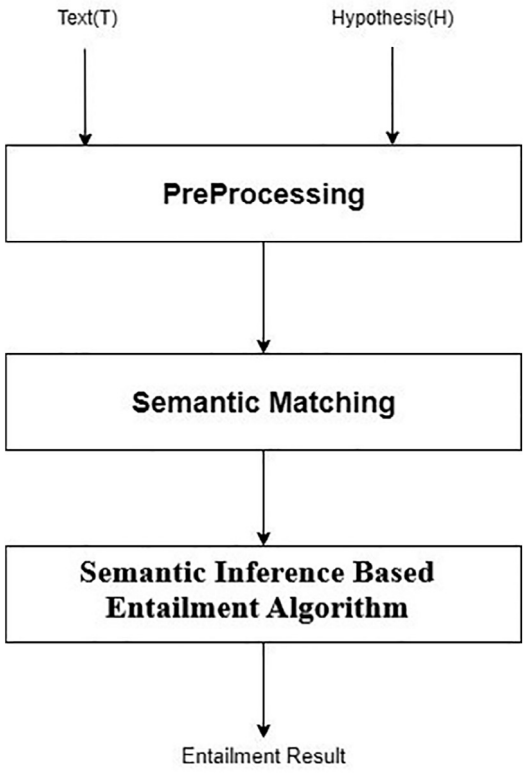

Fig. 5. Semantic Approach.

Using the representation language, the knowledge required from the pair of text is obtained. Fig. 4 shows a graphical representation of the RTE using logical representation.

Answering Yes or No questions (Kim and Goebel, 2017): A two-phase approach for legal bar question and answering has been developed to use the entailment as the deciding factor. The first phase is ad hoc IR. After IR, TE is used for answering *Yes or No* questions. If the question is easy, the entailment is directly obtained from the logic representation. If the question is difficult, then entailment is obtained using unsupervised learning such as K-means clustering algorithm. Although there are errors due to the inability to capture the semantic similarity, this article achieved the best performance in COLIEE 2017.

A subtask of answering yes or no questions has been identified in the logical representation for question entailment. And in Table 2, the results achieved and the limitations are drafted. The semantic constraint of the pair of texts was not fully captured by this method.

5.3. Semantic Approach

The semantic-based approach considers the meaning of the text to process TE rather than considering only the input surface string in the lexical based approach. Fig. 5 depicts a graphical representation of the RTE using a semantic approach.

Answer Extraction (Sacaleanu et al., 2008): For answer extraction in the structured data, a TE engine that captures semantic inference is used. Given a question and a set of relational answer patterns, the QA system selects the patterns that are entailed by the question. The answer pattern is associated with a high precision procedure for retrieving the answer. The entailed patterns for the input question are used for answering the input user question. This approach not only yields good results in the monolingual but also in the cross-language setting.

Answer Retrieval (Ou et al., 2009): The difficulties faced in his previous work (Ou et al., 2008) are addressed by question classification. Question classification in the system helps the entailment engine to find the entailed hypothesis question. A syntactic or semantic engine is used to find the textually-entailed hypothesis question. Once the entailed hypotheses questions are found, the corresponding query templates are used to retrieve the answers for the user question. This method is suitable for Ontology-based QA in the restricted domain. The performance has been enhanced compared to its previous model.

Answering Polar Questions (Kim et al., 2013): A hybrid method was proposed using TE to have better performance in answering the polar questions. After identifying the relevant legal document to the exam questions, RTE is used to answer the questions. A simple rule-based model is used to solve easy questions. But for the hard questions, unsupervised learning is used. The learning method on the linguistic features is based on confirming the semantic entailment features. It is the first paper to employ a hybrid method using textual entailment to have better performance.

Semantic Web Technology Question Answering (Ou and Zhu, 2011): The task of converting a natural language query into a compliant ontology query using deep linguistic processing is difficult. A method has been proposed to avoid the difficulty. For a considered domain ontology, a set of question templates and their corresponding query template are generated offline. For TE recognition, either a syntactic engine or semantic engine is used. The TE recognition is

Table 2
Logical Representation – Analysis.

