Multiway Attention Networks for Modeling Sentence Pairs

Chuanqi Tan†, Furu Wei‡, Wenhui Wang†, Weifeng Lv†, Ming Zhou‡
†State Key Laboratory of Software Development Environment, Beihang University, China
‡Microsoft Research, Beijing, China
Peking University, Beijing, China

Abstract
Modeling sentence pairs plays the vital role for judging the relationship between two sentences, such as paraphrase identification, natural language inference, and answer sentence selection. Previous work achieves very promising results using neural networks with attention mechanism. In this paper, we propose the multiway attention networks which employ multiple attention functions to match sentence pairs under the matching-aggregation framework. Specifically, we design four attention functions to match words in corresponding sentences. Then, we aggregate the matching information from each function, and combine the information from all functions to obtain the final representation. Experimental results demonstrate that the proposed multiway attention networks improve the result on the Quora Question Pairs, SNLI, MultiNLI, and answer sentence selection task on the SQuAD dataset.

1 Introduction
In this paper, we study the task of modeling a pair of sentences, which aims to compare two sentences and identify the relationship between them. It is a fundamental technology for a variety of tasks. For example, in the paraphrase identification task, it is used to determine whether two sentences are paraphrase or not [Madnani et al., 2012]. In the natural language inference task, it is utilized to judge whether a hypothesis sentence can be inferred from a premise sentence [Bowman et al., 2015]. In the answer sentence selection task, it is employed to assess the relevance between question-answer pairs and rank all candidate answer sentences [Yang et al., 2015]. Table 1 shows corresponding examples of above-mentioned three tasks.

Previous works using neural networks with attention mechanism show effectiveness on this task. These methods can be organized into two frameworks. The first framework is to model sentence pairs by encoding each sentence separately and then making the decision based on the two representations [Yu et al., 2014; Bowman et al., 2016]. The limitation of this framework is that two sentences do not interact during the encoding part. Some methods apply the attention mechanism to improve the interaction of two sentences [Tan et al., 2016]. However, it usually uses the representation of one sentence to attend another sentence, which still works at sentence level but lacks word level interactions. The second framework is based on the matching-aggregation framework [Wang and Jiang, 2016a; Wang et al., 2017b]. It applies the attention mechanism at the word level to improve the matching between words in two sentences. Then the matching information is aggregated to the sentence level for making the decision. This framework enables the word level interaction and leads to promising results. Inspired by this framework, we consider that the word level matching is very important for modeling sentence pairs.

To this end, we propose a multiway attention network (MwAN) for modeling sentence pairs. We propose using multiple attention functions to match two sentences at the word level. Specifically, we use four kinds of attention functions, including the concatenated attention function which is used in Rocktäschel et al. [2015] for natural language infer-

<table>
<thead>
<tr>
<th>Paraphrase Identification</th>
<th>S: she struck a deal with RH to pen a book today</th>
</tr>
</thead>
<tbody>
<tr>
<td>+: she signed a contract with RH to write a book</td>
<td></td>
</tr>
<tr>
<td>-: she denied today that she struck a deal with RH</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Natural Language Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>S: children smiling and waving at camera</td>
</tr>
<tr>
<td>E: there are children present</td>
</tr>
<tr>
<td>C: the kids are frowning</td>
</tr>
<tr>
<td>N: they are smiling at their parents</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Answer Sentence Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q: how are glacier caves formed?</td>
</tr>
<tr>
<td>+: a glacier cave is a cave formed within the ice of a glacier.</td>
</tr>
<tr>
<td>-: the ice facade is approximately 60m high</td>
</tr>
</tbody>
</table>

Table 1: For paraphrase identification, + means it is the paraphrase of S, otherwise -. For natural language inference, E, C, and N mean entailment, contradiction, and neutral, respectively. For answer sentence selection, + means it can answer the question Q, otherwise -.
ence, and the bilinear attention function which is utilized in Chen et al. [2016] to match the question and passage in the reading comprehension. In addition to these two widely-used attention functions, we use two other attention functions that calculate the word relation by the element-wise dot product and difference of two vectors. Such element-wise comparisons are conducted on distributed representations of different text granularities in previous works [Wang and Jiang, 2016a; Chen et al., 2017] to model the interactions between words and sentences. We directly apply these two functions between words, which enhances the word level relation modeling.

