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

Golshan Afzali Boroujeni[a], Heshaam Faili[a,*]
[a] *School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran;* golshan.afzali@ut.ac.ir

Abstract. With the increasing popularity of knowledge graphs (KGs), many applications such as sentiment analysis, trend prediction, and question answering use KG for matching the entities mentioned in the text to entities in the KG. Despite the usefulness of commonsense knowledge or factual background knowledge in the KGs, to the best of our knowledge, these KGs have been rarely used for 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 KG. Our method is a deep neural network based model that besides using KG, uses a latent-variable model for learning the representations of the question and answer, by jointly optimizing generative and discriminative objectives. Specifically, the proposed model leverages external background knowledge from KG to help identify entity mentions and their relations. It also uses the question category for producing a context-aware representation for each of the question and answer. Moreover, the model uses variational autoencoders (VAE) in a multi-task learning process with a classifier to produce a class-specific representation for each answer. The experimental results on three widely used datasets demonstrate that the proposed method significantly outperforms all existing models in this field.

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

1. Introduction

Knowledge graphs (KGs), such as DBpedia [1], BabelNet [2], etc., are multi-relational graphs consisting of entities and relationships among them. Many applications such as sentiment analysis, recommender systems, relation extraction, and question answering use KG to link the entities mentioned in the text to entities in the KG.

On the other hand, community question answering (CQA) forums, such as Stack Overflow and Yahoo! Answer, which are very popular nowadays, provide a new opportunity for users to share knowledge. Due to the few restrictions in these communities, anyone can freely ask any question, and also, a question is answered by one or more members. Unfortunately, there is often no evaluation of the given answers; it means one has to go through all possible answers for assessing them, which is exhausting and time-consuming, especially for the answers which are lengthy and have low readability. So, it is essential to propose a method for automatically identifying the best answers for each question.

The main difficulty lies in how to bridge the semantic gaps 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. One important feature of CQA is that in these communities, each question has at least two parts: 1) question subject that summarizes the question and in fact, contains the main information of the question; and 2) question body, that describes the question in details, and commonly has useless or noisy parts which do not provide useful information. Furthermore, most of the questions and answers in these forums are often lengthy, informal, and contain abbreviations and grammatical mistakes. A typical question example and two of its answers from the SemEval 2015 dataset are shown in Table 1.

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

Early works in this area used feature-based methods for explicitly modeling the semantic relation between the question and answer [3, 4]. With great advances in deep learning neural networks, considerable recent researches have applied deep learning based methods to answer classification in question answering communities [5-9]. These methods typically use a Convolutional Neural Network (CNN) [10] or Long Short term Memory (LSTM) [11] network for matching the question and answer, usually addressed as a classification problem. 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 knowledge 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 used in recent deep neural networks in CQA. The knowledge of these KGs provides rich information about entities, specially named entities, and relations between them, and therefore, is helpful for representing the question and answer in CQA. Considering the example in Table 1, named entities "Armada" and "Infiniti FX35" in the question and answer, do not exist in the available word embedding methods such as Word2vec [12] or Glove [13] and so, are out-of-vocabulary. Therefore, the conventional methods assign a negative score to the first answer because of not understanding these named entities and the relation between them; however, by using a comprehensive and well-designed KG like BabelNet, the model can correctly assign the correct label to answer due to the entities and facts exist in it.

Table 1

Example of one question and two of its answers from the SemEval 2015 dataset.

Question category	Qatar Living Lounge
Question subject	Nissan Offer
Question body	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	call them again and check how much is Safari or Infiniti FX35
Answer 2	take a guess !

– 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 in the current context can be extracted and a more accurate representation of the question and answer would be generated.

– The previous methods are unable to encoding all semantic information of the question and answer. Also, in [14] 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 should have two main properties; first, the representation should preserve the text's fundamental details. Second, each learned representation should 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; this makes the representations well suited for the question-answer relation extraction.

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 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 was first proposed in [15] for paragraph representation learning, the representations of the question and answer 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 procedure which is a multi-task learning process, the question-answer relevance label information is also considered in the representations learning; this makes it possible to produce class-specific representation.

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

– We leverage external knowledge from KG 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 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.
- We introduce a new architecture in which a classifier is in combination with the variational autoencoders to make each representation class-specific.
- Our proposed model achieves state-of-the-art performance on three CQA datasets, SemEval 2015, SemEval 2016, and SemEval 2017, outperforming all competitors in this manner, which is an indicator of the effectiveness of our proposed model.