Task	Entailment Method	Results Achieved	Limitations
Answering *Yes or No* questions	Logic representation from syntactic analysis tree	71.79%	Inability to capture semantic constraint

found between the user question and the set of question templates. The question templates, which entails the user question, are discovered. The entailed query templates are used for completing the query for the user question. Though it is suitable for Ontology-based QA in a restricted domain, this method requires a profound question analysis.

Five subtasks have been discussed in the Semantic Approach for question entailment. Table 3 examines the semantic methods used to approach the identified subtask in QA, as well as the outcomes and limitations. Although various techniques are used to capture semantic relationships, the lack of a required structured dataset is a disadvantage.

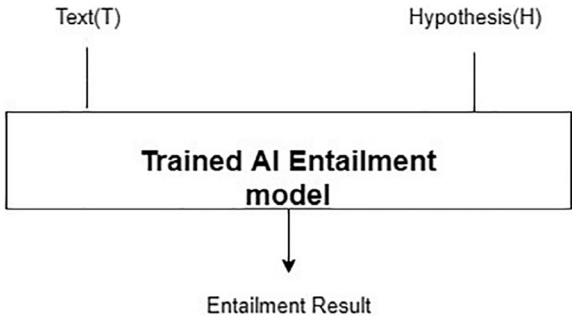

Fig. 6. AI Approach.

5.4. AI Approach

In the AI method, a model of the world is built to recognize TE and then a pair of text is given as input to check whether one text can be inferred from the other or not. Fig. 6 shows a graphical representation of the RTE using the AI approach.

Filtering question–answer pair (Liu et al., 2020): To improve the quality of question and answer pairs generated from an unlabeled text corpus ACS-QG (Answer-Clue-Style-aware Question generation) model is proposed. This model generates high-quality and diverse question–answer pairs. In the ACS-QG model, to ensure the generated question–answer pairs' quality, a filter is used to remove low-quality question–answer pairs. The filter consists of the entailment model and QA model. For the BERT-based entailment model, the SQuAD 2.0 dataset is used to train a classifier, which tells us whether a pair of question and answer match the original sentence. By combining both the entailment model and QA model low-quality QA pairs are removed. The dataset from the other entailment task can be used to enhance the entailment model. The model is significantly better than most of the advanced neural question generation models in terms of quality and scalability.

Answer Ranking (Wang and Nyberg, 2017): The QA system has been improved by introducing a paraphrase identification module based on the neural dual entailment model known as Bidirectional recurrent neural network model. To determine whether two different questions have the same semantic meaning the paraphrase identification module is used. In real-time, when questions are posted the paraphrase identification module will identify the questions having the same meaning semantically. It significantly improves the performance in answer ranking. The system was recognized as the model with the best performance at the TREC 2017 LIVEQA challenge.

Medical Question Answering (Abacha and Demner-Fushman, 2019): Both IR and TE are combined for medical QA. A dataset has been created that consists of 47K question–answer pairs. The authors have previously done works related to RTE. In the previous work, a clinical dataset consisting of consumer health questions was introduced. And also, a feature-based method for RTE was adopted, which achieved around 75% accuracy. The entailment between two questions is defined as if every answer of A is also a correct answer for B, then question B entails the question A. Hence to retrieve the answer for the user/premise question, it is done by retrieving inferred/entailed questions called hypothesis questions which are already associated with the answers. Here the architecture proposed by the author first uses IR for selecting candidate questions because classifying the entire QA collection in real-time is quite not feasible. Next, the retrieved candidate questions are classified depending on whether they are entailed by the user question. Candidate questions are ranked by combining the scores of IR and TE. By using Recognizing Question Entailment (RQE), overall performance is improved.And outperforms the best results from MedicalLiveQA 17 by a factor of 29.8%. But this model has low performance on consumer health care questions.

