Answer Selection in Community Question Answering Exploiting Knowledge Graph and Context Information

Golshan Afzali Boroujeni[a], Heshaam Faili[a,b,*], Yadollah Yaghoobzadeh[a]
[a] *School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran;*
[b] *Iran and School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Iran*
Emails: golshan.afzali@ut.ac.ir, hfaili@ut.ac.ir, y.yaghoobzadeh@ut.ac.ir

Editor(s): Mehwish Alam, FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Germany; Davide Buscaldi, LIPN, Université Sorbonne Paris Nord, France; Michael Cochez, Vrije University of Amsterdam, the Netherlands; Francesco Osborne, Knowledge Media Institute, (KMi), The Open University, UK; Diego Reforgiato Recupero, University of Cagliari, Italy; Harald Sack, FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
Solicited review(s): Ruijie Wang, University of Illinois at Urbana-Champaign, USA; Simon Gottschalk, L3S Research Center, Germany; 3 anonymous reviewers

Abstract. With the increasing popularity of knowledge graph (KG), many applications such as sentiment analysis, trend prediction, and question answering use KG for better performance. Despite the obvious usefulness of commonsense and factual information in the KGs, to the best of our knowledge, KGs have been rarely integrated into the task of answer selection in community question answering (CQA). In this paper, we propose a novel answer selection method in CQA by using the knowledge embedded in KGs. We also learn a latent-variable model for learning the representations of the question and answer, jointly optimizing generative and discriminative objectives. It also uses the question category for producing context-aware representations for questions and answers. Moreover, the model uses variational autoencoders (VAE) in a multi-task learning process with a classifier to produce class-specific representations for answers. The experimental results on three widely used datasets demonstrate that our proposed method is effective and outperforms the existing baselines significantly.

Keywords: community question answering, knowledge graph, context, convolutional-deconvolutional, variational autoencoder.

1. Introduction

Knowledge graphs (KGs), such as DBpedia [1] and BabelNet [2], are multi-relational graphs. They consist of entities and relationships among them. Many applications such as sentiment analysis [3], recommender systems [4], relation extraction [5], and question answering integrate the information in KGs by linking the entities mentioned in the text to entities in the KGs.

Community question answering (CQA) forums, such as Stack Overflow and Yahoo! Answer provide new opportunities for users to share knowledge. In these forums, anyone can ask any question, and a question is answered by one or more members. Unfortunately, there is often no evaluation of the given answers in how much they are related to the question. It means one has to go through all possible answers for assessing them, which is exhausting and time-consuming. Thus, it is essential to automatically identify the best answers for each question.

In this paper, we address the task of answer selection. As defined in SemEval 2015 [6], in this task, the goal is to classify the answers given a question into three categories: (i) good, which are the answers that address the question well (ii) potentially useful to the user (e.g., because they can help educate him/her on the subject)

*Corresponding author. E-mail: hfaili@ut.ac.ir

(iii) bad or useless. It should be noted that a good answer is an answer semantically relevant to the question, not necessarily the correct answer.

Table 1 shows two examples of questions, each with four answers, taken from the SemEval 2015 [6] dataset[1]. As shown in this table, in CQA, each question has at least three parts: (i) question category, which is the category that the question belongs to; (ii) question subject, which summarizes the question, and (iii) question body, which describes the question in details, and might contain useless or noisy parts as well. Most of the questions and answers in these forums are often

Table 1

Example of two questions and four of their answers from the SemEval 2015 dataset.

	Example 1	Example 2
Question category	Life in Qatar	Qatar Living Lounge
Question subject	Vodka cost	Nissan Offer
Question body	hey guys just wondering , what is the cost of a bottle of vodka in doha ? i dont mean from a hotel , but from the single bottle shop that is set up there . this is my favourite tipple ... ... thanks	Saw an ad in today's GT.. some offer for Nissan Vehicles. Pathfinder for QR. 89,000/- onwards...and Xterra is QR.93,000. and Armada is QR.118,000. I thought Pathfinder is more expensive than Xttera. Anyone know why Pathfinder is so cheap? Did the prices come down or is it a good offer price?
Answer 1 (good)	good to clean house piping russian standard , zubrovka , movscoscaya not passing 100 qr . go for the first two , extra frozen and do n't forget the caviar check mate	basic ones fs , fully loaded ones will cost much more
Answer 2 (good)	ketel 1 is qar240	2009 models are on offer . basic xe qr 101 000 automatic transmission , power window no cd player xe qr 111 000 cd player m cruise control , alloy wheel , power window..etc
Answer 3 (bad)	thanks weasal - i would also like to see the same drive thru style bottle o as well !	take a guess !
Answer 4 (potentially useful)	you need to make sure you have your rp and alcohol permit before you purchase any vodka however i have heard there are some brands we have not heard of in oz that are pretty cheap if you are willing to try them	call them again and check how much is Safari or Infiniti FX35

lengthy, informal, and contain abbreviations and grammatical mistakes. In these examples, for each question, the first and the second answers are labeled as ‘good’ as both try to answer the question. These two answers are both semantically relevant to the question, even though they might be inaccurate or completely wrong. The third answer is ‘bad’ because it is completely irrelevant to the question. The final answer is labeled as ‘potentially useful’ because while it does not provide a relevant answer to the question, it contains a relevant advice for the user.

The main difficulty is how to bridge the semantic gap between question-answer pairs. In other words, by recognizing the semantic relatedness of the question and answer, one can decide about the relevance of the question and its answers.

Early work in this area includes feature-based methods for explicitly modeling the semantic relation between the question and answer [6, 7]. With the great advances in deep neural networks, most recent researches apply deep learning based methods to answer classification in question answering communities [8-12]. These methods typically use a Convolutional Neural Network (CNN) [13] or Long Short term Memory (LSTM) [14] network for matching the question and answer. However, these methods have not achieved high accuracy due to some reasons. The **main challenges** remaining in this field are as follows:

- Despite the usefulness of commonsense or factual background knowledge in the KGs (such as DBpedia [1], BabelNet [2], etc.), to the best of our knowledge, these KGs have been rarely integrated in the recent deep neural CQA networks. KGs provide rich information about entities, specially named entities, and relations between them. Considering the examples in Table 1, named entities “Armada” and “Infiniti FX35” in the question and answer, do not exist in the available

[1] http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools

word embedding methods such as Word2vec [15] or Glove [16] and so, are out-of-vocabulary. Therefore, the conventional methods assign a negative score to the first answer due to their misunderstanding of named entities and their relations. However, by using a comprehensive KG like BabelNet, the model can assign the correct label to the answer due to the entities and facts exist in it.

- There are some words that may have different meanings in different contexts. By using the category of the question as the context representative, the correct meaning of the question and answer words can be extracted, and so a more accurate representation of the question and answer would be generated.
- The previous methods are unable to encode all semantic information of the question and answer. Also, in [17] it has been shown that it is difficult to encode all semantic information of a sequence into a single vector;

In semantic matching problems, the learned representations must contain two main properties. First, the representation must preserve the important details mentioned in the text. Second, each learned representation must contain discriminative information regarding its relationship with the target sentence. Following this motivation, by leveraging the external background knowledge and question category, we use deep generative models for question-answer pair modeling. Due to their ability to obtain latent codes that contain essential information of a sequence, we expect that their resulting representations can suite the question-answer relation extraction better.