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 for obtaining sentence representation are sequence-to-sequence models, due to their ability to leverage information from unlabeled data [16]. In these models, at first, an encoder encodes the input sentence x into a fixed-length vector z, and then, the output sequence is reconstructed from z through a decoder network. Specifically, in the autoencoder model, the encoder is a deterministic usually nonlinear 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 autoencoder (VAE) which was first introduced by [17] provides a probabilistic manner for describing an observation in latent space, instead of a vector.

In VAE, 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 follow [17]:

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

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

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

2.2. Challenges of VAE for text

Typically, the LSTM network is used as the decoder in VAEs for text generation [18]; but due to the recurrent nature of the LSTM, the decoder tends to ignore the information of the latent code; this is because of providing the ground-truth words of the previous time steps during training process; this prevents the encoded input to contain enough information about the input [18]. For this problem, we use a deconvolutional decoder network which is shown to have the best performance among the other methods. As said in [19], deconvolutional networks are typically used in deep learning networks for up-sampling fix-length latent representations usually made by a convolutional network.

3. Related works

In this section, we briefly review the related works on the applications of KG and the answer selection in CQA.

3.1. Applications of KG

Many applications such as sentiment analysis, recommender systems, relation extraction, and question answering use KG. In [20] a novel framework is proposed to model aspect-opinion pair identification and aspect-level sentiment classification which uses the knowledge from KG. Also, in [21] a two-layered attention network is proposed that leverages KG to improve sentiment prediction. In the work done in [22], a KG-based recommender system is proposed that transfers the relation information in KG to understand the reasons that a user likes an item. In [23] the KG is used to enhance the data representation in conversational recommender systems. The author in [24] proposed a supervised relation extraction method for long-tailed, imbalanced data by using the knowledge from KG. For the entity linking problem, in [25] a KG-based approach is presented for entity linking (EL) and word sense disambiguation (WSD), named Babelfy. Also, in [26] a method is proposed for entity linking that leverages KG for exploiting the sufficient context as a source of background knowledge. For the question answering problem, the authors in [27] used KG embedding for answering the questions, especially simple questions. The work done in [28] is also in the question answering field which leverages relation phrase dictionaries and KG embedding for answering the questions in natural language. In [29] a model is presented that uses the KG for question routing in CQA. In this model, topic representations with network structure are integrated into a unified KG question routing framework.

3.2. Answer selection in CQA

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

Feature-based methods which have a long research duration, 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. For example in [30], the authors use general tree matching methods based on tree edit distances. Later, authors in [31] employed a logistic together with a tree kernel function and extracted features to learn the associations between the question/answer pair. In [32] a wide range of feature types such as translation features, frequency features, and similarity features for answer ranking of non-factoid questions are considered. Also, methods presented in [3], [4], [33], [34], [35], and [36], are all in this category.

In 2015, SemEval organized a similar task titled "answer selection in community question answering". Thirteen teams participated in that challenge. The participant mainly focused on defining new features to capture the semantic similarity between the question and its answers. Word matching feature, special component feature, topic-modeling-based feature, non-textual feature, etc. were 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 includes the JAIST [37], KeLP model [34, 38], and Beihang-MSRA [39].

In contrast to feature engineering methods, deep learning based methods which have been widely used can learn to select features by end-to-end training which greatly reduce the needs of heavy feature engineering. The model presented in [10] uses two convolutional neural network (CNN) to capture the similarity between the question and the answer, and based on it, label the answer. In [40] a convolutional sentence model is proposed to identify the answer content of a question. Wang and Nyberg [11] presented a method that successfully employed Recurrent Neural Networks (RNNs) and proposed LSTM based model for this task. In addition to modeling the similarity of the answer and its question, context modeling is also considered in some recent studies. [7] and [41] proposed models which in them, in addition to the similarity of the answer and the question which is modeled by parallel CNNs, the label of previous and next answer is considered as context information and is modeled through LSTM networks. These two models achieve better results than methods which not consider context information in this field. Authors in [6] proposed an attentive deep neural network which employs attention mechanism besides CNN and LSTM network for answer selection in community question answering. In [42] 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 [8] 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 [9] an advanced deep neural network is proposed that leverage text categorization to improve the performance of question-answer relevance classification. Also, external knowledge is used to capture important entities in question and answer. A hierarchical attentional model named KHAAS is proposed in [43] for answer selection in CQA. This model exploits the knowledge from the knowledge base.

