The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

CA-RNN: Using Context-Aligned Recurrent Neural Networks for Modeling Sentence Similarity

Qin Chen,[1] **Qinmin Hu,**[1] **Jimmy Xiangji Huang,**[2] **Liang He**[1]
[1]Department of Computer Science & Technology, East China Normal University, China
[2]Information Retrieval & Knowledge Management Research Lab, York University, Canada
qchen@ica.stc.sh.cn, {qmhu, lhe}@cs.ecnu.edu.cn, jhuang@yorku.ca

Abstract

The recurrent neural networks (RNNs) have shown good performance for sentence similarity modeling in recent years. Most RNNs focus on modeling the hidden states based on the current sentence, while the context information from the other sentence is not well investigated during the hidden state generation. In this paper, we propose a context-aligned RNN (CA-RNN) model, which incorporates the contextual information of the aligned words in a sentence pair for the inner hidden state generation. Specifically, we first perform word alignment detection to identify the aligned words in the two sentences. Then, we present a context alignment gating mechanism and embed it into our model to automatically absorb the aligned words' context for the hidden state update. Experiments on three benchmark datasets, namely TREC-QA and WikiQA for answer selection and MSRP for paraphrase identification, show the great advantages of our proposed model. In particular, we achieve the new state-of-the-art performance on TREC-QA and WikiQA. Furthermore, our model is comparable to if not better than the recent neural network based approaches on MSRP.

Introduction and Motivation

Sentence similarity modeling plays an important role in various Natural Language Processing (NLP) tasks, such as answer selection and paraphrase identification. For the answer selection task, all the candidate answers are ranked by the sentence similarity with the given question (Wang, Smith, and Mitamura 2007; Yang, Yih, and Meek 2015). As to paraphrase identification, sentence similarity is used to determine whether two sentences have the same meaning (Yin and Schütze 2015; He, Gimpel, and Lin 2015).

Most traditional methods rely on the feature engineering and linguistic tools, which are labour consuming and prone to the errors of NLP tools such as dependency parsing (Yih et al. 2013; Wan et al. 2006). Recently, the recurrent neural network (RNN) based approaches have attracted more attention due to the good performance and less human interventions. Specifically, a sequential hidden states were generated and aggregated for each sentence with RNN, and the similarity score was calculated according to the hidden representations (Mueller and Thyagarajan 2016). To capture the salient information for better sentence representations, the attention based RNN models that produce a weight for each hidden state start to arouse more interest. (Santos et al. 2016) proposed the attentive pooling networks, which incorporated the word-by-word interactions for the attentive sentence representations. In (Tan et al. 2015), the representation of the question was utilized for the attentive weight generation for the answer.

To the best of our knowledge, most attention based RNNs focus on generating the attentive weights after obtaining all the hidden states, while the contextual information from the other sentence is not well studied during the internal hidden state generation (Santos et al. 2016; Tan et al. 2015; Hermann et al. 2015). Noting that the inner activation units in RNN controls the information flow over a sentence, (Wang, Liu, and Zhao 2016) proposed an IARNN-GATE model, which incorporated the question representation into the active gates to influence the hidden state generation for the answer. However, it utilized all the information of the question sentence, which would bring noises if the current hidden state was not relevant to the question. To alleviate this problem, (Bahdanau, Cho, and Bengio 2014) presented an alignment model, which measured how well the input at each position matched the output for neural machine translation. Whereas, the alignment model in fact implemented the attention mechanism, and also leveraged all the input information to generate the output. Moreover, it is still unknown how to integrate the alignment information into RNN for sentence similarity modeling.

In this paper, we propose a context-aligned RNN (CA-RNN) model, where the context information of the aligned words is incorporated into the hidden state generation. To be specific, we first perform word alignment detection to identify the aligned words that are potentially relevant in a sentence pair. Then, a context alignment gating mechanism is presented and embedded into our model, which consists of two steps, namely relevance measurement and context absorption. The relevance measurement step aims to determine how much context can be absorbed, by measuring the relevance between the other sentence and the current hidden state. In the context absorption step, the context information of the aligned words in the other sentence is absorbed for the current hidden state generation. It is worth noting that the absorbed context will be naturally propagated across the whole sentence when modeling within the bidirectional RNN (Schuster and Paliwal 1997) framework. After that, the sentence representation is obtained by aggregating all the hidden states with the classical pooling (Wang and Nyberg 2015) or attention methods (Tan et al. 2015). We perform experiments on three datasets for two well-known sentence similarity tasks, namely TREC-QA (Wang, Smith, and Mitamura 2007) and WikiQA (Yang, Yih, and Meek 2015) for answer selection and MSRP (Dolan, Quirk, and Brockett 2004) for paraphrase identification as demonstrated in (Wang, Mi, and Ittycheriah 2016). The results show that our proposed CA-RNN model significantly outperforms the classical RNN model, by utilizing the context alignment information for sentence similarity modeling. Furthermore, compared with the recent progress in answer selection, we achieve the new state-of-the-art performance on TREC-QA and WikiQA. Regarding to paraphrase identification, our model is also comparable to if not better than the recent neural approaches on MSRP.

The main contributions of our work are as follows: (1) we propose a new context-aligned RNN model, where the contexts of the aligned words in two sentences are well utilized for better hidden state generation; (2) a context alignment gating mechanism is presented and nicely embedded into our model, which can automatically absorb the relevant context and reduce the noise for generating a specific hidden state; (3) we conduct elaborate analyses of the experimental results on two sentence similarity tasks, which provides a better understanding of the effectiveness of our model.