Multi-Hop Question Answering (Trivedi et al., 2019): TE is used in the sentence relevance module and the aggregation module for QA, requiring multiple sentence reasoning. The first component consists of the sentence relevance module. Only the relevant sentence is focused by using the local hypothesis between each premise and the answer hypothesis sentence. The second component consists of the multi-layer aggregate module, which uses global entailment between all premises and the answer hypothesis. Here, the sentence relevance module is used to sort out the sentences that are not relevant. The multi-layer aggregate module is used to combine the obtained entailment information from the sentence relevance module. This model achieves the best result on OpenBookQA and MultiRC datasets.

Visual Entailment (Xie et al., 2019):Visual Question Answering (VQA) goal is to answer natural language questions based on the visual information provided. (Xi et al., 2020; Wu et al., 2021). A new inference task called the Visual Entailment(VE) is introduced to overcome the limitations in the VisualQA datasets. The goal of the model is to predict whether an image semantically entails a text. SNLI-VE is a dataset consisting of the image-sentence pairs. The premise is characterized by an image. Explainable Visual Entailment (EVE) has been proposed to address the VE task. EVE is composed of text and image branches similar to the attention top-down or bottom-up model. Attention top-down or bottom-up is the winner of VQA challenge 2017. The text branch extracts the features from the input text hypothesis through an RNN. The

Table 3
Semantic Approach – Analysis.

Task	Entailment Method	Results Achieved	Limitations
Answer Extraction	Entailment engine to capture semantic inference	58.25% (Average of different languages)	Lack of corresponding answer pattern
Answer Retrieval	Semantic information obtained from domain ontology aligned with wordnet	65% (Semantic Engine)	Can not cover all kind of user questions
Answering Polar Questions	Semantic Equivalence between input question and relevant law article	61.13%	Difficult to extract all semantic entailment features
Semantic Web Technology Question Answering	Semantic information obtained from domain ontology aligned with wordnet	65.6%	Requires deep question analysis

image branch generates image features from the premise image. The two branches' features have used an input to the fully connected layers for the final conclusion.Though it achieves accuracy, it is not trained to distinguish all the fine-grained information.

Progressive Visual Question Answering (Si et al., 2021):Language prior problem is the superficial correlation bias caused by the accidental correlation between answers and questions. This paper proposes a select and rerank progressive framework based on visual entailment to address the language prior problem. First, the top N candidate answers relevant to the question and the image is selected using visual question entailment. Then the candidate answers are reranked based on the visual entailment task that verifies whether the image semantically entails the synthetic statement of the question and the candidate answer.

Cooking Domain- Question Answering (Pathak et al., 2021): An automated system using entailment in QA is explored in this paper. In the cooking domain QA system, the SVM classifier is used to detect entailment relation between end-user questions and the questions contained inside the Knowledge Base (KB). It is followed by the retrieval of the answer corresponding to the prominently entailed KB questions.

MEDIQA 2019 (Abacha et al., 2019): The three tasks represented in MEDIQA are NLI, RQE, and QA. NLI labels the relation between two sentences, such as *Entailment, Neutral, and Contradiction*. RQE gives whether two sentences are entailed or not. QA filters and improves the ranking of the automatically retrieved answers. Here the overview of the approaches and results of the participants of the MEDIQA Challenge are discussed. For NLI, most teams had built their models on BERT since it is pre-trained on an extensive domain corpus. Also, since the dataset for NLI is from the clinical domain, the variations of BERT were used. Another standard model used for NLI was Multi-Task Deep Neural Network(MT-DNN). For RQE approaches combining the ensemble methods and transfer learning of multi-task language, models had given the best results. For QA, many teams have used their RQE and/or NLI models.

ANU-CSIRO (Nguyen et al., 2019): A new system has been proposed that combines both open-domain and biomedical domain approaches to improve semantic understanding and ambiguity resolution. For the open-domain ensemble approach and biomedical domain BERT and BioBERT models are used. Due to the similarity of the three tasks, a shared model was utilized.

ARS-NITK (Agrawal et al., 2019): An InferSent model was adopted for the task of NLI. For the task RQE, XGBoost with PubMed Embeddings was introduced. And finally, Bi-directional LSTMs trained on the SquAD dataset are used for the QA task.