Recently, various attention models based on the transformer model were proposed for sentence representation [44]. Also, some models were introduced which used the encoder or the decoder of the transformer model for different NLP tasks [45, 46]. BERT [47] and ELMo [48] which are contextualized embeddings are widely used nowadays. BERT which is a transformer-based model showed state-of-the-art results for question answering in the Stanford Question Answering Dataset (SQuAD) by fine-tuning the pre-trained model [49]. In [50] authors proposed the gated self-attention network which was combined with BERT along with transfer learning from a large-scale online corpus and provided improvement in the TREC-QA [51] and WikiQA [52] datasets for the answer selection task. In [49], a model is presented which integrates contextualized embeddings with the transformer encoder (CETE) for sentence similarity modeling. In this paper by utilizing contextualized embeddings (BERT, ELMo, and RoBERTA [53]), 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 competitor.

There are many limitations w.r.t the aforementioned methods that have made the answer selection in CQA still 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 context in question-answer representation. Furthermore, there are named entities in the questions and answers which are not considered in the representations because of not existing in available 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 KG to capture the meaning of the question and answer words and the relation between them. Our **other contribution** is considering the context in the representation which leads to having a more accurate representation and so, better performance. Furthermore, we **contribute** to jointly learning the 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 KG. In our proposed model which is depicted in Figure 1, at the first step, the question and its answer are represented by using WSD and leveraging external background knowledge from KG. By using KG, the entities (especially named-entities) and the relations between them are captured. As we know, there is noisy information in both the question and answer which doesn't provide meaningful information. So, at the next step, we employ an attention mechanism to extract the important and useful information of the question and answer. Finally, to infer the label of question-answer relevance, we propose a procedure in which a classifier is in a multi-task learning process with two separate VAEs. These VAEs are for learning the class-specific representations for each of question and answer. Next, we elaborate on three key components of the model which are initial representation, attention, and multi-task learning, in detail. The main notations used in Figure 1 are summarized in Table 2 for clarity.

4.1. Initial representation

Some words may have different meanings in different contexts. Usual word embedding methods, such as word2vec or Glove, don't address this issue and may lead to the incorrect representation of the sentence. Furthermore, sometimes, there are named entities in the sentence which are not defined in the common word embedding methods (such as "Armada" and "Infiniti FX35" in Table 1) and so, they are ignored in sentence representation. Considering these two problems, we