Related Work

Sentence similarity is a fundamental problem in many NLP tasks, such as information retrieval (Huang and Hu 2009; Wang et al. 2017), question answering (Yih et al. 2013; An et al. 2017) and paraphrase identification (Wan et al. 2006; Ji and Eisenstein 2013). In this paper, we focus on the latter two tasks as demonstrated in (Wang, Mi, and Ittycheriah 2016). Most previous work relies on feature engineering. (Yih et al. 2013) utilized the WordNet based semantic features to enhance lexical features for question answering. (Wan et al. 2006) showed that the dependency-based features were particularly useful for classifying cases of false paraphrase in the Microsoft Research Paraphrase Corpus (Dolan, Quirk, and Brockett 2004). (Heilman and Smith 2010) introduced the tree edit models and used the dependency parse trees for modeling sentence pairs. However, these methods are labor consuming due to the excessive dependence on the handcraft features.

Recently, there have been many studies about using the deep neural networks for sentence similarity modeling. (Zhao et al. 2017) and (Fang et al. 2016) applied the long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) based RNN model to obtain the semantic relevance between question-answer pairs for the community-based question answering. To capture the salient information for better sentence representations, the attention mechanism was introduced into the neural networks (Wang, Liu, and Zhao 2016; Santos et al. 2016; Chen et al. 2017). (Zhang et al. 2017) proposed an attentive interactive neural network, which focused on the interactions between text segments for answer selection. In addition, the interactions in sentence-level or word-level are incorporated for the attentive weight generation within the RNN framework. In (Tan et al. 2015), the attentive weights for an answer sentence relied on the interactions with the question sentence. In (Santos et al. 2016), the word-by-word interactions were utilized for the attentive sentence representations. Whereas, these attention based RNN models mainly focus on the attentive weight generation after obtaining all the hidden states, while the information interactions are not well investigated during the hidden state generation. (Wang, Liu, and Zhao 2016) proposed an IARNN-GATE model, where the question information was added to the active gates in RNN to influence the hidden state generation for answers. However, all the information of the question sentence is utilized in their model, which brings some noise for modeling the hidden states that are not relevant to the question.

To reduce the noise introduced into the RNN model, (Bahdanau, Cho, and Bengio 2014) proposed an alignment model, which measured how well the input at each position matched the output at the current position for neural machine translation. However, their alignment model in fact implements the attention mechanism, and still leverages the weighted sum of all the inputs to generate each output. In contrast, (Wang, Smith, and Mitamura 2007) utilized the syntactic alignment based features for answer selection, where each word in the candidate answer was softly aligned with a word in the question. (Wang and Ittycheriah 2015) proposed a word alignment based method, which found the best word alignment between an input question and a candidate one to answer the Frequently Asked Questions (FAQ). Despite the effectiveness of alignment, the contextual information of the aligned words has not well been investigated within the RNN framework. In this paper, we propose a context-aligned RNN model that incorporates the contextual information of the aligned words in two sentences for the inner hidden state generation. In this way, only the related context is used for sentence modeling, which is more accurate by our intuition.

Context-Aligned Recurrent Neural Networks

In this section, we will introduce our context-aligned RNN (CA-RNN) model for sentence similarity modeling in detail. For ease of understanding, we first give a brief introduction of the traditional RNN model as well as some notations used in this paper.

Given a sentence pair as $X = x_1, x_2, ..., x_m$ and $Y = y_1, y_2, ..., y_n$, we let $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_j \in \mathbb{R}^d$ denote the embedded representations of the words x_i and y_j respectively. The traditional RNN model models each sentence separately. For example, for the sentence as Y, we can obtain a sequential hidden states as $\mathbf{h}_1^y, \mathbf{h}_2^y, ..., \mathbf{h}_n^y$, and each $\mathbf{h}_j^y \in \mathbb{R}^k$ can be formulated as:

$$\mathbf{h}_j^y = f(\mathbf{y}_j, \mathbf{h}_{j-1}^y) \tag{1}$$

where f can be defined with the long short-term memory (LSTM) model or the gated recurrent unit (GRU), and $\mathbf{h}_j^y$ contains the context information from the first word to the

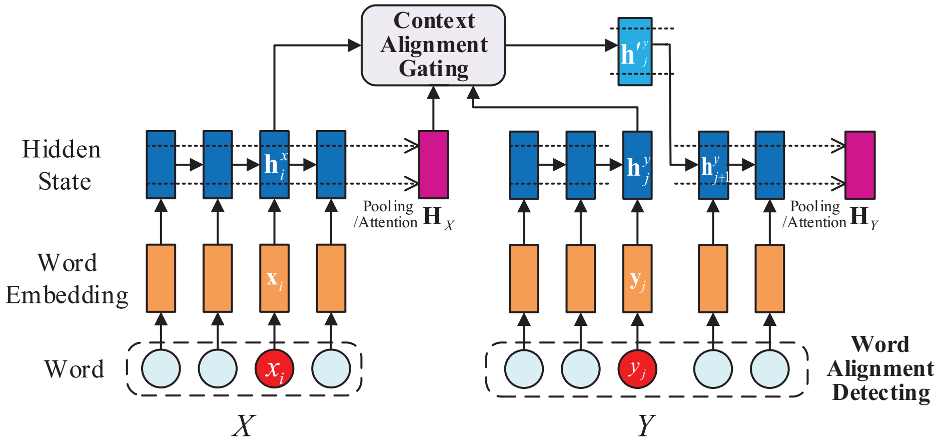


Figure 1: Framework of CA-RNN for sentence similarity modeling. Red circles denote the aligned words in the two sentences.

current one (Jozefowicz, Zaremba, and Sutskever 2015). Then, each sentence is represented by the pooling (Wang and Nyberg 2015) or attention (Tan et al. 2015) method over the hidden states, and the similarity score is calculated according to the two sentence representations ($\mathbf{H}_X$ and $\mathbf{H}_Y$).