Our model is built on the matching-aggregation framework. Given two sentences, we encode them with a bi-directional RNN to obtain contextual word representation of words in two sentences based on the word embeddings. Then we apply four above-mentioned attention functions to match two sentences at the word level. Next, we aggregate the matching information from multiway attention functions with two steps. The matching information of each function is first aggregated by a bi-directional RNN along with all words. Then the outputs of all attention functions are adaptively combined by applying the attention mechanism. We apply another bi-directional RNN to pass through the combined representation to aggregate the mixture matching information. Finally, we apply the attention-pooling to the matching information for a fix-length vector and feed it into a multilayer perceptron for the final decision.

We conduct experiments on three tasks with four standard benchmark datasets. Experimental results show that our multiway attention network achieves state-of-the-art results on the Quora Question Pairs dataset for paraphrase identification, the SNLI and MultiNLI datasets for natural language inference, and the SQuAD dataset for answer sentence selection. We also conduct ablation tests by using single attention function and removing any of functions. Results justify that all of four attention functions help the model performance. Moreover, we observe that different attention functions are good at samples of different categories, which explains why our framework improves the result since it combines the strengths from multiple attention functions.

2 Related Work

Recently, deep learning and neural network models achieve promising results in modeling sentence pairs. Basically, two sentences are encoded into sentence vectors with a neural network encoder, and then the relationship between two sentences was decided solely based on the two sentence vectors [Bowman et al., 2015; Yang et al., 2015; Tan et al., 2016]. However, this kind of framework ignores the lower level interaction between two sentences.

The matching-aggregation framework is therefore proposed matching two sentences at the word level and then aggregating the matching information based on the attention mechanism for the final decision. Rocktäschel et al. [2015] use the attention-based technique to improve the performance of LSTM-based recurrent neural network. They employ the word-by-word attention to obtain a sentence-pair encoding from fine-grained reasoning via soft-alignment of words and phrases in the premise and hypothesis, which achieves very promising result on the SNLI dataset. Wang and Jiang [2016b] propose match-LSTM for the natural language inference that tries to match the current word in the hypothesis with an attention-weighted representation of the premise calculated by the word-by-word attention. Wang et al. [2017b] propose BiMPM that matches two sentences by a bilateral matching with attention mechanism in multiple perspectives, which achieves the state-of-the-art results on the paraphrase identification, natural language inference, and answer sentence selection. They propose four types of representation instead of the attention-weighted representation to improve the result. Moreover, Cheng et al. [2016] propose LSTMN, which improves the attention mechanism by a memory network for the natural language inference. Parikh et al. [2016] match two sentences by attending, comparing, and aggregating for the SNLI dataset. Sha et al. [2016] propose re-reading two sentences with the attention mechanism to improve the memory of the other sentence for a better understanding. Munkhdalai and Yu [2016] propose using the tree structure to improve the recurrent or recursive architecture for the natural language inference and answer sentence selection. Chen et al. [2017] incorporate syntactic parsing information to basic attention-based neural networks, which achieves the state-of-the-art result on the SNLI dataset.

Furthermore, CNN is also applied on the sentence pairs modeling and achieves very promising results [Gong et al., 2017], Yin et al. [2016] propose the attention-based CNN for the paraphrase identification, natural language inference, and answer sentence selection. Wang and Jiang [2016a] use the CNN to aggregate the matching information for the answer sentence selection.

Our model belongs to the matching-aggregation framework. Based on the attention mechanism, we propose using multiple functions to match words in two sentences, and make the decision by aggregating the matching information from multiple attention functions.

3 Task Definition

Formally, we can represent each example of sentence pairs as a triple $(Q, P, y)$, where $Q = (q_1, ..., q_l, ..., q_N)$ is a sentence with a length $N$, $P = (p_1, ..., p_j, ..., p_M)$ is another sentence with a length $M$, and $y \in \mathcal{Y}$ is the label representing the relationship between $Q$ and $P$.

Specifically, for a paraphrase identification task, $Q$ and $P$ are two sentences, $\mathcal{Y} = \{0, 1\}$, where $y = 1$ means that $Q$ and $P$ are paraphrase of each other, and $y = 0$ otherwise.

For a natural language inference task, $Q$ is a premise sentence, $P$ is a hypothesis sentence, and $\mathcal{Y} = \{\text{entailment, contradiction, neutral}\}$ where entailment indicates $P$ can be inferred from $Q$, contradiction indicates $P$ cannot be true condition on $Q$, and neutral means $P$ and $Q$ are irrelevant to each other.