Double Transfer (Xu et al., 2019): For the Natural Language Understanding(NLU) tasks in the medical domain, a multi-source transfer learning approach is used to transfer the knowledge from the MT-DNN and SciBERT. The authors propose a method of fine-tuning both MT-DNN and SciBERT using multi-task learning and integrating models from both domains with ensembles.

Dr.Quad (Kumar et al., 2019): MT-DNN is used for the NLI and RQE tasks. The MT-DNN model combines the multi-task learning (MTL) strength and the language model pre-training. MT-DNN uses BERT as the encoder and uses MTL to fine-tune the multiple task-specific layers. Question Answering re-ranking system uses a simple model to combine the NLI and RQE models.

DUT-BIM (Zhou et al., 2019a): The performance of the QA task can be improved by understanding the semantic relation between the question and answer. To extract semantic relation between different words in the question and answer, a BioBERT Transformer is used. BioBERT is used to encode contextual information. Transformers are used to learn the long-range dependency information between the words in question and answer.

DUT-NLP (Zhou et al., 2019b): The Adversarial Multi-task Network (AMTN) model was proposed to jointly model the RQE and QA. AMTN utilizes a BioBERT and an Interactive Transformer, which can be viewed as MTL to learn the shared semantic representation. Adversarial Training is intoduced in the Multi-task framework so that shared representation has more common information and reduces the mixing of the task-specific information.

IIT-KGP (Sharma and Roychowdhury, 2019): QSpider model is proposed for medical QA through TE. This method captures the entailment between two questions by capturing the question types. Qspider consists of the state-of-the-art model Sci-BERT used to capture the question type and the semantic relation followed by the Gradient Boosting classifier to check for entailment. The question type which was captured is used as the feature to detect the question entailment.

IITP (Bandyopadhyay et al., 2019): Multiple deep learning-based systems are utilized for the three tasks proposed in the MEDIQA challenge. Five system results are submitted for each of the NLI and RQE tasks. Four system results are presented for the QA task. Most of the models proposed have BERT/BioBERT embeddings and BM25.

Ku_ai (Cengiz et al., 2019): A BERT encoder combined with the classification head is used for the NLI task. The classification head outputs the probabilities from a three-way softmax. The probabilities corresponds to the labels that the sentence pair can have.

lasigeBioTM (Lamurias and Couto, 2019): A common architecture is used for all three tasks, which are then fine-tuned for the given specific task using the training data presented. BioBERT was used for all three tasks, with minimal changes made for each task.

MSIT_SRIB (Chopra et al., 2019): A Biomedical Multi Task Deep Neural Network(Bio-MTDNN) is adopted for the NLI task. The model utilizes a transfer learning paradigm. Also, the model integrates the knowledge from external sources.

NCUEE (Lee et al., 2019): BERT-BiLSTM-Attention architecture is used for the NLI task. BERT is used as the word embedding method to integrate the BiLSTM network with an attention mechanism. In the end, a softmax activation function is used to classify the sentence pair.

PANLP (Zhu et al., 2019): Various sets of models have been experimented for each task. Knowledge distillation has been integrated to boost the performance of the single models. Model ensemble, Transfer learning, and re-ranking mechanism have also been implemented to improve the performance. Either BERT or MT-DNN has been used in most of the proposed models.

Pentagon (Pugaliya et al., 2019): It is an entailment-based approach for re-ranking and filtering answers in the medical domain. For NLI and RQE, MT-DNN models are proposed. The MT-DNN based models are used as a feature extractor. The features obtained from the NLI and RQE are used for candidate answer selection and re-ranking on the medical dataset.

Saama Research (Kanakarajan et al., 2019): A BERT pre-trained on various data is adopted for the task of NLI. And finally, BERT fine-tuned on the MIMIC III v1.4 is proposed for solving the NLI task.

Sieg (Bhaskar et al., 2019): A MT-DNN model is used for NLI and RQE. The lower levels of MT-DNN are standard, whereas the upper layers are task-specific. An Infersent model has been adopted as the baseline for the NLI Task. An SVM model is considered the baseline for the RQE task. The shared layers for both tasks consist of the variants of the BERT model.