The Framework of CA-RNN

It is notable that the hidden states in traditional RNN only depend on the current sentence, while the information interactions between the two sentences are neglected. With the assumption that the contextual information of the aligned words can facilitate sentence similarity modeling, we propose a context-aligned RNN (CA-RNN) model, which automatically absorbs the context of the aligned words for the hidden state generation.

The framework of our proposed CA-RNN model is presented in Figure 1. Compared with the traditional RNN model which models each sentence independently, our model can better capture the internal interactions of the aligned words in two sentences. Specifically, we first perform word alignment detection in the word level. With this step, the aligned words that are potentially relevant in the two sentences will be discovered and retained. Then, a context alignment gating mechanism is presented and embedded into our model, which controls the information flow between the two sentences by absorbing the context of the aligned words for hidden state generation. As shown in Figure 1, if the word y_j in sentence Y aligns with the word x_i in sentence X, the originally generated hidden state (i.e., $\mathbf{h}^y_j$) by RNN will be updated with the context information (i.e., $\mathbf{h}^x_i$) of the aligned word via context alignment gating. The representation of sentence X is also an input of the gating to help determine how much contextual information to be absorbed. We will give a detailed description of each step in the following sections.

Word Alignment Detecting

Word alignment aims at indicating the related words in a parallel text, and has been extensively studied in statistical machine translation (Och and Ney 2000; Bahdanau, Cho, and Bengio 2014). Recently, it has attracted more attention for finding the semantically related words such as synonyms in many NLP tasks (Van der Plas and Tiedemann 2006; Wu and Zhou 2003). Since the main concern of this paper is how to utilize the context information of the aligned words for better sentence similarity modeling, we mainly explore two simple but effective approaches for word alignment detecting, namely the overlap-based and semantic-based.

Overlap-based For many sentence similarity tasks, there are usually some lexical overlaps in the sentence pair. For example, given a question like "*Who was president of the United States in 1922?*", the words as "*president*" and "*1922*" often appear frequently in the candidate answers. Thus, one direct approach is to align the same words that occur in the two sentences. In other words, the lexically overlapped words are deemed to be aligned in the two sentences.

Semantic-based In some cases, there are few lexical overlaps in the sentence pairs. Instead, the semantically related words, such as synonyms (e.g., "*United States*" and "*USA*", "*say*" and "*state*"), are used for expressing the same meaning. Therefore, we adopt an additional semantic-based approach to find more aligned words in the two sentences. Specifically, the monolingual word aligner[1] algorithm is utilized, which is based on the Stanford Core NLP tools[2] and exploits the word semantics to make alignment decisions.

It is worth noting that the above two approaches for word alignment detecting can be replaced by other alternatives presented in the literature (Och and Ney 2000; Liu, Liu, and Lin 2005). With the detected aligned words, the aligned information can be further utilized for sentence similarity modeling with our proposed CA-RNN model.

Context Alignment Gating

Intuitively, the aligned words in a relevant sentence pair usually have the similar context. While for the irrelevant pair,

[1] https://github.com/ma-sultan/monolingual-word-aligner

[2] https://stanfordnlp.github.io/CoreNLP/

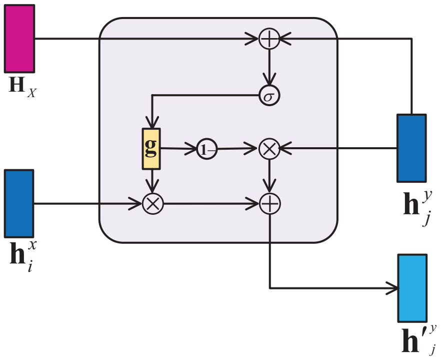


Figure 2: Context alignment gating.

the contexts of the aligned words will be probably different. Noting that the hidden state in RNN contains the contextual information from the first word to the current one (Jozefowicz, Zaremba, and Sutskever 2015), we are motivated to decrease the distance between the hidden states of the aligned words in a relevant sentence pair, and increase the distance for an irrelevant pair. In particular, we present a context alignment gating mechanism for our model, which adapts the hidden state generation for the aligned words and well boosts the context interactions in the hidden space.

The details of our context alignment gating are illustrated in Figure 2. Different from the traditional RNN model, our gating mechanism makes full use of the information in two sentences rather than a single one for the hidden state generation. To be specific, to model the hidden state of word y_j that is aligned with the word x_i, our context alignment gating mechanism mainly performs the following two steps, namely relevance measurement and context absorption.

Relevance Measurement The relevance measurement step measures the relevance between the representation of the sentence ($\mathbf{H}_X$) that the aligned word lies in and the hidden state ($\mathbf{h}_j^y$) corresponding to the current word, which serves as a good criteria to decide how much context information of the aligned words in the other sentence to be absorbed. More concretely, the relevance is formulated as:

$$\mathbf{g} = \sigma(\mathbf{W}_H\mathbf{H}_X + \mathbf{W}_h\mathbf{h}_j^y + \mathbf{b}) \tag{2}$$

where $\mathbf{W}_H$ and $\mathbf{W}_h$ are weight matrices, $\mathbf{b}$ is a bias vector, and $\sigma(\cdot)$ is an element-wise sigmoid function. It is worth noting that the obtained $\mathbf{g}$ is a vector, which reflects the relevance in each hidden dimension.