For an answer sentence selection task, $Q$ is a question, $P$ is a candidate answer sentence, and $\mathcal{Y} = \{0, 1\}$ where $y = 1$ means $P$ is a correct answer sentence for $Q$, and $y = 0$ otherwise.
4 Our Approach

We show the multiway attention networks in Figure 1. It consists of five parts under the matching-aggregation framework. Specifically, we obtain contextual word representation for two sentences using a bi-directional RNN. Then we match words between two sentences in multiple ways. For every word pair from $Q$ and $P$, we can obtain four matching scores using multiple attention functions. Next, we aggregate the matching information along with words in $P$ in two steps. We first match two sentences inside each attention function, and then combine the matching information from all functions. The bi-directional RNN is applied to aggregate the matching information both in the inside aggregation and mixed aggregation. Finally, we use the attention pooling to aggregate the matching information in $P$ for a fix-length vector and apply a multilayer perceptron (MLP) classifier for the final decision.

4.1 Gated Recurrent Unit

We use Gated Recurrent Unit (GRU) [Cho et al., 2014] instead of basic RNN. Equation 1 describes the mathematical model of the GRU. $r_t$ and $z_t$ are the gates and $h_t$ is the hidden state.

\begin{align*}
    z_t &= \sigma(W_{hz}h_{t-1} + W_{zx}x_t + b_z) \quad (1a) \\
    r_t &= \sigma(W_{hr}h_{t-1} + W_{xr}x_t + b_r) \quad (1b) \\
    \hat{h}_t &= \Phi(W_{h}(r_t \odot h_{t-1}) + W_{x}x_t + b) \quad (1c) \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \quad (1d)
\end{align*}

4.2 Encoding Layer

Consider two sentences $Q = \{w^q_i\}_{i=1}^N$ and $P = \{w^p_i\}_{i=1}^M$, we first convert the words to their respective word-level embeddings and contextual embeddings. The contextual embeddings are generated by taking the output of a pre-trained language model [Peters et al., 2018], which has shown the effectiveness on many NLP tasks. We then use a bi-directional GRU to produce new representation $h^q_1, \ldots, h^q_N$ and $n^p_1, \ldots, h^p_M$ of all words in two sentences respectively:

\begin{align*}
    w^q_i &= [e^q_i, ln^q_i] \quad (2a) \\
    \overrightarrow{h}^q_i &= \text{GRU}(\overrightarrow{h}^q_{i-1}, w^q_i) \quad (2b) \\
    \overleftarrow{h}^q_i &= \text{GRU}(\overleftarrow{h}^q_{i+1}, w^q_i) \quad (2c)
\end{align*}

and $h^q_i = [\overrightarrow{h}^q_i, \overleftarrow{h}^q_i]$. Meanwhile, we use another bi-directional GRU to encode each $w^p_i$ to the representation $h^p_i$.

4.3 Multiway Matching

Previous works show the effectiveness of the word-level attention in modeling sentence pairs [Rocktäschel et al., 2015; Wang and Jiang, 2016b]. In our model, we design multiple attention functions to compare two vectors of the words in two sentences, namely concat attention, bilinear attention, dot attention, and minus attention. Therefore, at each position $t$ of $P$, the word can match $Q$ using four attention functions to obtain corresponding weighted-sum representations of $Q$.

\begin{align*}
    q^k_t &= f_k(h^q_i, h^p_t, W_k) \quad (3)
\end{align*}

where $h^p_t$ is the representation of $P$ at the position $t$, $h^q_i$ is the representation of all positions in $Q$, and $q^k_t$ is the corresponding representation of $Q$ using the attention function $f_k$ based on the parameter $W_k$, where $k = (c, b, d, m)$, which represents the concat attention, bilinear attention, dot attention, and minus attention, respectively.
Concat Attention:
\[ s_j = v_c^T \tanh(W_c^1 h_q^j + W_c^2 h_p^j) \]  
\[ a_t^i = \exp(s_t^i)/\sum_{j=1}^{N} \exp(s_j^i) \]  
\[ q_t^i = \sum_{i=1}^{N} a_t^i h_q^i \]

Bilinear Attention:
\[ s_j = h_j^T W_p h_q^j \]  
\[ a_t^i = \exp(s_t^i)/\sum_{j=1}^{N} \exp(s_j^i) \]  
\[ q_t^i = \sum_{i=1}^{N} a_t^i h_q^i \]