Surf (Nam et al., 2019): The pretrained language models such as BioBERT and PubMed-ELMo along with the transfer learning method is used for the NLI task to deal with the expanding medical abbreviation which can be considered as the general problem in medical domain.

UU_Tails (Tawfik and Spruit, 2019): BERT Embeddings and different ensemble methods were proposed for the NLI task. Whereas for the RQE task, a transformer base architecture of the USE (universal sentence encoder) is used during the submission.

UW-BHI (Kearns et al., 2019): BERT, Cui2Vec, and Embeddings of Semantic Predictions (ESP) are used for comparing the performance and the internal representation for the NLI task. As a result, BERT embeddings fine-tuned using PubMed and MIMIC-III gave the best performance.

WTMED (Wu et al., 2019): A hybrid approach has been adopted for the NLI task. The model consists of the text encoder as the core component. The syntax encoder and the feature encoder are used to capture the syntactic and domain-specific information. Then the output of the three baseline models are combined by an ensemble module. At the end of the hybrid approach, a conflict resolution strategy has been integrated.

Seventy-Two teams participated in the challenge on the Alcrowd platform. The official score includes only the teams who have sent the working notes paper describing the approach.Out of the twenty teams that have been recognized for their official score, only eight have submitted all three tasks. Seven teams submitted solutions for the NLI task, while only one team submitted for the RQE task. Two teams participated in both NLI and RQE, one team only in QA, and the last team in both RQE and QA. In MEDIQA 2019 challenge, 98 percent accuracy was achieved in the NLI task, 74.9 percent in the RQE task, and 78.3 percent in the QA task.Table 4 lists all of the details about the accuracies achieved and the tasks completed by the teams. Several research works have been discussed in the AI Approach for question entailment. Table 5 examines the semantic methods used to approach the identified subtask in QA, as well as their outcomes and limitations.

6. Discussion

The survey is structured based on the method adopted by the researchers for a given task to recognize TE. By categorizing in this

Table 4
Accuracies Achieved in MEDIQA Tasks.

Team	Accuracy Achieved for the Task		
	NLI	RQE	QA
ANU-CSIRO	80.0%	48.9%	58.4%
ARS-NITK	87.7%	66.7%	53.6%
Double Transfer	93.8%	66.2%	78.0%
Dr.Quad	85.5%	66.7%	56.5%
DUT-BIM	×	×	74.5%
DUT-NLP	×	63.6%	74.5%
IIT-KGP	×	68.4%	×
IITP	81.8%	53.2%	71.7%
KU_ai	84.7%	×	×
lasigeBioTM	72.4%	48.5%	63.7%
MSIT_SRIB	81.3%	×	×
NCUEE	84.0%	×	×
PANLP	96.6%	74.9%	77.7%
Pentagon	85.7%	67.1%	76.5%
Saama Research	78.3%	×	×
Sieg	91.1%	70.6%	×
Surf	90.6%	×	×
UU_Tails	85.2%	58.4%	×
UW-BHI	81.3%	×	×
WTMED	98.0%	×	×

Table 5
AI Approach – Analysis.

Task	Entailment Method	Results Achieved	Limitations
Filtering question–answer Pair	BERT based neural model	53.25% (ROUGE-L)	Error found in questions due to the problem of semantic mismatching, meaningless or information incompleteness
Answer Ranking	Bidirectional recurrent neural network model	56.00%	Used only the quora training data set
Medical Question Answering	Logistic Regression Model	94.33%	Obtained lower performance on Consumer health questions
Multi-Hop Question Answering	Enhanced Sequential Inference Model	55.80%	Problem in attention in the model.
Visual Entailment	CNN and RNN	71.16%	Not Trained to distinguish fine grained information
Progressive Visual Question Answering	LXMERT (Visual Entailment Degree)	66.73%	Resource limitation
Cooking Domain- Question Answering	SVM Classifier	58.8%(Average Accuracy)	Unavailability of structured cooking data

Table 6
TE based QA – Analysis.