Context Absorption In the context absorption step, the original hidden state ($\mathbf{h}_j^y$) obtained by RNN will directly absorb the context information ($\mathbf{h}_i^x$) of the aligned word in the other sentence according to the measured relevance. As a result, a new hidden state will be generated, which is formulated as:

$$\mathbf{h}_j^{'y} = \mathbf{g} \odot \mathbf{h}_i^x + (1 - \mathbf{g}) \odot \mathbf{h}_j^y \tag{3}$$

where $\mathbf{g}$ is an interpolated relevance parameter obtained by Formula (2), $\odot$ denotes the element-wise multiplication, and $\mathbf{h}_j^{'y}$ is the new generated hidden state.

With the above two steps, the contextual gap between the aligned words in relevant sentence pairs will be narrowed down, while it is opposite for the irrelevant ones. Furthermore, the new generated hidden state will be passed to the next time step by sequential modeling of the sentence. Therefore, it will have a big influence on the whole sentence modeling. In other words, the relevant sentences will be closer in the hidden space, while the irrelevant ones will be farther away by the context interactions in our context alignment gating mechanism.

Experimental Setup

Datasets and Evaluation Metrics

To evaluate the effectiveness of our proposed model, we conduct experiments on two well-known sentence similarity tasks, namely answer selection and paraphrase identification as demonstrated in (Wang, Mi, and Ittycheriah 2016).

Answer Selection. Given a question and a list of candidate answers, the answer selection task is to rank the candidates according to their similarities with the question. Two widely used datasets, namely TREC-QA and WikiQA, are adopted in our experiments. TREC-QA was created by Wang et al. (Wang, Smith, and Mitamura 2007) based on the QA track (8-13) data of Text REtrieval Conference. WikiQA (Yang, Yih, and Meek 2015) is an open domain QA dataset in which all answers were collected from the Wikipedia. Both TREC-QA and WikiQA have the train, development and test sets, and each sample is labeled as 1 or 0 to indicate whether the candidate answer is right or wrong for a given question. The statistics of the datasets are presented in Table 1. The performance of answer selection is usually measured by the mean average precision (MAP) and mean reciprocal rank (MRR) (Santos et al. 2016; Wang, Liu, and Zhao 2016).

Table 1: Statistics of the datasets for answer selection. We remove all the questions with no right or wrong answers. "Avg QL" and "Avg AL" denote the average length of questions and answers.

Dataset		# of quetions	Avg QL	Avg AL
TREC-QA	Train	1162	7.57	23.21
	Dev	65	8.00	24.90
	Test	68	8.63	25.61
WikiQA	Train	873	7.16	25.29
	Dev	126	7.23	24.59
	Test	243	7.26	24.59

Paraphrase Identification. The paraphrase identification task can be treated as a binary classification problem, and the goal is to judge whether two sentences are paraphrases or not according to their similarity. We utilize the Microsoft Research Paraphrase corpus (MSRP) (Dolan, Quirk, and Brockett 2004) for experiment, which is constructed from a large corpus of temporally and topically clustered news articles. The MSRP dataset contains 4,076 sentence pairs in the training set, and 1,725 ones in the test set. Each sentence pair is labeled with 1 or 0 to indicate whether the two sentences are paraphrases or not. Since no development set is provided, we randomly select 100 positive pairs (labeled as 1) and 100 negative pairs (labeled as 0) from the training set as the development set. To evaluate the performance, two widely used metrics, namely accuracy (Acc) and F1 score are adopted (Yin and Schütze 2015; He, Gimpel, and Lin 2015).

Training

We use the bidirectional LSTM (BLSTM) (Graves and Schmidhuber 2005) model as the function in Formula (1) to obtain the original hidden states, which can effectively mitigate the gradient vanish problem (Tan et al. 2015). For the answer selection task, the context of the aligned words in the *question* is utilized for the new hidden state generation for the *answer*, since we usually select answers according to the question. As to paraphrase identification, the context-aligned information of the *left* sentence is used to model the *right* sentence for the model consistency. For ease of description, we call the above two alignment directions as **Q2A** and **L2R** respectively. Then, the widely used max pooling (Wang and Nyberg 2015) method and the recently proposed attention mechanism (Tan et al. 2015) are investigated to obtain the sentence representations. After that, we utilize the Manhattan distance similarity function with l_1 norm and restrict it to a range of $[0, 1]$ for sentence similarity calculation (Mueller and Thyagarajan 2016):

$$S(X, Y) = exp(-||\mathbf{H}_X - \mathbf{H}_Y||_1) \tag{4}$$

The predicted probability of a sentence pair labeled as 1 or 0 is defined according to the relevance score: $\hat{p(}c = 1|X, Y) = S(X, Y)$ and $\hat{p(}c = 0|X, Y) = 1 - S(X, Y)$.