Dot Attention:
\[ s_j = v_d^T \tanh(W_d(h_q^j \odot h_p^j)) \]  
\[ a_t^i = \exp(s_t^i)/\sum_{j=1}^{N} \exp(s_j^i) \]  
\[ q_t^i = \sum_{i=1}^{N} a_t^i h_q^i \]

Minus Attention:
\[ s_j = v_m^T \tanh(W_m(h_q^j - h_p^j)) \]  
\[ a_t^i = \exp(s_t^i)/\sum_{j=1}^{N} \exp(s_j^i) \]  
\[ q_t^i = \sum_{i=1}^{N} a_t^i h_q^i \]

4.4 Aggregation

We aggregate the matching information from multiway attention functions in two steps. The inside aggregation is to aggregate the matching information along with words in the sentence \( P \) inside each attention function. For each position \( t \), we concatenate the word representation \( h_p^t \) in \( P \) with its corresponding representation \( h_q^j \) of \( Q \) [Wang and Jiang, 2016b] and add a gate to determine the importance of the concatenated representation [Wang et al., 2017a]. Then we use the bi-directional GRU to pass through each position in \( P \). We take the concat attention as an example,
\[ x_t^c = [q_t^i, h_t^p] \]  
\[ g_t = \text{sigmoid}(W_p x_t^c) \]  
\[ x_t^{ce} = g_t \odot x_t^c \]  
\[ \vec{h}_t^c = \text{GRU}(\vec{h}_{t-1}^c, x_t^{ce}) \]  
\[ \overrightarrow{h}_t^c = \text{GRU}(\overrightarrow{h}_{t-1}^c, x_t^{ce}) \]

and \( h_t^c = [\overrightarrow{h}_t^c, \vec{h}_t^c] \). For the bilinear, dot, and minus attention, we obtain the \( h_t^b, h_t^d, \) and \( h_t^m \), respectively.

The mixed aggregation is to combine the matching information from all attention functions. We apply the attention mechanism to adaptively combine four representations with the parameter \( v^p \) as the input.
\[ s_j = v^T \tanh(W_q^1 h_q^j + W_p^2 v^p) \]  
\[ a_t^i = \exp(s_t^i)/\sum_{j=1}^{N} \exp(s_j^i) \]  
\[ x_t = \sum_{i=1}^{N} a_t^i h_t^j \]

Then, we feed \( x_t \) into a bi-directional GRU to aggregate the matching information from multiple attention functions. We obtain \( h_t^q = [\overrightarrow{h}_t^q, \vec{h}_t^q] \) for all positions in \( P \).
\[ \overrightarrow{h}_t^q = \text{GRU}(\overrightarrow{h}_{t-1}^q, x_t) \]  
\[ \vec{h}_t^q = \text{GRU}(\vec{h}_{t-1}^q, x_t) \]

4.5 Prediction Layer

After aggregating the information from multiway matching, we convert the resulting representations of all positions in \( P \) to a fixed-length vector with pooling and feed it into a classifier to determine the relation between two sentences. We first apply an attention pooling with the parameter \( v^q \) to select the important information in \( Q \):
\[ s_j = v^T \tanh(W_q^1 h_q^j + W_p^2 v^q) \]  
\[ a_t = \exp(s_t)/\sum_{j=1}^{N} \exp(s_j) \]  
\[ r^q = \sum_{i=1}^{N} a_i h_q^i \]

Then we use this representation \( r^q \) to select the information in the matching vectors.
\[ s_j = v^T \tanh(W_q^1 h_p^j + W_p^2 r^q) \]  
\[ a_t = \exp(s_t)/\sum_{j=1}^{N} \exp(s_j) \]  
\[ r^p = \sum_{i=1}^{N} a_i h_p^i \]

Finally, we feed \( r^p \) into a multilayer perceptron (MLP) classifier for the probability \( p_i \) of each label in the corresponding task.

For all tasks, the objective function is to minimize the following cross entropy:
\[ \mathcal{L} = \sum_{i=1}^{N} [-y_i \log p_i + (1 - y_i) \log (1 - p_i)] \]

where \( y_i \) denotes a label, in paraphrase detection it is \( (0, 1) \) to represent whether two sentences are paraphrase, in natural language inference it is the relation of two sentences of entailment, contradiction, and neutral, in answer sentence selection it is \( (0, 1) \) to represent whether \( P \) can answer the question \( Q \).