In the proposed model, at the first step, the question and answer words are disambiguated based on the question category and external background knowledge from our selected KG. At the end of this step, the correct meaning of each word in the current context is captured. In the second step, by using the representation of the question subject as the attention source, the noisy parts of the question and answer are discarded and the useful information of them is extracted. At the final step, by using the convolutional-deconvolutional autoencoding framework, which is first proposed in [18] for paragraph representation learning, the representations of questions and answers are learned. This framework, which uses the deconvolutional network as its decoder, is used to model each of the question and answer separately. In this multi-task learning process, the question-answer relevance label information is also considered in the representations learning, enabling class-specific representations.

The **main contributions** of our work can be summarized as follows:

- We leverage external knowledge from KGs to capture the meaning of the question and answer words and extract the relation between them.
- We propose to use the category of the question as context to understand the correct meaning of the question and answer words in the current context. To the best of our knowledge, we are the first to use the question category to have context-aware representations in CQA.
- We propose to use two convolutional-deconvolutional autoencoding frameworks that attempt to make separate representations of the question and answer. To the best of our knowledge, we are the first to use this deconvolutional VAE in answer selection problem.
- We introduce a new architecture for answer selection, in which a classifier combined with variational autoencoders to make the representations class-specific.
- Our proposed model achieves state-of-the-art performance in three CQA datasets: SemEval 2015, SemEval 2016 [19], and SemEval 2017 [20].

In the next section, we provide preliminaries in this field. Then we review some previous researches in Section 3. The proposed idea is presented in Section 4. In Section 5, experimental results and analyses are presented. The conclusion is given in Section 6.

2. Preliminaries

2.1. Latent-variable model for text processing

The most common way to obtain sentence representations is to use sequence-to-sequence models, due to their ability to leverage information from unlabeled data [21]. In these models, first an encoder encodes the input sentence x into a fixed-length vector z, and then the output sentence $\tilde{x}$ is reconstructed from z through a decoder network. Specifically, in the autoencoder models, the encoder is a deterministic function and the output of the decoder is the reconstruction of the input sentence x. A problem with autoencoders for text is the deterministic nature of the encoder function, which results in poor model generalization. Variational autoencoders (VAEs) [22] provide a probabilistic manner for describing an observation in a latent space, instead of a vector.

In VAEs, the decoder network reconstructs the input conditioning on the samples from the latent code (via its posterior distribution). Given an observed sentence x, the VAE objective is to maximize the variational lower bound, as follows [22]:

$$z \sim Enc(x) = q(z|x),\ \tilde{x} \sim Dec(z) = p(x|z) \quad (1)$$

$$\begin{aligned} & L_{VAE} \quad (2) \\ & = E_{q_\emptyset(z|x)}[log\ p_\theta(x|z)] \\ & - D_{KL}(q_\emptyset(z|x)|p(z)) \\ & = E_{q_\emptyset(z|x)}[\log p_\theta(x|z) \\ & + \log p(z) - \log q_\emptyset(z|x)] \\ & \leq log \int p_\theta(x|z)p(z)dz = \log p_\theta(x) \end{aligned}$$

In Eq. (1), q and p are the encoder and decoder probabilistic functions, respectively. In Eq. (2), $\emptyset$ and θ are the encoder and decoder parameters, respectively. The lower bound $L_{VAE}(\theta, \emptyset; x)$ is maximized with respect to these parameters.

2.2. Challenges of VAEs for text

Typically, the LSTM networks is used as the decoder in VAEs for text generation [23]. However, due to the recurrent nature of LSTMs, the decoder tends to ignore the information of the latent variable. Providing the ground-truth words of the previous time steps during training process, prevents the learned sentence embeddings to have enough information about the input [23]. To resolve this problem, we use a deconvolutional network as decoder shown to have the best performance among the other methods [24]. As said in [24], deconvolutional networks are typically used in deep learning networks for up-sampling fix-length latent representations usually made by a convolutional network.

3. Related work

3.1. Applications of knowledge graphs

In many NLP and ML applications, KGs are integrated in the models, e.g., sentiment analysis [25], [3], recommender systems [4, 26], relation extraction [5], entity linking [27], and question answering (QA). For the QA problem, the authors in [28] use KG embeddings for answering the questions, especially simple questions. The work done in [29] is also in the QA field which leverages relation phrase dictionaries and KG embeddings for answering the questions in natural language. In [30], a model is presented that uses KGs for question routing in CQA. In this model, topic representations with network structure are integrated into a unified KG question routing framework. The work done in [31] presents a survey on the representation, acquisition, and applications of KGs.

3.2. Answer selection in CQA

In the literature, the methods for answer classification can be roughly divided into two main groups: feature-based and deep learning methods.

Feature-based methods, with a long research history, employ a simple classifier with manually constructed features. In these methods, some textual and structural features are selected and a simple classifier such as support vector machine (SVM) or KNN is applied to them. The methods presented in [6], [7], [32], [33], [34], [35], [36], [37], and [38], are all in this category. Some of these papers along with their features are summarized in Table 2.

In 2015, SemEval organized a similar task to ours, titled “answer selection in community question answering”. Thirteen teams participated in that challenge. The participants mainly focused on defining new features to capture the semantic similarity between the question and its answers. Word matching features, special component features, topic-modeling-based features, non-textual features, etc. are typical features used by the participants. This shared task was repeated by SemEval in 2016 and 2017 as SemEval 2016 task 3 and SemEval 2017 task 3. The best system in SemEval 2015/2016/2017 are the JAIST [39], KeLP [33, 40], and Beihang-MSRA [41].

In contrast to feature engineering methods, deep learning based methods learn features automatically by end-to-end training, greatly reducing the needs of feature engineering. Some of these methods are summarized in Table 2.

The model presented in [13] uses two convolutional neural networks (CNNs) to capture the similarity between the questions and answers, and based on it, label the answer. In [42], a convolutional sentence model is proposed to identify the answer content of a question. Wang and Nyberg [14] present a method that successfully employs recurrent neural networks (RNNs) for this task.

In addition to modeling the similarity of the answer and its question, context modeling is also considered in some recent studies. [10] and [43] propose models in which the labels of the previous and next answers are

considered as context information. These methods outperform their counterparts which do not consider context information.

Attention is another method used for answer selection. Authors in [9] proposed an attentive deep neural network which employs attention mechanism besides CNN and LSTM networks for answer selection in CQA. In [44], a network called Question Condensing is proposed. In this method, which is based on the question's subject-body relationship, the question's subject is considered as the main part and the question's body is aggregated with it based on their similarity and disparity. Joint modeling of users, questions, and answers is proposed in [11], in which a hybrid attention mechanism is used to model question-answer pairs. User information is also considered in answer classification in this model. In [12], an advanced deep neural network is proposed that leverages text categorization to improve the performance of question-answer relevance classification. Also, external knowledge is used to capture important entities in questions and answers. A hierarchical attentional model named KHAAS is proposed in [45] for answer selection in CQA.

Recently, various attention models based on the Transformer architecture are proposed for learning sentence representation [47]. Also, some models are introduced with Transformer network as their encoders or decoders [48, 49]. BERT [50] and RoBERTA [51] as contextualized word embeddings are widely used nowadays. BERT outperformed the previous state-of-the-art results significantly, for question answering in the Stanford question answering dataset (SQuAD) by fine-tuning the pre-trained model [46]. In [52], authors propose the gated self-attention network along with transfer learning from a large-scale online corpus, and provide improvements in the TREC-QA [53] and WikiQA [54] datasets for the answer selection task. In [46], a model is presented with the Transformer encoder (CETE) for sentence similarity modeling. In this paper, by utilizing contextualized embeddings (BERT, ELMo, and RoBERTA [51]), two different approaches, namely, feature-based and fine-tuning-based, are presented. CETE model has achieved state-of-the-art performance in answer selection task in CQA and is our main baseline.