For the answer selection task, the candidate answers are ranked by $\hat{p(}c = 1|X, Y)$. As to the paraphrase identification task, the predicted label is 1 when $\hat{p(}c = 1|X, Y) > \hat{p(}c = 0|X, Y)$ (i.e., $S(X, Y) > 0.5$). Otherwise, the predicted label is 0. For each sentence pair, the loss function is defined by the cross-entropy of the predicted and true label distributions for training:

$$L(X, Y; c) = -\sum_{j=0}^{C-1} p(c = j|X, Y) \log \hat{p(}c = j|X, Y) \tag{5}$$

where C is the number of classes, and $p(c = j|X, Y)$ is the gold probability of label c, which equals to 1 with ground truth and otherwise is 0.

Parameter Settings

We use different word embeddings for different tasks. Specifically, for the answer selection task, we use the 100-dimensional GloVe word vectors[3], which are trained based on the global word co-occurrence (Pennington, Socher, and Manning 2014). For the paraphrase identification task, we concatenate the GloVe vectors with the 25-dimensional PARAGRAM vectors[4] that are developed for paraphrase tasks (Wieting et al. 2015). The dimension of the hidden state is set to 50. We use the AdaDelta (Zeiler 2012) algorithm for parameter update when training. The optimal parameters are obtained based on the best performance on the development set, and then used for evaluation on the test set.

Results and Analyses

Effectiveness of CA-RNN

To investigate the effect of our CA-RNN model, the BLSTM based RNN model which does not involve any context alignment information is utilized for comparisons. Table 2 and Table 3 show the performance of various models for answer selection and paraphrase identification, where the superscripts as "O" and "S" denote the overlap-based and semantic-based alignment methods respectively. The best result obtained on each data set is marked in bold. It is observed that we achieve significant improvements over the classical BLSTM model on all datasets, by incorporating the context alignment information into the hidden state generation. The maximum improvement is up to 15.6% in terms of MAP on TREC-QA. Regarding to the various methods used for word alignment detection, the semantic-based method slightly outperforms the overlap-based one, since it can identify more semantic relevant words between two sentences.

It is also notable that the classical BLSTM model relies more on the attention mechanism to capture the salient information for sentence similarity modeling. However, the attention method mainly focuses on measuring the weight of each hidden state, while does not pay specifical attention to the surrounding context of the aligned words in a sentence pair. Moreover, the attentive weight is produced after obtaining all the hidden states, which neglects the internal interactions during hidden state generation. In contrast, our proposed CA-RNN model can explicitly capture the internal relations between two sentences, by incorporating the contextual information of the aligned words into hidden state generation. Therefore, our CA-RNN model integrated with the simple max pooling method can yield similar performance as the attention mechanism. In particular, we achieve the best performance on TREC-QA when integrated with max pooling.

Table 2: Performance of various models for answer selection on TREC-QA and WikiQA: (1) the superscripts "O" and "S" denote the overlap-based and semantic-based alignment methods; (2) "MAX" and "ATT" represent the max pooling and attention method respectively and (3) all improvements over the BLSTM model are **significant** according to the paired t-test at the 0.05 level.

	Model	**TREC-QA**		**WikiQA**	
		MAP	**MRR**	**MAP**	**MRR**
MAX	BLSTM	0.7117	0.8118	0.6761	0.6943
	CA-RNN^O	0.8205	0.8853	0.7211	0.7334
	CA-RNN^S	**0.8227**	**0.8886**	0.7226	0.7373
ATT	BLSTM	0.7528	0.8226	0.6979	0.7107
	CA-RNN^O	0.8115	0.8808	0.7282	0.7436
	CA-RNN^S	0.8159	0.8821	**0.7358**	**0.7450**

[3] http://nlp.stanford.edu/data/glove.6B.zip

[4] http://ttic.uchicago.edu/ wieting/

Table 3: Performance of various models for paraphrase identification on MSRP: (1) the superscripts “O” and “S” denote the overlap-based and semantic-based alignment methods; (2) “MAX” and “ATT” represent the max pooling and attention method respectively and (3) all improvements over the BLSTM model are **significant** according to the paired t-test at the 0.05 level.

	Model	MSRP	
		Acc	F1
MAX	BLSTM	73.8	81.9
	CA-RNN^{O}	75.6	82.2
	CA-RNN^{S}	75.7	82.8
ATT	BLSTM	74.4	82.0
	CA-RNN^{O}	76.5	83.3
	CA-RNN^{S}	**77.3**	**84.0**

Comparison with the Recent Progress

In addition to the classical BLSTM model, we compare our model with the recent progress in answer selection and paraphrase identification.

Results on TREC-QA and WikiQA Table 4 and Table 5 report the results from prior work on TREC-QA and WikiQA respectively. (Yin et al. 2015; Wang, Liu, and Zhao 2016; Santos et al. 2016; Chen et al. 2017) are the recent attention based models that focus on the attentive sentence representations. Since (Wang, Liu, and Zhao 2016) reported the results (marked with †) on the manually cleaned TREC-QA dataset with only 78 questions in the training set, we also present the corresponding results for a fair comparison in the last row. It is observed that our proposed model achieves the new state-of-the-art performance on both TREC-QA and WikiQA. Specifically, we outperform the best result on TREC-QA (full training set) with an absolute improvement of 0.02 in terms of MAP. For WikiQA, our model slightly outperforms the state-of-the-art inner attention based RNN models (Wang, Liu, and Zhao 2016). Whereas, the inner attention based RNN models did not perform very well on the cleaned TREC-QA dataset that has less training samples. In contrast, our model is more robust by directly absorbing the contextual information of the aligned words in questions for the hidden state generation in answers. Moreover, compared with the method which integrated the word alignment features into the learning-to-rank framework (Wang and Ittycheriah 2015), our model is also much more effective and does not rely on the laboursome feature engineering.