4.6 Implementation Details

We use 300-dimensional uncased pre-trained GloVe embeddings [Pennington et al., 2014] without update during training. We use zero vectors to represent all out-of-vocabulary words. Hidden vector length is set to 150 for all layers. We apply dropout [Srivastava et al., 2014] between layers, with dropout rate 0.2. The model is optimized using AdaDelta [Zeiler, 2012] with initial learning rate of 1.0.

5 Experiment

We conduct experiments on three tasks with four datasets. Experimental results show that our model outperforms our baseline and other competing approaches. We also conduct ablation tests to analyze the contribution of each attention function.

5.1 Dataset and Evaluation Metrics

**Quora Question Pairs** This dataset consists of over 400,000 question pairs, and each question pair is annotated with a binary value indicating whether the two questions are paraphrase of each other. We select 5,000 paraphrases and 5,000 non-paraphrases as the development set, and use another 5,000 paraphrases and 5,000 non-paraphrases as the test set. We keep the remaining instances as the training set.

4414
SNLI It is a natural language inference dataset [Bowman et al., 2015]. The original data set contains 570,152 sentence pairs, each labeled with one of the following relationships: entailment, contradiction, neutral and —, where — indicates a lack of human annotation and is usually discarded [Wang and Jiang, 2016b]. In the end, we have 549,367 pairs for training, 9,842 pairs for development and 9,824 pairs for test.

MultiNLI It is a natural language inference dataset [Williams et al., 2017]. Models can be evaluated on both the matched test examples, which are derived from the same sources as those in the training set, and the mismatched examples, which do not closely resemble any seen at training time. This dataset contains 392,702 pairs for training, 9,815 matched pairs and 9,832 mismatched pairs for development, 9,796 matched pairs and 9,847 mismatched pairs for test.

SQuAD It is a reading comprehension dataset, where the answer to each question is a span of text from the corresponding passage [Rajpurkar et al., 2016]. To evaluate our answer selection task, we split the sentences from the passage using the Stanford CoreNLP Toolkit [Manning et al., 2014] and treat the sentence which contains the correct span as the answer sentence. As the author of SQuAD only publishes the training set and the development set, we split the 10,570 instances in the development set to 5,000 for development and 5,570 for test. We use mean reciprocal rank (MRR) for evaluation.

5.2 Result on Paraphrase Identification
We compare our model with several baselines shown in Table 2. Siamese-CNN and Siamese-LSTM [Wang et al., 2016a] are based on the sentence encoding framework. Multi-Perspective-CNN and Multi-Perspective-LSTM improve them by multiple perspective cosine matching functions [Wang et al., 2017b]. We also compare our MwAN with models under the matching-aggregation framework, such as LDC [Wang et al., 2016b] and BiMPM [Wang et al., 2017b] under RNN and DIIN [Gong et al., 2017] under CNN. Our MwAN outperforms all baselines with 89.60% accuracy on the development set and 89.12% accuracy on the test set.

5.3 Result on Natural Language Inference
Table 3 shows results on the SNLI. Our MwAN achieves the state-of-the-art result with 88.3% test accuracy. Moreover, we also report the ensemble result. Following the ensemble strategy in Wang et al. [2017b], we train the model 4 times with the same setting, and sum the probability of each single model to decide the result. The test accuracy is 89.4%, which outperforms all baselines and achieves the best test score on the SNLI benchmark to date.

Table 4 shows results on the MultiNLI test set. *indicates that the model participating in the competition on June 15th, 2017.

5.4 Result on Answer Sentence Selection
For the answer sentence selection task, we do not use the WikiQA dataset because it is too small to train a complex deep learning model. We implement two basic deep learning baselines Bigram-CNN [Yu et al., 2014] and Attentive LSTM [Tan et al., 2016], and we also report results of two state-of-the-art methods BiMPM [Wang et al., 2017b] and CNN-MULT [Wang and Jiang, 2016a]. All these baselines achieve very promising results on the WikiQA dataset for this task. As shown in Table 6, our MwAN achieves 91.35 in terms of MRR, which outperforms all baseline methods.