There are still some limitations in the aforementioned methods that make the answer selection in CQA a **challenge**. In feature engineering methods, the main problem is that extracting informative features is tedious and time-consuming. Also, they do not achieve high performance in most of the time. In the deep learning methods, the representations of the question-answer pair are learned independently which results in insufficient exploitation of the semantic correlation between them. Also, none of the existing methods have considered the question category as context information in question-answer representation. Furthermore, sometimes the named entities in the questions and answers are disregarded when learning the representations because they do not exist in the word embedding methods such as Glove or Word2vec.

Different from the aforementioned studies, in our proposed model, we **contribute** to use external background knowledge from KGs to capture the meaning of the question and answer words and the relation between them. We also consider the context in the representation which leads to having a more accurate representation and so better performance. Furthermore, we **contribute** to learning the joint representations of question-answer pair. This allows us to find compact representations of them in the latent space which benefits the semantic matching between question-answer sentences.

4. Proposed method

The main principle of this paper is to address the question-answer relevance classification in CQA by using KGs. In our proposed model, depicted in Figure 1, at the first step, the words in the question and the answer are disambiguated using WSD and leveraging external knowledge from a KG. By using the KG, the entities (especially named-entities) and the relations between them are captured. As we know, noisy information exists in the questions and answers, so in the next step, we employ an attention mechanism to extract the important information. Finally, to infer the label of question-answer relevance, we propose a classifier in a multi-task learning process with two separate VAEs for the question and answer. These VAEs help learning the class-specific representation.

Next, we elaborate on three key components of the model in more details: initial representation, attention, and multi-task learning. The main notations used in Figure 1 are summarized in Table 3 for clarity.

Table 2

Summarization of previous community question answering approaches.

Answer selection approaches	references	Methods
Feature-based	[36]	Used general tree matching methods based on tree edit distances
	[37]	Used logistic together with a tree kernel function and extracted features to learn the associations between the question/answer pair
	[38]	Used translation features, frequency features, and similarity features
Deep learning	[13]	Used two convolutional neural network (CNN) to capture the similarity between the question and the answer
	[14]	Used Recurrent Neural Networks (RNNs) and LSTM based model
	[10], [43]	Used CNNs for similarity matching and the label of previous and next answer for context modeling through LSTM
	[11]	Used joint modeling of users, questions, and answers and also, attention mechanism for modeling question-answer pair
	[45]	Used hierarchical attentional model and also, the knowledge from the knowledge base
	[46]	Integrates contextualized embeddings with the transformer encoder (CETE) for sentence similarity modeling

4.1. Initial representations

Some words may have different meanings in different contexts. Static word embedding methods, such as word2vec or Glove, do not address this issue and may lead to incorrect sentence representations. Furthermore, there are sometimes named entities in sentences not defined in the common word embedding vocabularies (such as “Armada” and “Infiniti FX35” in Table 1) and so, they are ignored in sentence representations. Considering these two problems, we propose to disambiguate each word of the question subject, question body, and answer body by leveraging KG. We also use the question category, as the context representative. In this disambiguation procedure, the meaning of each disambiguated word (including named entities) is captured through KG, and the relation between them is extracted. We use Babelfy, a unified graph-based approach to entity linking (EL) and word sense disambiguation (WSD) [55], for disambiguating the question and answer.

The Babelfy algorithm is a KG based model that requires the availability of a semantic network, such as BabelNet, which encodes structural and lexical information. In this semantic network, each vertex is an entity. The Babelfy algorithm has three main steps. At first, given a lexicalized semantic network, it assigns each vertex a semantic signature, i.e., a set of related vertices. For the relatedness notion in this step, the global structure of the semantic network is exploited and a more precise and higher coverage measure of relatedness is obtained. To address this issue, a structural weighting of the network’s edge is provided. Then for each vertex, a set of related vertices is created by using random walks with restart. In the second step, for a given text, by applying part-of-speech tagging and identifying all the textual fragments, it lists all possible meanings of the extracted fragments. Finally, by creating a graph-based semantic interpretation of the

Table 3

Notation list

Notation	Description
Q^{cat}	Question category
Q^{sub}	Question subject
Q^{body}	Question body, containing the details of the question
A^{body}	Answer body, containing the details of the answer
sub^{init}	Initial representation of the question subject
Q^{init}	Initial representation of the question
A^{init}	Initial representation of the answer
Q^{rep}	Attentional representation of the question and subject
A^{rep}	Attentional representation of the answer and subject
Z_a	The sampled latent feature vector of answer
Z_q	The sampled latent feature vector of question
y	Question-answer relevance label