Table 4: Performance comparisons on TREC-QA. Note that the work marked with † used the manually cleaned training set which had less questions.

System	MAP	MRR
(Wang and Nyberg 2015)	0.7134	0.7913
(Wang, Liu, and Zhao 2016) †	0.7369	0.8208
(Wang and Ittycheriah 2015)	0.7460	0.8200
(Santos et al. 2016)	0.7530	0.8511
(Wang, Mi, and Ittycheriah 2016)	0.7714	0.8447
(Chen et al. 2017)	0.7814	0.8513
(Rao, He, and Lin 2016)	0.8010	0.8770
CA-RNN^{S} (MAX)	**0.8227**	**0.8886**
CA-RNN^{S} (MAX) †	0.8131	0.8818

Table 5: Performance comparisons on WikiQA

System	MAP	MRR
(Yang, Yih, and Meek 2015)	0.6520	0.6652
(Santos et al. 2016)	0.6886	0.6957
(Yin et al. 2015)	0.6921	0.7108
(Rao, He, and Lin 2016)	0.7010	0.7180
(Wang, Mi, and Ittycheriah 2016)	0.7058	0.7226
(Chen et al. 2017)	0.7212	0.7312
(Wang, Liu, and Zhao 2016)	0.7341	0.7418
CA-RNN^{S} (ATT)	**0.7358**	**0.7450**

Results on MSRP The results from recent work on MSRP are summarized in Table 6, which can be divided into 2 groups: non-neural network (non-NN) based and neural network (NN) based. It is observed that (Ji and Eisenstein 2013) achieved the state-of-the-art performance with a non-NN based method that relied on various hand-crafted features. Recently, the NN based methods tend to outperform or rival many traditional methods. To be specific, (Yin and Schütze 2015) presented a convolutional neural network (CNN) based deep learning architecture, which modeled interaction features at multiple levels of granularity. However, their model relied much on the pretraining step. In (He, Gimpel, and Lin 2015), a similar model was proposed, which also used a CNN model for feature extraction at a multiplicity of perspectives. They achieved the best performance among the NN based methods by incorporating various components, such as the part-of-speech (POS) embeddings, the PARAGRAM (Para.) embeddings, multiple widths and similarity layers. We observe that our model is comparable to theirs that contains all the components except the POS embeddings. Furthermore, our model has less parameters to be tuned, while their model needs to consider the variables such as the window size and perspective granularity.

Table 6: Performance comparisons on MSRP. The results in the first group are non-NN based, and others are NN based.

System	Acc	F1
(Blacoe and Lapata 2012)	73.0	82.3
(Wan et al. 2006)	75.6	83.0
(Madnani, Tetreault, and Chodorow 2012)	77.4	84.1
(Ji and Eisenstein 2013)	**80.4**	**86.0**
(Hu et al. 2014)	69.9	80.9
(Socher et al. 2011)	76.8	83.6
(Yin and Schütze 2015) - without pretraining	72.5	81.4
(Yin and Schütze 2015) - with pretraining	78.1	84.4
(He, Gimpel, and Lin 2015) - without POS embeddings	77.8	N/A
(He, Gimpel, and Lin 2015) - with POS and Para. embeddings	78.6	84.7
CA-RNN^{S} (ATT)	77.3	84.0

Investigation of Context Alignment Directions

In the answer selection task, the question intuitively plays an initiative role, since we usually select answers according to the question. Thus, we apply the **Q2A** context alignment direction, which absorbs the context information of the aligned words in the *question* for the hidden state generation of the *answer* as demonstrated previously. As to paraphrase identification, it aims to determine whether two sentences have the same meaning, and the order of the two sentences does not influence the results theoretically. For the model consistency, we also utilize the aligned context of the *left* sentence to model the hidden states in the *right* one, namely **L2R**. For a better understanding, we investigate the influence of different context alignment directions in this section. In other words, the context of the aligned words in the *answer* (or *right*) sentence is utilized for hidden state modeling in the *question* (or *left*) sentence, namely **A2Q** (or **R2L**) in the answer selection (or paraphrase identification) task.

Figure 3 shows the performance of our CA-RNN^S model with max pooling and various alignment directions on each dataset. The results of other configurations such as CA-RNN^O with max pooling or attention are similar to Figure 3. It is observed that there is a great performance decline in answer selection with the A2Q context alignment direction. This corresponds to our previous intuition that we should model the candidate answer with the aligned contextual information from the question. Moreover, since not all candidate answers are relevant to the question, the noisy answer will lead to an inaccurate question modeling and decreases the performance based on the A2Q strategy. Regarding to the paraphrase identification task, the models with different context alignment directions (L2R and R2L) perform similarly, which is also in line with our previous intuition.

Case Studies

To have an insight of why our model is more effective, we randomly sample a similar and dissimilar sentence pair from the datasets, and draw the word-by-word similarity heatmaps based on CA-RNN^S and BLSTM in Figure 4 and Figure 5 respectively. The results of CA-RNN^O are similar to CA-RNN^S, and we do not present them here for the limited space. A deeper color indicates a larger similarity value. The aligned words are presented as tuples under each figure.

We observe that for the similar sentence pair, the number of similar words in our proposed model is larger than that in the BLSTM model. In addition, the aligned words in the two sentences, such as (“*could*”, “*could*”), (“*not*”, “*not*”) and (“*reached*”, “*reached*”), are more similar in our proposed model. Since the context information of the aligned words will be propagated to the whole sentence by sequential modeling, the similarity between the surroundings of the aligned words is also much higher based on our model. Moreover, the color of the cell in the bottom right is deeper in CA-RNN^S, which indicates a larger sentence similarity if we use the last hidden state for sentence representations.