Table 2

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

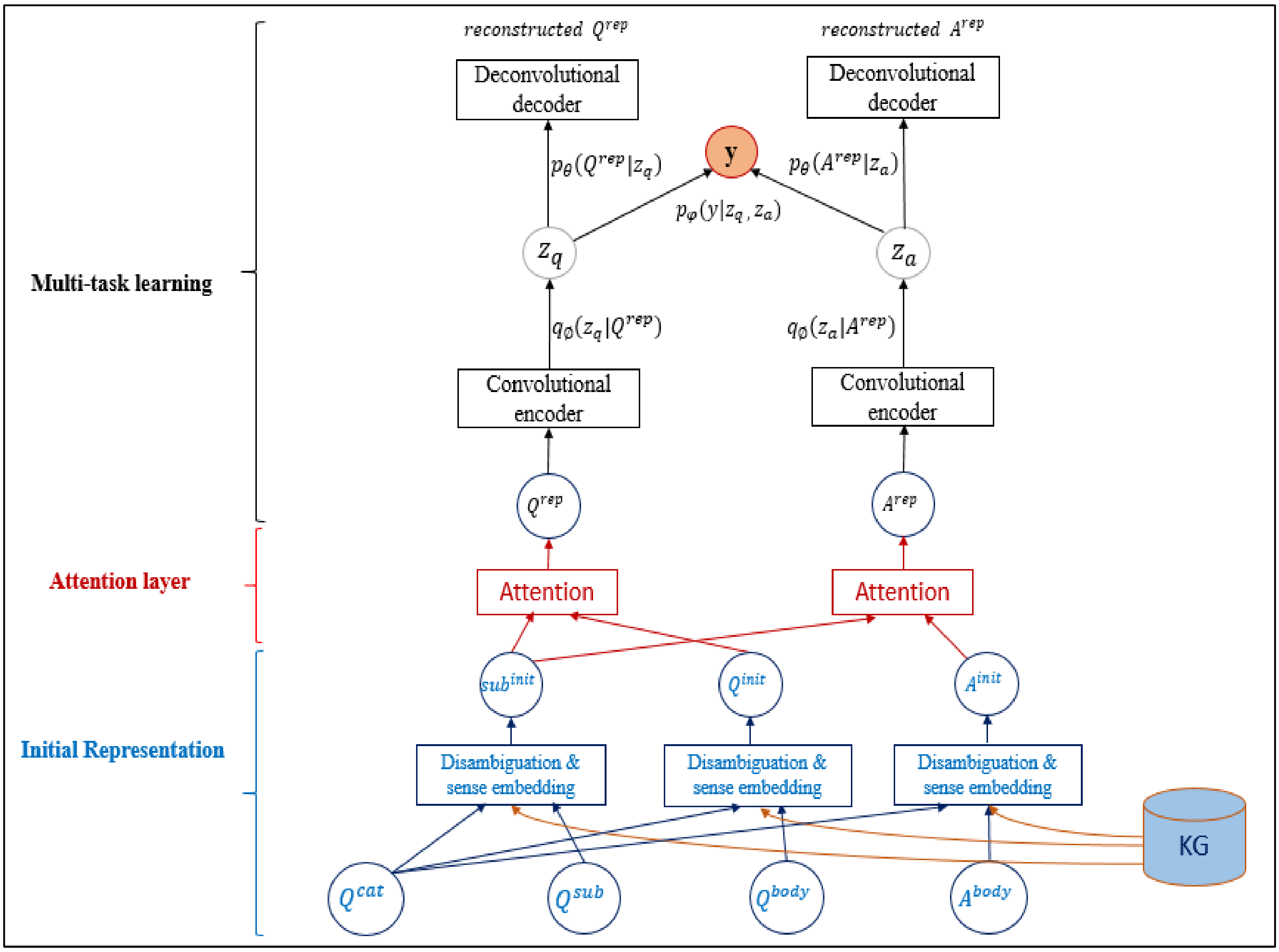

Fig. 1. Proposed model architecture

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 (include named entities) is captured through KG, and the relation between them is extracted. We use Babelfy, a unified graph-based approach to EL and WSD [25], for disambiguating the question and answer. The Babelfy algorithm which is a KG based model, has three main steps [25]: at first, given a lexicalized semantic network, it assigns each vertex, a semantic signature which is a set of related vertices. In this semantic network, each vertex is an entity. Then, for a given text, it lists all possible meaning of the extracted fragments from text according to the semantic network. Finally, by creating a graph-based semantic interpretation of the whole text and using a previously-computed semantic signature, it selects the best candidate meaning for each fragment [25]. Based on this process, it can be said that Babelfy uses each word's surrounding words to disambiguate it in a text. So, in our proposed method, for considering the question category as the context information, we simply concatenate it to question subject, question body, and answer body and apply Babelfy on them. We use BabelNet, the largest multilingual KG [2], as our lexicalized semantic network in the disambiguating 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 [54]. After disambiguating and capturing the correct sense in the current context from KG, we use NASARI [55] for each sense representation. NASARI is a multilingual vector representation, enables accurate representation of word sense with a high coverage, including both concepts and named entities [55]. The output of this step is the initial representation of the question subject, question body, and answer which are denoted as sub^{init}, Q^{init}, and Q^{init}, respectively, in Figure 1.

4.2. Attention layer

The problem of redundancy and noise is prevalent in community question answering [56]. There is noisy information in both the question and answer which doesn't provide meaningful information. 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 this noisy information, we use the representation of the question subject, which is denoted as sub^{init} in Figure 1, as the attention source to capture the important and useful information of the question and answer. As the output, Q^{rep} and A^{rep} will be produced which are attentional representations of the question and answer, respectively, and contain useful information.

4.3. Multi-task learning

The multi-task learning module which is shown in Figure 1, is based on Siamese architecture [57]. Siamese networks are a type of neural network that appeared in vision (face recognition [58]) and have recently been extensively studied to learn representations of sentences and predict similarity or entailment between pairs as an end-to-end differentiable task [59-62].