Approach	Outline	Drawbacks
Lexical Approach	To improve the performance by using only linguistic information for RTE.	Fails to capture Semantic relation
Logical Representation	To improve performance uses representation language to make inference	Unable to capture deep semantic constraint
Semantic Approach	Considers the meaning of the text instead of just using the linguistic information	Lack of the required structured data set
AI Approach	Building the model where the model is trained to recognize TE	Unavailability of datasets for training TE

way, the enhancement experienced in RTE over the past few years is seen. Also, the limitations of a particular method for RTE introduce the succeeding method, which performs better by overcoming the shortcomings encountered. The four approaches identified in this survey are Lexical Approach, Logical Representation, Semantic Approach, AI Approach. Table 6 lists the subtasks identified, advantages, and drawbacks of all the four approaches described in this survey.

Over the years, TE-based QA has developed and achieved positive results. In this survey, we can see that all subtasks that adopt TE are domain-specific. The future direction of TE-based QA calls for building datasets for all required domains where RTE can be beneficial. By generating more TE datasets, the quality of training will be improved.

7. Conclusion

A QAS is used to provide a user with a natural language response to a specific question.NLP is a field of computer science concerned with the interactions between the computer and human language. The QA application in NLP has many subtasks. This survey aims to make the reader understand about using TE to achieve the desired goal in the QA subtask. In addition, the survey is organized based on the methods used by the researchers to recognize TE for each task. Several different entailment methods have been identified for RTE. Though each technique for identifying TE has its own advantages and disadvantages, the lack of an entailment dataset for training is a significant impediment to the development of TE-based QE.

The advancement of modern technology has improved the lives of humans, and without doubt, artificial intelligence plays a significant role in making our lives easier. Entailment can be considered a significant task that can be incorporated in any NLP application in the emerging digital era. Deep learning, the subfield of artificial intelligence, is booming in the 21st century due to the availability of voluminous data. We can use this big data to create an entailment dataset. The ability to process large data sets using deep learning empowers us to use entailment to improve the performance of any QA subtask. The goal of future research would be to enhance the performance of the NLP task by using entailment.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abacha, A., Demner-Fushman, D., 2019. A question-entailment approach to question answering. BMC Bioinformatics 20, 1–23.

Abacha, A.B., Shivade, C., Demner-Fushman, D., 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 370–379.

Agrawal, A., George, R., Ravi, S., Kamath, S., Kumar, A., 2019. Ars_nitk at mediqa 2019: Analysing various methods for natural language inference, recognising question entailment and medical question answering system. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 533–540.

Bandyopadhyay, D., Gain, B., Saikh, T., Ekbal, A., 2019. Iitp at mediqa 2019: Systems report for natural language inference, question entailment and question answering. arXiv preprint arXiv:1906.06332.

Bentivogli, L., Clark, P., Dagan, I., Giampiccolo, D., 2009. The fifth pascal recognizing textual entailment challenge. In: TAC.

Bentivogli, L., Clark, P., Dagan, I., Dang, H., Giampiccolo, D., 2010. The sixth pascal recognizing textual entailment challenge. In: TAC.

Bentivogli, L., Clark, P., Dagan, I., Giampiccolo, D., 2011. The seventh pascal recognizing textual entailment challenge. In: TAC, Citeseer.

Bhaskar, S., Rungta, R., Route, J., Nyberg, E., Mitamura, T., 2019. Sieg at mediqa 2019: Multi-task neural ensemble for biomedical inference and entailment. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 462–470.

Bowman, S., Angeli, G., Potts, C., Manning, C., 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326..

Cengiz, C., Sert, U., Yuret, D., 2019. Ku_ai at mediqa 2019: Domain-specific pre-training and transfer learning for medical nli. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 427–436.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., Wang, W., 2019. Tabfact: A large-scale dataset for table-based fact verification. arXiv preprint arXiv:1909.02164..

Chierchia, G., McConnell-Ginet, S., 2000. Meaning and Grammar: An Introduction to Semantics. MIT press.