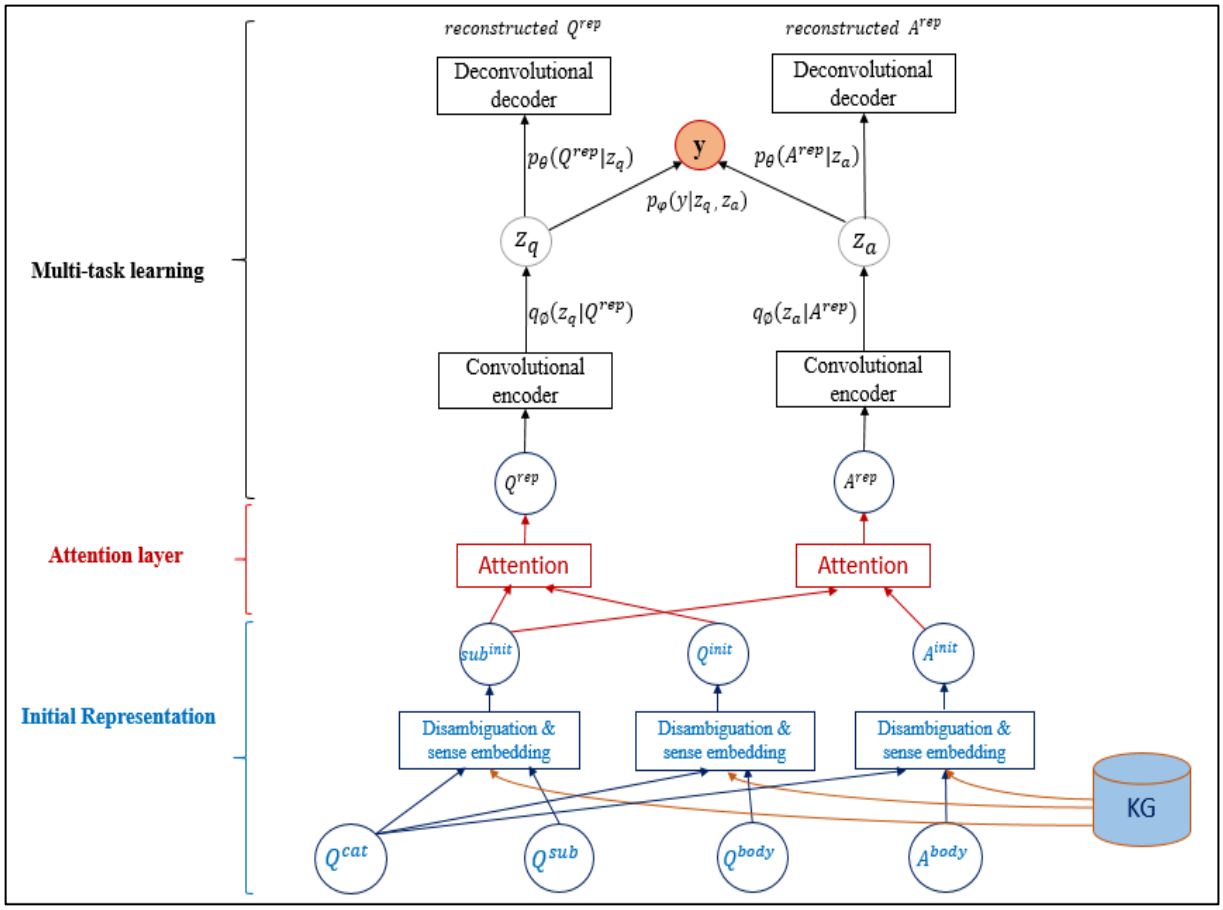

Fig. 1. Proposed model architecture. The inputs to this architecture are question's category, question's subject, question's body, answer's body and also, the KG. The output is question-answer relevance label.

whole text and using a previously-computed semantic signature, it selects the best candidate meaning for each fragment [55].

Based on this process, it can be said that Babelfy uses the context of a word to disambiguate it in a text. In our proposed method, to consider the question category as the contextual information, we simply concatenate it to the question subject, question body, and answer body. The concatenation of these three parts is considered as the input text.

To apply Babelfy to our problem, at its first step, we use BabelNet, the largest multilingual KG [2], as our lexicalized semantic network in the disambiguation procedure. The BabelNet, which contains both concepts and named entities as its vertices, is obtained from the automatic seamless integration of Wikipedia[1]and WordNet [56]. Then, independently of the input texts which are the question category, question subject, question body, and the answer body, we assign each vertex of the BabelNet a set of related vertices as its semantic signature. As said before, for the relatedness notion in this step, the global structure of the semantic network is exploited and a more precise and higher coverage measure of relatedness is obtained. This is done by using a structural weighting of the network's edge and after that, applying random walks with restart method. At the second step, given the input texts, all the textual fragments of them are identified, and at the final step, each fragment is disambiguated.

After disambiguating and capturing the correct sense in the current context from KG, we represent it using NASARI [57]. NASARI is a multilingual vector

[1] www.wikipedia.org

representation of word senses with high coverage, including both concepts and named entities [57]. More specifically, NASARI combines the structural knowledge from semantic networks with the statistical information derived from text corpora. This makes it possible to have an effective representation of millions of BabelNet synsets. The output of this step is the initial representation of the question subject, question body, and answer denoted as sub^{init}, Q^{init}, and A^{init}, respectively, in Figure 1.

4.2. Attention layer

The problem of redundancy and noise is prevalent in CQA [58]. On the other hand, the question subject summarizes the main points of the question and so can be used to extract useful information from the question and answer.

In order to reduce the impact of redundancy and noise, we use the representation of the question subject, sub^{init} in Figure 1, as the attention source to capture the important and useful information of the question and answer. Q^{rep} and A^{rep} are the outputs, which are the attentional representations of the question and answer, respectively. By defining w_i^q and w_i^a as the *i*-th word of the question and answer, respectively, Q^{rep} and A^{rep} are computed as follow:

$$\alpha_i^q = \frac{\exp(\rho([w_i^q\ ;\ sub^{init}]))}{\sum_{j=1}^{m} exp(\rho([w_j^q\ ;\ sub^{init}]))} \tag{3}$$

$$Q^{rep} = \sum_{i=1}^{m} \alpha_i^q . w_i^q \tag{4}$$

$$\alpha_i^a = \frac{\exp(\rho([w_i^a\ ;\ sub^{init}]))}{\sum_{j=1}^{l} exp(\rho([w_j^a\ ;\ sub^{init}]))} \tag{5}$$

$$A^{rep} = \sum_{i=1}^{l} \alpha_i^a . w_i^a \tag{6}$$

Where α_i^q and α_i^a indicates the importance of *i*-th word in the question and answer, respectively. Also, *m* is the length of question and *l* is the length of answer. ρ is the attention function and is computed as follow:

$$\rho([x\ ;\ y]) = U_d^T \tanh(W_d[x; y]) \tag{7}$$

Where U_d and W_d are projection parameters to be learned.

4.3. Multi-task learning

The multi-task learning module in Figure 1 is based on Siamese architecture [59]. Siamese neural architecture first appeared in vision (face recognition [60]). It has recently been extensively studied to learn representations of sentences and to predict similarity or entailment relation between sentence pairs as an end-to-end differentiable task [61-64].

Our model consists of deconvolutional-based twin networks. This proposed model is used for question-answer relevance extraction by employing the discriminative information encoded by the encoder network.

As shown in Figure 1, Q^{rep} and A^{rep}, the question and answer representations, are fed into separate VAEs. The encoder, i.e., a convolutional network, starting encodes the representation to the latent code z. Then the decoder, i.e., a deconvolutional network, starting by the latent code z, tries to arrive at the initial representation. These two VAEs are trained with shared weights.

To infer the label of the question-answer relevance, two latent features are sampled from the inference network, as z_q and z_a, and after concatenation, are fed into a classifier in a multi-task learning process with the two VAEs. The classifier is an MLP network. It generates the probability for each label ("good", "bad", and "potentially useful"), to model the conditional distribution $p_\varphi(y|z_q, z_a)$ with parameters φ.

To balance between maximizing the variational lower bound and minimizing the classifier loss, the model training objective is defined as follow:

$$\begin{aligned} L^{labeled} = \ & \alpha L_{classifier}\left(\varphi;\ z_a, z_q, y\right) \\ & - L_{VAE}(\theta, \emptyset; a) \\ & - L_{VAE}(\theta, \emptyset; q) \end{aligned} \tag{8}$$

Here, α is an annealing parameter between 0 to 1 (treated as a hyper-parameter), balancing the importance of the classifier loss. φ represents the classifier parameters. By changing the value of α, the learned latent variable can gradually focus only on retraining those features useful for answer classification.

5. Experimental results and analysis

In this section, we demonstrate the implementation details and analysis of our proposed framework and the comparison of experimental results.

5.1. Data

We conduct experiments on three widely used CQA datasets, SemEval-2015 Task 3[1] [6], SemEval-2016 Task 3[2] [19], and SemEval-2017 Task 3[3] [20], which contain real data from the QatarLiving forum. This forum is organized as a set of independent question-comment threads. Table 4 shows the statistics of these three datasets. For SemEval-2017 dataset, the training set is exactly the same as SemEval-2016, but the test set does not contains the ‘‘Potentially Useful’’ class.

Each question in the datasets consists of a short title or subject and a detailed description or body. Questions are followed by a list of comments (or answers), each of which is classified in one of three categories: ‘‘Definitely Relevant’’ (Good), ‘‘Potentially Useful’’ (Potential), or ‘‘Bad’’ (bad, dialog, non−English, other). “Good” label indicates that the answer is relevant to the question and answers it, even though it might be a wrong answer. “Potential” indicates that the answer contain potentially useful information about the question, and “Bad” indicates that the answer is irrelevant or useless. Besides three-class classification experiments, we also conducted experiments for two-class classification. Similar to the previous work, for two-class classification, we merge “Potentially Useful” and “Bad” labels to one label, “Bad”, in our experiments.