Regarding to the dissimilar sentence pair, the words are less similar based on our model. Even for the aligned words as (“*states*”, “*said*”) that can be both about expressions when treating each word independently, our model can yield a very low similarity score by incorporating the aligned contextual information into hidden state modeling. It is also notable that the last cell in our proposed model is in lighter color, indicating a smaller sentence similarity when using the last hidden state for sentence representations.

Conclusions and Future Work

In this paper, we propose a novel context-aligned RNN (CA-RNN) model for sentence similarity modeling. To be specific, a context alignment gating mechanism is presented and well embedded into our model, which can automatically absorb the contextual information of the aligned words in two sentences for better hidden state generation. The experimental results on three datasets for two well-known sentence similarity tasks show the great advantages of our proposed CA-RNN model. In particular, we achieve the new state-of-the-art performance on two widely-used answer selection datasets. In addition, our model is comparable to if not better than the state-of-the-art neural network based approaches for paraphrase identification. It is also interesting to find that different sentence similarity tasks prefer different context alignment directions. Furthermore, the case studies provide an insight of why our proposed model is more effective. In the future, we will investigate the effect of our model on more tasks, such as information retrieval and textual entailment. Another interesting research direction is to study the influence of context alignment directions on more tasks.

Acknowledgements

This research is funded by the National Nature Science Foundation of China (No. 61602179 and 61572193), and Shanghai Municipal Commission of Economy and Information Under Grant Project (No. 201602024). This research is also supported by the Natural Sciences & Engineering Research Council (NSERC) of Canada, an NSERC CREATE award in ADERSIM, the York Research Chairs (YRC) program and an ORF-RE award in BRAIN Alliance. We also acknowledge our team members Jie Zhou and Weijie An.

References

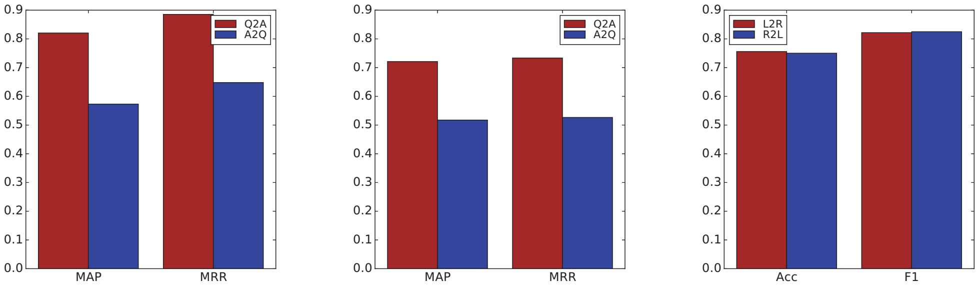
An, W.; Chen, Q.; Tao, W.; Zhang, J.; Yu, J.; Yang, Y.; Hu, Q.; He, L.; and Li, B. 2017. ECNU at 2017 LiveQA track: Learning question similarity with adapted long short-term memory networks. In *TREC*, 1–9.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Blacoe, W., and Lapata, M. 2012. A comparison of vector-based representations for semantic composition. In *EMNLP-CoNLL*, 546–556.

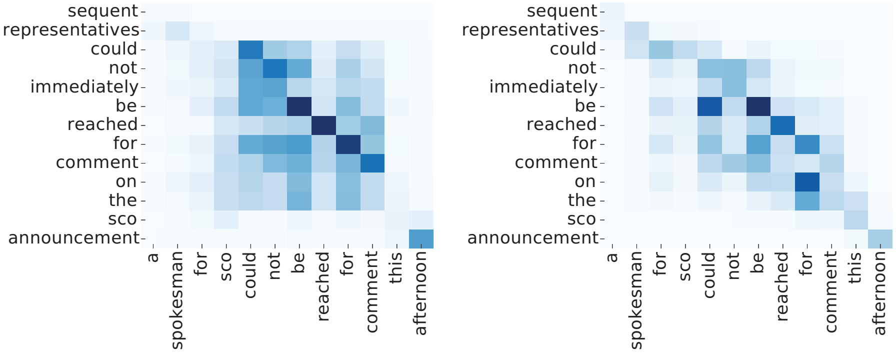
Chen, Q.; Hu, Q.; Huang, J. X.; He, L.; and An, W. 2017. Enhancing recurrent neural networks with positional attention for question answering. In *SIGIR*, 993–996.

Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING*, 350–356.



(a) TREC-QA (b) WikiQA (c) MSRP

Figure 3: Performance of CA-RNN^S (MAX) with various context alignment directions.



(a) CA-RNN^S (b) BLSTM

Figure 4: Word-by-word similarity for a similar sentence pair based on CA-RNN^S and BLSTM. Aligned words: ("*could*", "*could*"), ("*not*", "*not*"), ("*be*", "*be*"), ("*reached*", "*reached*"), ("*for*", "*for*"), ("*comment*", "*comment*"), ("*sco*", "*sco*").

Fang, H.; Wu, F.; Zhao, Z.; Duan, X.; Zhuang, Y.; and Ester, M. 2016. Community-based question answering via heterogeneous social network learning. In *AAAI*, 122–128.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5):602–610.

He, H.; Gimpel, K.; and Lin, J. J. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *EMNLP*, 1576–1586.