Chopra, S., Gupta, A., Kaushik, A., 2019. Msit_srib at mediqa 2019: Knowledge directed multi-task framework for natural language inference in clinical domain. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 488–492.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S., Schwenk, H., Stoyanov, V., 2018. Xnli: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053.

Cooper, R., Crouch, R., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., 1996. Using the framework. The FraCaS Consortium. Technical Report. Technical Report, FraCaS deliverable D-16.

Dagan, I., Glickman, O., Magnini, B., 2005. The pascal recognising textual entailment challenge. In: Machine Learning Challenges Workshop. Springer, pp. 177–190.

Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H., 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 263–274.

Giampiccolo, D., Dang, H., Magnini, B., Dagan, I., Cabrio, E., Dolan, B., 2008. The fourth pascal recognizing textual entailment challenge. In: TAC, Citeseer..

Giampiccolo, D., Magnini, B., Dagan, I., Dolan, W., 2007. The third pascal recognizing textual entailment challenge. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 1–9.

Glickman, O., Dagan, I., Koppel, M., 2005. A probabilistic lexical approach to textual entailment. In: International Joint Conference on Artificial Intelligence, Lawrence Erlbaum Associates Ltd. p. 1682.

Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., Szpektor, I., 2006. The second pascal recognising textual entailment challenge. In: Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment.

Harabagiu, S., Hickl, A., 2006. Methods for using textual entailment in open-domain question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 905–912.

raj Kanakarajan, K., Ramamoorthy, S., Archana, V., Chatterjee, S., Sankarasubbu, M., 2019. Saama research at mediqa 2019: Pre-trained biobert with attention visualisation for medical natural language inference. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 510–516.

Kearns, W., Lau, W., Thomas, J., 2019. Uw-bhi at mediqa 2019: An analysis of representation methods for medical natural language inference. arXiv preprint arXiv:1907.04286.

Khot, T., Sabharwal, A., Clark, P., 2018. Scitail: A textual entailment dataset from science question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Kim, M., Goebel, R., 2017. Two-step cascaded textual entailment for legal bar exam question answering. In: Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, pp. 283–290.

Kim, M., Xu, Y., Goebel, R., Satoh, K., 2013. Answering yes/no questions in legal bar exams. In: JSAI International Symposium on Artificial Intelligence. Springer, pp. 199–213.

Kodra, L., Mece, E., 2017. Question answering systems: A review on present developments, challenges and trends. International Journal of Advanced Computer Science and Applications 8, 217–224.

Kouylekov, M., Negri, M., Magnini, B., Coppola, B., 2006. Towards entailment-based question answering: Itc-irst at clef 2006. In: Workshop of the Cross-Language Evaluation Forum for European Languages. Springer, pp. 526–536.

Kumar, V., Srinivasan, A., Chaudhary, A., Route, J., Mitamura, T., Nyberg, E., 2019. Dr. quad at mediqa 2019: Towards textual inference and question entailment using contextualized representations. arXiv preprint arXiv:1907.10136.

Lai, A., Bisk, Y., Hockenmaier, J., 2017. Natural language inference from multiple premises. arXiv preprint arXiv:1710.02925.

Lamurias, A., Couto, F., 2019. Lasigebiotm at mediqa 2019: Biomedical question answering using bidirectional transformers and named entity recognition. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 523–527.

Lee, L., Lu, Y., Chen, P., Lee, P., Shyu, K., 2019. Ncuee at mediqa 2019: medical text inference using ensemble bert-bilstm-attention model. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 528–532.

Liu, B., Wei, H., Niu, D., Chen, H., He, Y., 2020. Asking questions the human way: Scalable question-answer generation from text corpus. In: Proceedings of The Web Conference 2020, pp. 2032–2043.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., 2014. A sick cure for the evaluation of compositional distributional semantic models. In: Lrec, Reykjavik. pp. 216–223.

Nam, J., Yoon, S., Jung, K., 2019. Surf at mediqa 2019: Improving performance of natural language inference in the clinical domain by adopting pre-trained language model. arXiv preprint arXiv:1906.07854.