Table 4

Statistics of SemEval 2015, 2016, and 2017 datasets

Statistics		Number of questions	Number of answers
SemEval 2015	Train	2600	16541
	Dev	300	1654
	Test	329	1976
SemEval 2016	Train	4879	36198
	Dev	244	2440
	Test	327	3270
SemEval 2017	Train	4879	36198
	Dev	244	2440
	Test	293	2930

5.2. Baselines

In the experiments, we compare our proposed method with several baselines:

- **JAIST** [39]: this method, which had the best performance in SemEval-2015, investigates various features. SVM classifier is then used to predict the question-answer relation.
- **KeLP** [33]: It uses three kinds of features, including linguistic similarities between texts, syntactic trees, and task-specific information. This model was the winner of the SemEval-2016 and SemEval-2017 Task 3.
- **CNN** [65]**:** this model is a basic Siamese model with CNNs as encoder.
- **BiLSTM-attention** [8]: A biLSTM network for building the embeddings of question and answer followed by an attention mechanism are used to learn the question and answer representations.
- **CNN**-LSTM-**CRF** [10]: This model is a hierarchical architecture combining CNN, biLSTM, and CRF to model the context information, including content correlation and label dependency.
- **RCNN** [43]: In this model, a CNN is used to capture the semantic matching between the question and answer and an RNN is used for capturing the semantic correlations embedded in the sequence of answers.
- **Question Condensing** [44]: In this model, the question subject is considered as the main source and the information in the question body is aggregated based on that.
- **MKMIA-CQA** [12]: This model is a multi-task network that uses interactive attention and external knowledge to classify the answer in CQA. The knowledge base used in this model is a subset of Freebase[4] (FB5M3).
- **KHAAS** [45]: This model is a hierarchical attentional model that exploits the knowledge in the knowledge base for answer selection in CQA. The knowledge base used in this model is Freebase for the English dataset.
- **UIA-LSTM-CNN** [11]: This model calculates inter and intra sentence attentions between

[1] http://alt.qcri.org/semeval2015/task3/index.php?id=data-and-tools

[2] https://alt.qcri.org/semeval2016/task3/index.php?id=data-and-tools

[3] http://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools

[4] http://www.freebase.com/

questions and answers. It also exploits the user information.

- **CETE** [46]: In this model, contextualized word embeddings with the transformer encoder are utilized for sentence similarity modeling in answer selection in CQA.

5.3. Implementation details

As mentioned before, we use BabelNet as our KG, which contains both concepts and named entities. Then NASARI is used for capturing the embedding of each disambiguated word (sense). The max length is set to 50 and the vocab size is set to 5000. For the training procedure, we use a convolutional encoder with three layers followed by a deconvolutional encoder with the same number of layers. We try these hidden sizes: 100, 300, and 500. The weight parameters are randomly sampled from a uniform distribution $U(-0.01,0.01)$, and the bias parameters are set to zero. The batch size is set to 128.

The model is trained using RMSProp optimizer [66]. Dropout is employed on the latent variable layer with the dropout rate of 0.5.

5.4. Quantitative evaluation

For the answer selection task, the standard metrics used in previous work for benchmarking are macro-averaged F1 and Mean Average Precision (MAP). We measure the performance using these metrics on three datasets: SemEval 2015, SemEval 2016, and SemEval 2017.

Table 5, Table 6, and Table 7 show the performance comparison of our proposed model with other baselines for three-class classification, on SemEval 2015, SemEval 2016, and SemEval 2017, respectively. It should be noted that in three-class classification, for the baselines in which their results are reported for two-class classification, we modified their source code for three-class classification (KeLP [33], Question Condensing [44], MKMIA-CQA [12], KHAAS [45], UIA-LSTM-CNN [11], and CETE [46]). Also, CNN [65] and BiLSTM-attention [8] models, which their original implementations are for datasets other than ours, were re-implemented for SemEval datasets. Table 8, Table 9, and Table 10 are for two-class classification results.

Table 5
Quantitative evaluation results on SemEval 2015 for three-class classification

Method	F1 score	MAP
JAIST	57.19	66.23
KeLP	59.71	68.42
CNN	54.42	64.09
BiLSTM-attention	58.63	67.86
CNN-LSTM-CRF	58.96	68.03
RCNN	58.77	69.15
Question Condensing	60.63	71.45
MKMIA-CQA	61.93	72.07
KHAAS	57.81	69.74
UIA-LSTM-CNN	61.37	69.89
CETE	69.08	78.63
Proposed model	**74.91***(p-value = 0.03)	**85.41***(p-value = 0.02)

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

Table 6
Quantitative evaluation results on SemEval 2016 for three-class classification

Method	F1 score	MAP
JAIST	46.65	57.89
KeLP	44.67	54.38
CNN	43.57	55.21
BiLSTM-attention	49.28	60.08
CNN-LSTM-CRF	50.08	61.57
RCNN	49.82	61.98
Question Condensing	52.47	61.49
MKMIA-CQA	56.68	64.25
KHAAS	53.06	61.05
UIA-LSTM-CNN	56.87	64.17
CETE	65.39	72.32
Proposed model	**68.79***(p-value = 0.04)	**77.48***(p-value = 0.03)

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

Table 7

Quantitative evaluation results on SemEval 2017 for three-class classification

Method	F1 score	MAP
JAIST	48.51	58.89
KeLP	49.83	60.24
CNN	50.02	61.97
BiLSTM-attention	52.97	63.09
CNN-LSTM-CRF	56.32	68.47
RCNN	55.84	68.54
Question Condensing	58.72	70.18
MKMIA-CQA	59.91	70.57
KHAAS	56.06	68.16
UIA-LSTM-CNN	59.24	70.74
CETE	68.12	79.07
Proposed model	**70.43***(p-value = 0.04)	**81.83***(p-value = 0.04)

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

Table 8
Quantitative evaluation results on SemEval 2015 for two-class classification

Method	F1 score	MAP
JAIST	78.96	86.2
KeLP	80.73	89.43
CNN	76.92	84.24
BiLSTM-attention	79.09	85.56
CNN-LSTM-CRF	81.33	89.91
RCNN	81.52	87.47
Question Condensing	83.91	90.1
MKMIA-CQA	84.85	91.01
KHAAS	81.85	88.76
UIA-LSTM-CNN	85.37	90.45
CETE	86.34	94.7
Proposed method	**88.45*** (p-value = 0.04)	**95.63**

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

Table 9

Quantitative evaluation results on SemEval 2016 for two-class classification

Method	F1 score	MAP
JAIST	62.16	77.56
KeLP	64.36	79.19
CNN	64.92	76.21
BiLSTM-attention	69.82	77.31
CNN-LSTM-CRF	70.04	77.45
RCNN	71.84	78.44
Question Condensing	73.75	80.98
MKMIA-CQA	74.35	81.27
KHAAS	71.02	79.12
UIA-LSTM-CNN	73.91	80.57
CETE	79.48	88.8
Proposed method	**81.67*** (p-value = 0.03)	**90.02*** (p-value = 0.03)

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

Table 10

Quantitative evaluation results on SemEval 2017 for two-class classification

Method	F1 score	MAP
JAIST	68.04	87.24
KeLP	69.87	88.43
CNN	72.14	86.21
BiLSTM-attention	74.82	88.05
CNN-LSTM-CRF	77.04	87.66
RCNN	76.33	87.80
Question Condensing	78.11	88.51
MKMIA-CQA	79.78	88.93
KHAAS	75.64	81.25
UIA-LSTM-CNN	77.43	87.92
CETE	85.45	94.3
Proposed method	**87.92*** (p-value = 0.03)	**94.9**

* Numbers mean that improvement from our model is statistically significant over the baseline methods (t-test, p-value < 0.05).