Our model consists of deconvolutional-based twin networks which are independent with shared parameters. 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}, which are question and answer representations, respectively, are fed into separate VAEs. The encoder which is a convolutional network, by starting with the representation, works upward and encodes it to the latent code z. Then the decoder which is a deconvolutional network, starts by latent code z, and by working downward, tries to arrive at the initial representation. These two VAEs are trained with shared weights for both encoder and decoder, to recover their corresponding input sentences.

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 to a classifier which is in a multi-task learning process with the two VAEs. The classifier which is an MLP network, outputs the probability for each label (good, bad, and potentially useful, in this task), to model the conditional distribution $p_{\varphi}(y|z_q, z_a)$ with parameter φ.

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

$$\begin{aligned} L^{labeled} = \alpha L_{classifier}\left(\varphi;\ z_a, z_q, y\right) \\ - L_{VAE}(\theta, \emptyset; a) \\ - L_{VAE}(\theta, \emptyset; q) \end{aligned} \tag{3}$$

In Eq. (3), α is an annealing parameter between 0 to 1 (treated as a hyper-parameter), balancing the importance of the classifier loss, and φ is the classifier parameter. By changing the value of α, the learned latent variable can gradually focus only on retraining those features that are 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[2] [3], SemEval-2016 Task 3[3] [63], and SemEval-2017 Task 3[4] [64], which contains real data from the QatarLiving forum. This

[1] www.wikipedia.org

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

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

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

forum is organized as a set of independent question-comment threads. Each question in the datasets consists of a short question title which is the question subject and a detailed question description, which is its body. Each question is followed by a list of comments (or answers), each of which is classified in one of three categories as ‘‘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 can answer it well; “Potential” indicates that the answer may contain useful information for the user about the question, and “Bad”, indicates that the answer is irrelevant or useless for the user. Table 3 shows the statistics of these three datasets.

5.2. Baseline methods

In the experiments, we compare our proposed method with several baseline methods which consist of:

- **JAIST** [37]: this method which had the best performance in SemEval-2015, investigated various features and, the SVM classifier was then used to predict the question-answer relation.
- **KeLP** [34]: It used 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 two CNN networks as encoder with shared weights and parameters.
- **BiLSTM-attention** [5]: A biLSTM network for building the embeddings of question and answer followed by an attention mechanism was used to learn the question and answer representations.
- **CNN-LSTM-CRF** [7]: This model is a hierarchy architecture combining CNN, biLSTM, and CRF to model the context information, including content correlation and label dependency.
- **RCNN** [41]: In this model convolutional neural network (CNN) is in combination with the recurrent neural network (RNN). CNN is used to capture both the semantic matching between question and answer and RNN is for capturing the semantic correlations embedded in the sequence of answers.
- **Question Condensing** [42]: In this model which uses deep learning based approach, the question subject is considered as the main part of it and the question body information is aggregated based on it.
- **MKMIA-CQA** [9]: 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[1] (FB5M3).
- **KHAAS** [43]: 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** [8]: This model calculates inter and intra sentence attention between question and answer. It also exploits the user information for answer selection.
- **CETE** [49]: In this model, contextualized word embeddings with the transformer encoder are utilized for sentence similarity modeling in answer selection in CQA.

Table 3

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

5.3. 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). For the training procedure, we use a 3-layer convolutional encoder followed by a 3-layer deconvolutional network as decoder. The array of hidden-size for trying is set to 100, 300, and 500.

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

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

5.4. Quantitative evaluation

For the answer classification task, the official scores are macro-averaged F1 and Mean Average Precision (MAP), which are reported in previous methods. Therefore, we make the comparison based on these measures and on three datasets, SemEval 2015, SemEval 2016, and SemEval 2017.

Table 4, Table 5, and Table 6 show the performance comparison of our proposed model with other baseline methods, on SemEval 2015, SemEval 2016, and SemEval 2017, respectively[1].

As it is obvious in Table 4, Table 5, and Table 6, our proposed model outperforms F1 of the state-of-the-art method (CETE) up to about 6% for SemEval 2015, about 4% for SemEval 2016, and about 3% for SemEval 2017. It also outperforms MAP of the state-of-the-art method (CETE) up to about 7% for SemEval 2015, about 6% for SemEval 2016, and about 3% for SemEval 2017. More specifically, incorporating the external commonsense knowledge from KG and using context information for initial representation, and then, training by convolutional-deconvolutional VAE (instead of common VAE), results in higher performance in comparison to other existing methods.