Negri, M., Kouylekov, M., 2009. Question answering over structured data: an entailment-based approach to question analysis. In: Proceedings of the International Conference RANLP-2009, pp. 305–311.

Nguyen, V., Karimi, S., Xing, Z., 2019. Anu-csiro at mediqa 2019: Question answering using deep contextual knowledge. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 478–487.

Ou, S., Mekhaldi, D., Orasan, C., 2009. An ontology-based question answering method with the use of textual entailment. In: 2009 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, pp. 1–8.

Ou, S., Orasan, C., Mekhaldi, D., Hasler, L., 2008. Automatic question pattern generation for ontology-based question answering. In: Flairs Conference, pp. 183–188.

Ou, S., Zhu, Z., 2011. An entailment-based question answering system over semantic web data. In: International Conference on Asian Digital Libraries. Springer, pp. 311–320.

Pathak, A., Manna, R., Pakray, P., Das, D., Gelbukh, A., Bandyopadhyay, S., 2021. Scientific text entailment and a textual-entailment-based framework for cooking domain question answering. Sādhanā 46, 1–19.

Pugaliya, H., Saxena, K., Garg, S., Shalini, S., Gupta, P., Nyberg, E., Mitamura, T., 2019. Pentagon at mediqa 2019: Multi-task learning for filtering and re-ranking answers using language inference and question entailment. arXiv preprint arXiv:1907.01643.

Sacaleanu, B., Orasan, C., Spurk, C., Ou, S., Ferrandez, O., Kouylekov, M., Negri, M., 2008. Entailment-based question answering for structured data. In: Coling 2008: Companion volume: Demonstrations, pp. 173–176.

Sharma, P., Roychowdhury, S., 2019. Iit-kgp at mediqa 2019: Recognizing question entailment using sci-bert stacked with a gradient boosting classifier. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 471–477.

Si, Q., Lin, Z., Zheng, M., Fu, P., Wang, W., 2021. Check it again: Progressive visual question answering via visual entailment. arXiv preprint arXiv:2106.04605.

Tawfik, N., Spruit, M., 2019. Uu_tails at mediqa 2019: Learning textual entailment in the medical domain. iin: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 493–499..

Trivedi, H., Kwon, H., Khot, T., Sabharwal, A., Balasubramanian, N., 2019. Repurposing entailment for multi-hop question answering tasks. arXiv preprint arXiv:1904.09380.

Wang, D., Nyberg, E., 2017. Cmu oaqa at trec 2017 liveqa: A neural dual entailment approach for question paraphrase identification. In: TREC.

Williams, A., Nangia, N., Bowman, S., 2017. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Wu, Y., Ma, Y., Wan, S., 2021. Multi-scale relation reasoning for multi-modal visual question answering. Signal Processing: Image Communication 96, 116319.

Wu, Z., Song, Y., Huang, S., Tian, Y., Xia, F., 2019. Wtmed at mediqa 2019: A hybrid approach to biomedical natural language inference. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 415–426.

Xi, Y., Zhang, Y., Ding, S., Wan, S., 2020. Visual question answering model based on visual relationship detection. Signal Processing: Image Communication 80, 115648.

Xie, N., Lai, F., Doran, D., Kadav, A., 2019. Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706.

Xu, Y., Liu, X., Li, C., Poon, H., Gao, J., 2019. Doubletransfer at mediqa 2019: Multi-source transfer learning for natural language understanding in the medical domain. arXiv preprint arXiv:1906.04382.

Zhou, H., Lei, B., Liu, Z., Liu, Z., 2019a. Dut-bim at mediqa 2019: Utilizing transformer network and medical domain-specific contextualized representations for question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 446–452.

Zhou, H., Li, X., Yao, W., Lang, C., Ning, S., 2019b. Dut-nlp at mediqa 2019: an adversarial multi-task network to jointly model recognizing question entailment and question answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 437–445.

Zhu, W., Zhou, X., Wang, K., Luo, X., Li, X., Ni, Y., Xie, G., 2019. Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowledge distillation. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 380–388.