As it is obvious in Table 5, Table 6, and Table 7, for three-class classification, our proposed model outperforms other baselines. It beats the state-of-the-art method, CETE, in F1 by about 6%, 4% and 3% for SemEval 2015, 2016, and 2017. Similarly, it outperforms the MAP results of the CETE in all three datasets. The p-values for these differences are less than 0.05, indicating that the improvements are statistically significant. It should be noted that considering the "potentially useful" label as a separate class, instead of merging it with the "bad" class and having a three-class classification model, needs the model to be more accurate and it is the superiority of our proposed approach over the competitors.

Similarly, for two-class classification, as indicated in Table 8, Table 9, and Table 10, our proposed method outperforms the baseline methods in F1 and MAP. Except the MAP of 2015 and 2017 datasets, the increase in other values is statistically significant. These results show that our model's improvements are not dependent on the number of classes only. The experimental results prove our hypothesis about the obtained representations for the question and answer. In other words, the results indicate that these representations are informative in predicting the relevance of the questions and answers.

5.5. Ablation study

To analyze the effects of each component of our model, we also report the ablation test of our model in terms of discarding external knowledge from KG (w/o KG), attention on the subject (w/o AS), question category (w/o category), deconvolutional decoder (w/o deconv), and VAE (w/o VAE). For w/o KG, we simply use word embedding instead of sense embedding in the initial representation. For w/o category, we disambiguate each question and answer themselves, without considering category information. Also, for w/o deconv and w/o VAE, we use LSTM for the decoder and simple autoencoder instead of VAE, respectively. The ablation results are summarized in Table 11 and Table 12 for the three datasets.

We also analyze the performance of the proposed method by starting from the baseline model, and incrementally add one component at a time. The baseline model is the vanilla version in which there are only two parallel autoencoders to obtain question and answer representations. Then, the concatenation of these representations are sent to an MLP to extract question-answer relevance. Table 13 and Table 14 demonstrate the results.

Generally, all five factors contribute to the results of our proposed model. It is obvious that F1 and MAP decrease sharply by discarding KG. This is within our expectation since using KG enriches overall text representation, by making it possible to consider all entities (especially named entities), the context, and focusing on useful information. In addition, deconvolutional VAE also has a great contribution. This verifies that using deconvolutional decoder results to have a more informative representation. Not surprisingly, combining all components achieves the best performance.

5.6. Parameter analysis

In this subsection, we analyze the model sensitivity to hyper-parameters specific to CNN: window size, stride, and filter-size (number of filters). Figure 2 and Figure 3 indicate the change of macro-averaged F1 values for different values of window size and filter-size, respectively.

For stride value, we observe that when it is 4 or greater, the system gets close to fully fit the training data (over fitting). The best value for stride is 2 for both datasets.

As it is obvious in Figure 2 and Figure 3, the best value obtained for macro-averaged F1 is 74.91 for SemEval 2015, 68.79 for SemEval 2016, and 70.43 for SemEval 2017, which are for window size, stride, and filter-size equal to 4, 2, and 300, respectively.

Table 11

Ablation test of the proposed model on SemEval 2015, SemEval 2016, and SemEval 2017 for three-class classification

Method	SemEval 2015		SemEval 2016		SemEval 2017	
	F1 score	MAP	F1 score	MAP	F1 score	MAP
Proposed model	**74.91**	**85.41**	**68.79**	**77.48**	**70.43**	**81.83**
w/o KG	69.21	80.52	62.16	69.56	64.51	75.42
w/o AS	74.67	84.09	67.91	75.12	69.93	78.97
w/o category	72.03	82.73	65.96	74.38	68.12	78.09
w/o deconv	71.26	82.11	64.89	73.47	66.87	77.07
w/o VAE	70.1	81.92	64.07	73.29	65.72	75.49

Table 12

Ablation test of the proposed model on SemEval 2015, SemEval 2016, and SemEval 2017 for two-class classification

Method	SemEval 2015		SemEval 2016		SemEval 2017	
	F1 score	MAP	F1 score	MAP	F1 score	MAP
Proposed model	**88.45**	**95.63**	**81.67**	**90.02**	**87.92**	**94.9**
w/o KG	83.67	89.87	75.01	81.89	80.59	88.35
w/o AS	87.91	93.27	79.90	87.91	86.01	90.83
w/o category	85.81	91.54	77.21	86.01	84.71	89.22
w/o deconv	86.23	92.48	78.12	86.45	83.12	90.07
w/o VAE	84.67	92.01	78.06	85.74	81.04	87.91

Table 13

Analysis of each component impact on SemEval 2015, SemEval 2016, and SemEval 2017 for three-class classification

Method	SemEval 2015		SemEval 2016		SemEval 2017	
	F1 score	MAP	F1 score	MAP	F1 score	MAP
Baseline	63.16	73.02	57.12	68.44	57.34	66.12
Add KG	67.42	78.12	61.89	72.01	61.99	72.67
Add category	69.35	79.01	63.54	72.67	64.41	75.01
Add AS	70.1	81.92	64.07	73.29	65.72	75.49
Add VAE	71.26	82.11	64.89	73.47	66.87	77.07
Add deconv	74.91	85.41	68.79	77.48	70.43	81.83

Table 14

Analysis of each component impact on SemEval 2015, SemEval 2016, and SemEval 2017 for two-class classification

Method	SemEval 2015		SemEval 2016		SemEval 2017	
	F1 score	MAP	F1 score	MAP	F1 score	MAP
Baseline	75.58	82.48	69.34	80.67	72.55	77.91
Add KG	80.33	87.76	73.23	85.60	77.90	84.80
Add category	83.01	89.11	75.88	86.87	82.15	87.81
Add AS	83.90	91.01	76.91	87.24	82.44	88.01
Add VAE	85.12	91.77	77.83	87.80	83.54	89.91
Add deconv	88.45	95.63	81.67	90.02	87.92	94.9