Heilman, M., and Smith, N. A. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *NAACL HLT*, 1011–1019.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *NIPS*, 1693–1701.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Hu, B.; Lu, Z.; Li, H.; and Chen, Q. 2014. Convolutional neural network architectures for matching natural language sentences. In *NIPS*, 2042–2050.

Huang, X., and Hu, Q. 2009. A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *SIGIR*, 307–314.

Ji, Y., and Eisenstein, J. 2013. Discriminative improvements to distributional sentence similarity. In *EMNLP*, 891–896.

Jozefowicz, R.; Zaremba, W.; and Sutskever, I. 2015. An empirical exploration of recurrent network architectures. In *ICML*, 2342–2350.

Liu, Y.; Liu, Q.; and Lin, S. 2005. Log-linear models for word alignment. In *ACL*, 459–466.

Madnani, N.; Tetreault, J.; and Chodorow, M. 2012. Re-examining machine translation metrics for paraphrase identification. In *NAACL HLT*, 182–190.

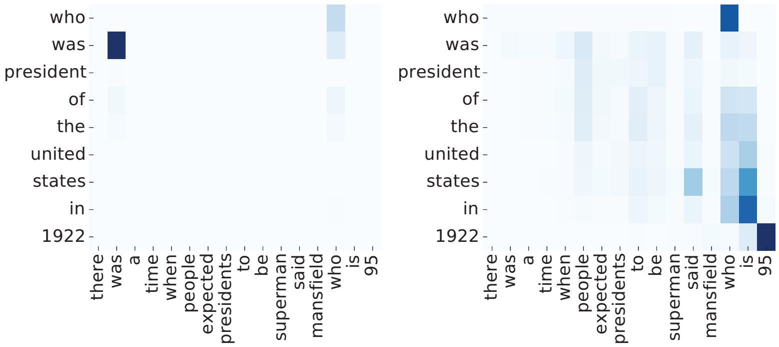
Mueller, J., and Thyagarajan, A. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, 2786–2792.

Och, F. J., and Ney, H. 2000. A comparison of alignment models for statistical machine translation. In *COLING*, 1086–1090.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Rao, J.; He, H.; and Lin, J. 2016. Noise-contrastive estimation for answer selection with deep neural networks. In *CIKM*, 1913–1916.

Santos, C. d.; Tan, M.; Xiang, B.; and Zhou, B. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.



(a) CA-RNN^S

(b) BLSTM

Figure 5: Word-by-word similarity for a dissimilar sentence pair based on CA-RNN^S and BLSTM. Aligned words: ("*president*", "*presidents*"), ("*states*", "*said*").

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Socher, R.; Huang, E. H.; Pennin, J.; Manning, C. D.; and Ng, A. Y. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, 801–809.

Tan, M.; Santos, C. d.; Xiang, B.; and Zhou, B. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.

Van der Plas, L., and Tiedemann, J. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *COLING-ACL poster*, 866–873.

Wan, S.; Dras, M.; Dale, R.; and Paris, C. 2006. Using dependency-based features to take the para-farce out of paraphrase. In *ALTW*, 131–138.

Wang, Z., and Ittycheriah, A. 2015. FAQ-based question answering via word alignment. *arXiv preprint arXiv:1507.02628*.

Wang, D., and Nyberg, E. 2015. A long short-term memory model for answer sentence selection in question answering. In *ACL*, 707–712.

Wang, Y.; Hu, Q.; Song, Y.; and He, L. 2017. Potentiality of healthcare big data: Improving search by automatic query reformulation. In *BigData*.

Wang, B.; Liu, K.; and Zhao, J. 2016. Inner attention based recurrent neural networks for answer selection. In *ACL*, 1288–1297.

Wang, Z.; Mi, H.; and Ittycheriah, A. 2016. Sentence similarity learning by lexical decomposition and composition. *arXiv preprint arXiv:1602.07019*.

Wang, M.; Smith, N. A.; and Mitamura, T. 2007. What is the jeopardy model? a quasi-synchronous grammar for QA. In *EMNLP-CoNLL*, 22–32.

Wieting, J.; Bansal, M.; Gimpel, K.; Livescu, K.; and Roth, D. 2015. From paraphrase database to compositional paraphrase model and back. *arXiv preprint arXiv:1506.03487*.

Wu, H., and Zhou, M. 2003. Optimizing synonym extraction using monolingual and bilingual resources. In *IWP*, 72–79.

Yang, Y.; Yih, W.-t.; and Meek, C. 2015. WikiQA: A challenge dataset for open-domain question answering. In *EMNLP*, 2013–2018.

Yih, S. W.-t.; Chang, M.-W.; Meek, C.; and Pastusiak, A. 2013. Question answering using enhanced lexical semantic models. In *ACL*, 1744–1753.

Yin, W., and Schütze, H. 2015. Convolutional neural network for paraphrase identification. In *NAACL HLT*, 901–911.

Yin, W.; Schütze, H.; Xiang, B.; and Zhou, B. 2015. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *arXiv preprint arXiv:1512.05193*.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, X.; Li, S.; Sha, L.; and Wang, H. 2017. Attentive interactive neural networks for answer selection in community question answering. In *AAAI*, 3525–3531.

Zhao, Z.; Lu, H.; Zheng, V. W.; Cai, D.; He, X.; and Zhuang, Y. 2017. Community-based question answering via asymmetric multi-faceted ranking network learning. In *AAAI*, 3532–3539.