Table 4

Quantitative evaluation results on SemEval 2015

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***	**85.41***

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

Table 5

Quantitative evaluation results on SemEval 2016

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***	**77.48***

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

Table 6

Quantitative evaluation results on SemEval 2017

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***	**81.83***

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

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 so informative that they can pinpoint the relevance of the question and answer.

5.5. Ablation study

In order to analyze the effectiveness of each component of the proposed 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), respectively. For the model without external

[1] It should be noted that for the baseline methods in which their experiments were for two-class classification, we re-implemented them for three-class classification.

knowledge from KG, we simply use word embedding instead of sense embedding in the initial representation. For the model without category, we disambiguate each question and answer themselves, without considering category information. Also, for the model without deconvolutional decoder and the model without VAE, we use LSTM for the decoder and simple autoencoder instead of VAE, respectively.

The ablation results are summarized in Table 7, Table 8, and Table 9, for three experimental datasets. Generally, all five factors contribute to great improvement to the proposed model. In the results, it is obvious that F1 score 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 so, focusing on useful information. In addition, deconvolutional VAE also has a great contribution to the effectiveness of the proposed model. 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 which are 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 observed that for this value equals 4 and greater, the system has got close to fully fit the training data (over-fitting) and the best value for it was 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 7

Ablation test of the proposed model on SemEval 2015

Method	F1 score	MAP
Proposed model	**74.91**	**85.41**
w/o KG	69.21	80.52
w/o AS	74.67	84.09
w/o category	72.03	82.73
w/o deconv	67.41	78.11
w/o VAE	66.15	77.92

Table 8

Ablation test of the proposed model on SemEval 2016

Method	F1 score	MAP
Proposed model	**68.79**	**77.48**
w/o KG	62.16	69.56
w/o AS	67.91	75.12
w/o category	65.96	74.38
w/o deconv	61.89	72.47
w/o VAE	61.07	70.29

Table 9

Ablation test of the proposed model on SemEval 2017

Method	F1 score	MAP
Proposed model	**70.43**	**81.83**
w/o KG	64.51	75.42
w/o AS	69.93	78.97
w/o category	68.12	78.09
w/o deconv	62.87	73.17
w/o VAE	61.02	73.49