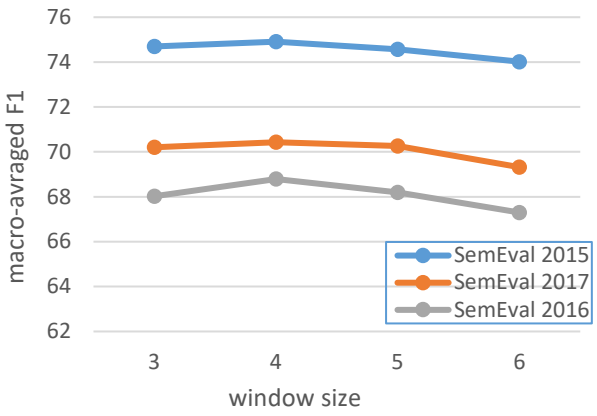

Fig. 2. The influence of window size on model performance

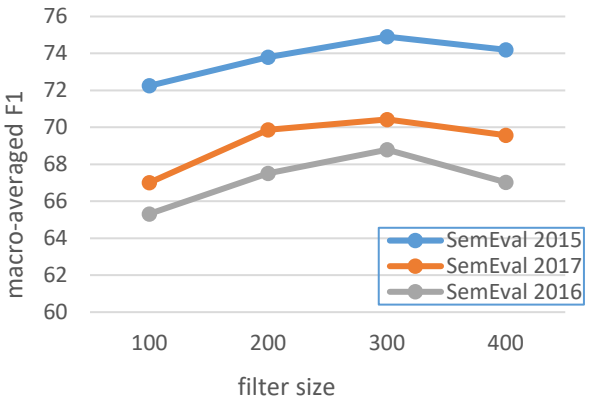

Fig. 3. The influence of filter size on model performance

6. Conclusion

In this article, we proposed a new model based on KGs for answer selection in community question answering forums. In the proposed architecture, external background knowledge is used to capture entity mentions and their relations in questions and answers. Also, by using the question category, a context-aware representation is generated for the question and answer. The model is trained in a multi-task learning procedure, in which there are two variational autoencoders in combination with a classifier to capture the semantic relatedness of the question and answer.

Quantitatively, the experimental results demonstrated that our model outperformed all existing baselines. We also conducted an ablation analysis to show the effectiveness of each component of the proposed model. The results confirm the choices we had in our architecture design because all of them, especially the KG integration, contribute positively.

Acknowledgement:

This work is based upon research funded by Iran National Science Foundation (INSF) under project No. 4002438

References

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*, ed: Springer, 2007, pp. 722-735. doi: 10.1007/978-3-540-76298-0_52

[2] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artifical Intellignece*, 2012, pp. 217-250. doi: 10.1016/j.artint.2012.07.001

[3] A. Kumar, D. Kawahara, and S. Kurohashi, "Knowledge-Enriched Two-Layered Attention Network for Sentiment Analysis," in *NAACL-HLT (2)*, 2018.doi: 10.18653/v1/N18-2041

[4] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu, "Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1006-1014.doi: 10.1145/3394486.3403143

[5] N. Zhang, S. Deng, Z. Sun, G. Wang, X. Chen, W. Zhang, *et al.*, "Long-tail Relation Extraction via Knowledge Graph Embeddings and Graph Convolution Networks," in *NAACL-HLT (1)*, 2019. doi: 10.18653/v1/N19-1306

[6] P. Nakov, L. Màrquez, W. Magdy, A. Moschitti, J. Glass, and B. Randeree, "Semeval-2015 task 3: Answer selection in community

question answering," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 269-281. doi: 10.18653/v1/S15-2047

[7] M. Nicosia, S. Filice, A. Barrón-Cedeno, I. Saleh, H. Mubarak, W. Gao, *et al.*, "QCRI: Answer selection for community question answering-experiments for Arabic and English," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 203-209. doi: 10.18653/v1/S15-2036

[8] L. Zhenzhen, J. Huang, Zh. Zhou, H. Zhang, Sh. Chang, and Zh. Huang. "LSTM-based deep learning models for answer ranking." In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pp. 90-97. IEEE, 2016. doi: 10.1109/DSC.2016.37

[9] Y. Xiang, Q. Chen, X. Wang, and Y. Qin, "Answer selection in community question answering via attentive neural networks," *IEEE Signal Processing Letters,* vol. 24, pp. 505-509, 2017. doi: 10.1109/LSP.2017.2673123.

[10] Zh. Xiaoqiang, B. Hu, Q. Chen, and X. Wang. "Recurrent convolutional neural network for answer selection in community question answering." *Neurocomputing* 274 (2018): 8-18. doi: 10.1016/j.neucom.2016.07.082

[11] J. Wen, H. Tu, X. Cheng, R. Xie, and W. Yin, "Joint modeling of users, questions and answers for answer selection in CQA," *Expert Systems with Applications,* vol. 118, pp. 563-572, 2019. doi: 10.1016/j.eswa.2018.10.038.

[12] M. Yang, W. Tu, Q. Qu, W. Zhou, Q. Liu, and J. Zhu, "Advanced community question answering by leveraging external knowledge and multi-task learning," *Knowledge-Based Systems,* vol. 171, pp. 106-119, 2019. doi: 10.1016/j.knosys.2019.02.006.

[13] Ph. Ngoc-Quan, G. Kruszewski, and G. Boleda. "Convolutional neural network language models." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1153-1162. 2016.doi: 10.18653/v1/D16-1123

[14] D. Wang and E. Nyberg, "A long short-term memory model for answer sentence selection in question answering," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 707-712. doi: 10.3115/v1/P15-2116

[15] T. Mikolov, K. Chen, G. Corrado, J. Dean, L. Sutskever, and G. Zweig, "word2vec," *URL https://code. google. com/p/word2vec,* vol. 22, 2013. doi: 10.1017/S1351324916000334

[16] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543. doi: 10.3115/v1/D14-1162

[17] S.R. Bowman, G. Angeli, C. Potts, and C.D. Manning, A large annotated corpus for learning natural language inference. doi: 10.18653/v1/D15-1075

[18] Y. Zhang, D. Shen, G. Wang, Z. Gan, R. Henao, and L. Carin, "Deconvolutional paragraph representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4169-4179. doi: 10.1109/TCYB.2020.2973300

[19] P. Nakov, L. Màrquez, A. Moschitti, W. Magdy, H. Mubarak, A. A. Freihat, *et al.*, "SemEval-2016 Task 3: Community Question Answering," Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 525–545, 2016.doi: 10.18653/v1/S16-1083

[20] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, *et al.*, "SemEval-2017 task 3: Community question answering," Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 27-48, 2017.doi: 10.18653/v1/S17-2003

[21] Gu, J., Lu, Z., Li, H. and Li, V.O., 2016, August. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1631-1640). doi: 10.18653/v1/P16-1154.

[22] Eikema, B. and Aziz, W., 2019, August. Auto-Encoding Variational Neural Machine Translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (pp. 124-141). doi: 10.18653/v1/W19-4315.

[23] S. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating Sentences from a Continuous Space," in *Proceedings of the Twentieth Conference on Computational Natural Language Learning (CoNLL)*. 2016. doi: 10.18653/v1/K16-1002

[24] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2010*, 2010.doi:10.1109/CVPR.2010.5539957

[25] F. Chen and Y. Huang, "Knowledge-enhanced neural networks for sentiment analysis of Chinese reviews," *Neurocomputing*, vol. 368, pp. 51-58, 2019. doi: 10.1016/j.neucom.2019.08.054.

[26] Y. Cao, X. Wang, X. He, Z. Hu, and T.-S. Chua, "Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences," in *The world wide web conference*, 2019, pp. 151-161. doi: 10.1145/3308558.3313705

[27] M. Asgari-Bidhendi, A. Hadian, and B. Minaei-Bidgoli, "Farsbase: The persian knowledge graph," *Semantic Web*, vol. 10, pp. 1169-1196, 2019. doi: 1570-0844/0-1900.

[28] X. Huang, J. Zhang, D. Li, and P. Li, "Knowledge graph embedding based question answering," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 105-113. doi: 10.1145/3289600.3290956

[29] R. Wang, M. Wang, J. Liu, W. Chen, M. Cochez, and S. Decker, "Leveraging knowledge graph embeddings for natural language question answering," in *International Conference on Database Systems for Advanced Applications*, 2019, pp. 659-675. doi: 10.1007/978-3-030-18576-3_39

[30] Z. Liu, K. Li, and D. Qu, "Knowledge graph based question routing for community question answering," in *International Conference on Neural Information Processing*, 2017, pp. 721-730. doi: 10.1007/978-3-319-70139-4

[31] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Transactions on Neural Networks and Learning Systems*, 2021. doi: 10.1109/TNNLS.2021.3070843

[32] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. Chiang, "Quality-aware collaborative question answering: methods and evaluation," in *Proceedings of the second ACM international conference on web search and data mining*, 2009, pp. 142-151. doi: 10.1145/1498759.1498820

[33] S. Filice, D. Croce, A. Moschitti, and R. Basili, "Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 1116-1123. doi: 10.18653/v1/S16-1172.

[34] K. Tymoshenko and A. Moschitti, "Assessing the impact of syntactic and semantic structures for answer passages reranking," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 1451-1460.doi: 10.1145/2806416.2806490

[35] Y. Hou, C. Tan, X. Wang, Y. Zhang, J. Xu, and Q. Chen, "HITSZ-ICRC: Exploiting classification approach for answer selection in community question answering," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 196-202.doi: 10.18653/v1/S15-2035

[36] H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua, "Question answering passage retrieval using dependency relations," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 400-407. doi: 10.1145/1076034.1076103

[37] M. Heilman and N. A. Smith, "Tree edit models for recognizing textual entailments, paraphrases, and answers to questions," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 1011-1019. doi: 10.18653/v1/N22-4401.

[38] V. Suzan, H. Halteren, D. Theijssen, S. Raaijmakers, and L. Boves. "Learning to rank for why-question answering." *Information Retrieval* 14, no. 2 (2011): 107-132. doi: 10.1007/s10791-010-9136-6.

[39] Q. H. Tran, V. Tran, T. Vu, M. Nguyen, and S. B. Pham, "JAIST: Combining multiple features for answer selection in community question answering," in *Proceedings of the 9th*

International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 215-219. doi: 10.18653/v1/S15-2038

[40] S. Filice, G. Da San Martino, and A. Moschitti, "Kelp at semeval-2017 task 3: Learning pairwise patterns in community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 326-333.doi: 10.18653/v1/S17-2053

[41] W. Feng, Y. Wu, W. Wu, Z. Li, and M. Zhou, "Beihang-msra at semeval-2017 task 3: A ranking system with neural matching features for community question answering," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 280-286.doi: 10.18653/v1/S17-2045

[42] Sh. Taihua, X. Kui, P. Zhang, and H. Chen. "Collaborative learning for answer selection in question answering." *IEEE Access* 7 (2018): 7337-7347. doi: 10.1109/ACCESS.2018.2890102.

[43] X. Zhou, B. Hu, Q. Chen, and X. Wang, "Recurrent convolutional neural network for answer selection in community question answering," *Neurocomputing*, vol. 274, pp. 8-18, 2018. doi: 10.1016/j.neucom.2016.07.082.

[44] W. Wu, S. Xu, and W. Houfeng, "Question condensing networks for answer selection in community question answering," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1746-1755.doi: 10.18653/v1/P18-1162

[45] M. Yang, L. Chen, Z. Lyu, J. Liu, Y. Shen, and Q. Wu, "Hierarchical fusion of common sense knowledge and classifier decisions for answer selection in community question answering," *Neural Networks*, vol. 132, pp. 53-65, 2020. doi: 10.1016/j.neunet.2020.08.005.

[46] M. T. R. Laskar, X. Huang, and E. Hoque, "Contextualized Embeddings based Transformer Encoder for Sentence Similarity Modeling in Answer Selection Task," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 5505-5514. doi: 10.1145/3209978.3210019.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008. doi: 10.1101/2021.01.31.428935.

[48] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, *et al.*, "Universal sentence encoder for english," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp.169-174. doi: 10.18653/v1/D18-2029

[49] W. Pengwei, L. Wei, Yong Cao, J. Xie, and Z. Nie. "Large-scale unsupervised pre-training for end-to-end spoken language understanding." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7999-8003. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053163.

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2018, pp.4171-4186. doi: 10.18653/v1/N19-1423

[51] A. Acheampong Francisca, N. Henry, and W. Chen. "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition." In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 117-121. IEEE, 2020. doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[52] H. Weiyi, Q. Qu, and M. Yang. "Interactive knowledge-enhanced attention network for answer selection." *Neural Computing and Applications* 32, no. 15 (2020): 11343-11359. doi: 10.1007/s00521-019-04630-x.

[53] Ch. Qin, Q. Hu, J. Xiangji Huang, L. He, and W. An. "Enhancing recurrent neural networks with positional attention for question answering." In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 993-996. 2017. doi: 10.1145/3077136.3080699.

[54] Y. Yang, W.-t. Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in *Proceedings of the 2015 conference on empirical methods in natural language*

processing, 2015, pp. 2013-2018.doi: 10.18653/v1/D15-1237

[55] A. Moro, A. Raganato, and R. Navigli, "Entity linking meets word sense disambiguation: a unified approach," *Transactions of the Association for Computational Linguistics,* vol. 2, pp. 231-244, 2014. doi: 10.1162/tacl_a_00179.

[56] F. Christiane. "WordNet." In *Theory and applications of ontology: computer applications*, pp. 231-243. Springer, Dordrecht, 2010. doi: 10.1007/978-90-481-8847-5_10.

[57] J. Camacho-Collados, M. T. Pilehvar, and R. Navigli, "Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities," *Artificial Intelligence,* vol. 240, pp. 36-64, 2016. doi: 10.1016/j.artint.2016.07.005.

[58] H. Jun, Sh. Qian, Q. Fang, and Ch. Xu. "Attentive interactive convolutional matching for community question answering in social multimedia." In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 456-464. 2018. doi: 10.1145/3240508.3240626.

[59] Ch. Davide. "Siamese neural networks: An overview." *Artificial Neural Networks* (2021): 73-94. doi: 10.1007/978-1-0716-0826-5_3.

[60] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 539-546.doi: 10.1109/CVPR.2005.202

[61] A. Ion, and P. Malakasiotis. "A survey of paraphrasing and textual entailment methods." *Journal of Artificial Intelligence Research* 38 (2010): 135-187. doi: 10.1613/jair.2985

[62] M. Seyed Vahid, M. Joodaki, M. Maleki Kahaki, and M. Salimi Sartakhti. "A method Based on an Attention Mechanism to Measure the Similarity of two Sentences." In *2021 7th International Conference on Web Research (ICWR)*, pp. 238-242. IEEE, 2021. doi: 10.1109/ICWR51868.2021.9443135.

[63] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp.670-680. doi: 10.18653/v1/D17-1070

[64] Y. Homma, S. Sy, and C. Yeh, "Detecting duplicate questions with deep learning," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS*, 2016, pp.25964-25975. doi: 10.1109/ACCESS.2020.2968391

[65] R. Jinfeng, H. He, and J. Lin. "Experiments with convolutional neural network models for answer selection." In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1217-1220. 2017. doi: 10.1145/3077136.3080648.

[66] Z. Fangyu, L. Shen, Z. Jie, W. Zhang, and W. Liu. "A sufficient condition for convergences of adam and rmsprop." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 11119. 2019. doi: 10.1109/CVPR.2019.01138.