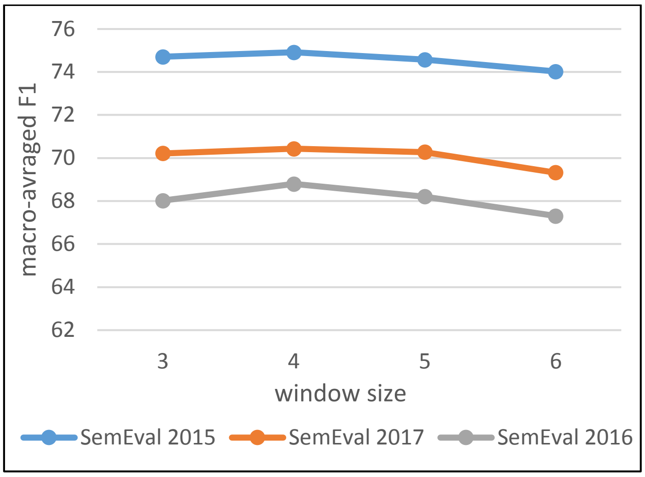

Fig. 2. The influence of window size on model performance

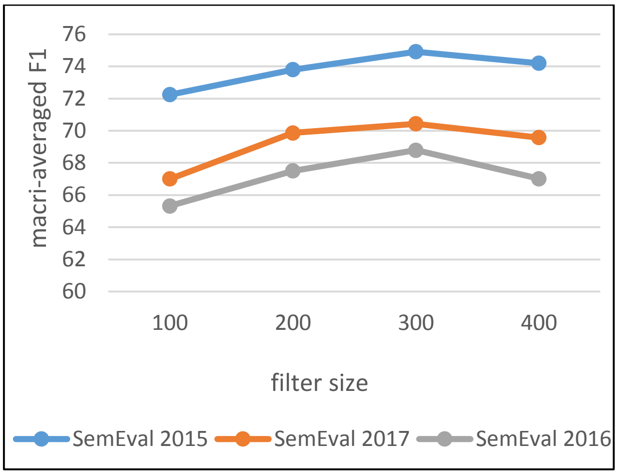

Fig. 3. The influence of filter size on model performance

6. Conclusion

In this paper, we proposed a new model based on KG for answer quality tagging in community question answering forums. In the proposed architecture, external background knowledge is used to capture entity mentions and their relations in each of the question and answer. Also, by using the question category, a context-aware representation is generated for the question and answer. Furthermore, 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 and then, classify their relevance. Quantitatively, the experimental results demonstrated that our model outperformed all the compared methods. We also conducted an ablation test to analyze the effectiveness of each component of the proposed model and the results showed that all the investigated factors, especially the KG, contributed a great improvement to the model.

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.

[2] R. Navigli and S. P. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 216-225.

[3] 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.

[4] 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.

[5] M. Tan, C. d. Santos, B. Xiang, and B. Zhou, "LSTM-based deep learning models for non-factoid answer selection," *arXiv preprint arXiv:1511.04108*, 2015.

[6] 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.

[7] Y. Xiang, X. Zhou, Q. Chen, Z. Zheng, B. Tang, X. Wang, *et al.*, "Incorporating label dependency for answer quality tagging in community question answering via CNN-LSTM-CRF," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1231-1241.

[8] 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.

[9] 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.

[10] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in neural information processing systems*, 2014, pp. 2042-2050.

[11] 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.

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

[13] 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.

[14] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[15] 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.

[16] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] 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.

[19] 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.

[20] F. Chen and Y. Huang, "Knowledge-enhanced neural networks for sentiment analysis of Chinese reviews," *Neurocomputing*, vol. 368, pp. 51-58, 2019.

[21] A. Kumar, D. Kawahara, and S. Kurohashi, "Knowledge-Enriched Two-Layered Attention Network for Sentiment Analysis," in *NAACL-HLT (2)*, 2018.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] I. O. Mulang, K. Singh, A. Vyas, S. Shekarpour, M.-E. Vidal, and S. Auer, "Encoding Knowledge Graph Entity Aliases in Attentive Neural Network for Wikidata Entity Linking," in *International Conference on Web Information Systems Engineering*, 2020, pp. 328-342.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] M. Surdeanu, M. Ciaramita, and H. Zaragoza, "Learning to rank answers on large online QA collections," in *proceedings of ACL-08: HLT*, 2008, pp. 719-727.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, "Deep learning for answer sentence selection," *arXiv preprint arXiv:1412.1632,* 2014.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175,* 2018.

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," ed, 2018.

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

[48] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365,* 2018.

[49] 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.

[50] T. Lai, Q. H. Tran, T. Bui, and D. Kihara, "A gated self-attention memory network for answer selection," *arXiv preprint arXiv:1909.09696,* 2019.

[51] M. Wang, N. A. Smith, and T. Mitamura, "What is the Jeopardy model? A quasi-synchronous grammar for QA," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 22-32.

[52] 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.

[53] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen*, et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692,* 2019.

[54] G. A. Miller, *WordNet: An electronic lexical database*: MIT press, 1998.

[55] 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.

[56] X. Zhang, S. Li, L. Sha, and H. Wang, "Attentive interactive neural networks for answer selection in community question answering," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[57] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737-744.

[58] 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.

[59] E. Dadashov, S. Sakshuwong, and K. Yu, "Quora Question Duplication."

[60] A. Sanborn and J. Skryzalin, "Deep learning for semantic similarity," *CS224d: Deep Learning for Natural Language Processing Stanford, CA, USA: Stanford University,* 2015.

[61] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364,* 2017.

[62] 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.

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

[64] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin*, et al.*, "SemEval-2017 task 3: Community question answering," *arXiv preprint arXiv:1912.00730,* 2019.

[65] R. Sequiera, G. Baruah, Z. Tu, S. Mohammed, J. Rao, H. Zhang*, et al.*, "Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering," *arXiv preprint arXiv:1707.07804,* 2017.

[66] M. C. Mukkamala and M. Hein, "Variants of rmsprop and adagrad with logarithmic regret bounds," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 2545-2553.