Table 2: Results for paraphrase identification on the Quora dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Siamese-CNN</td>
<td>-</td>
<td>79.60</td>
</tr>
<tr>
<td>Multi-Perspective-CNN</td>
<td>-</td>
<td>81.38</td>
</tr>
<tr>
<td>Siamese-LSTM</td>
<td>-</td>
<td>82.58</td>
</tr>
<tr>
<td>Multi-Perspective-LSTM</td>
<td>-</td>
<td>83.21</td>
</tr>
<tr>
<td>L.D.C.</td>
<td>-</td>
<td>85.55</td>
</tr>
<tr>
<td>BiMPM</td>
<td>88.69</td>
<td>88.17</td>
</tr>
<tr>
<td>DIIN</td>
<td>89.44</td>
<td>89.06</td>
</tr>
<tr>
<td>MwAN</td>
<td>89.60</td>
<td>89.12</td>
</tr>
</tbody>
</table>

Table 3: Results for natural language inference on the SNLI dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Matched</th>
<th>Mismatched</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBOW</td>
<td>64.8</td>
<td>64.5</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>66.9</td>
<td>66.9</td>
</tr>
<tr>
<td>ESIM</td>
<td>72.3</td>
<td>72.1</td>
</tr>
<tr>
<td>DIIN</td>
<td>78.8</td>
<td>77.8</td>
</tr>
<tr>
<td>DIIN (ensemble)</td>
<td>80.0</td>
<td>78.7</td>
</tr>
<tr>
<td>TALP-UPC*</td>
<td>67.9</td>
<td>68.2</td>
</tr>
<tr>
<td>LCF-MALTA*</td>
<td>70.7</td>
<td>70.8</td>
</tr>
<tr>
<td>Rivercorners*</td>
<td>72.1</td>
<td>72.1</td>
</tr>
<tr>
<td>Rivercorners (ensemble)*</td>
<td>72.2</td>
<td>72.8</td>
</tr>
<tr>
<td>alpha*</td>
<td>73.5</td>
<td>73.6</td>
</tr>
<tr>
<td>YixinNic-UNC-NLP*</td>
<td>74.5</td>
<td>73.5</td>
</tr>
<tr>
<td>alpha (ensemble)*</td>
<td>74.9</td>
<td>74.9</td>
</tr>
<tr>
<td>MwAN</td>
<td>78.5</td>
<td>77.7</td>
</tr>
<tr>
<td>MwAN (ensemble)</td>
<td>79.8</td>
<td>79.4</td>
</tr>
</tbody>
</table>

Table 4: Results for natural language inference on the MultiNLI test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM with attention [Rocktäschel et al., 2015]</td>
<td>85.3</td>
<td>83.5</td>
</tr>
<tr>
<td>mLSTM [Wang and Jiang, 2016b]</td>
<td>92.0</td>
<td>86.1</td>
</tr>
<tr>
<td>LSTMN [Cheng et al., 2016]</td>
<td>88.5</td>
<td>86.3</td>
</tr>
<tr>
<td>decomposable attention [Parikh et al., 2016]</td>
<td>89.5</td>
<td>86.3</td>
</tr>
<tr>
<td>Intra-sentence attention [Parikh et al., 2016]</td>
<td>90.5</td>
<td>86.8</td>
</tr>
<tr>
<td>NTI-SLSTM-LSTM [Munkhdalai and Yu, 2016]</td>
<td>88.5</td>
<td>87.3</td>
</tr>
<tr>
<td>re-read LSTM [Sha et al., 2016]</td>
<td>90.7</td>
<td>87.5</td>
</tr>
<tr>
<td>BiMMPM [Wang et al., 2017b]</td>
<td>90.9</td>
<td>87.5</td>
</tr>
<tr>
<td>btree-LSTM encoders [Parikh et al., 2016]</td>
<td>88.6</td>
<td>87.6</td>
</tr>
<tr>
<td>DIIN [Gong et al., 2017]</td>
<td>91.2</td>
<td>88.0</td>
</tr>
<tr>
<td>ESIM [Chen et al., 2017]</td>
<td>92.6</td>
<td>88.0</td>
</tr>
<tr>
<td>ESIM+Syntactic tree-LSTM [Chen et al., 2017]</td>
<td>93.5</td>
<td>88.6</td>
</tr>
<tr>
<td>ESIM+Elmo [Peters et al., 2018]</td>
<td>-</td>
<td>88.7</td>
</tr>
<tr>
<td>BiMPPM ensemble [Wang et al., 2017b]</td>
<td>93.2</td>
<td>88.8</td>
</tr>
<tr>
<td>DIIN ensemble [Gong et al., 2017]</td>
<td>92.3</td>
<td>88.9</td>
</tr>
<tr>
<td>MwAN</td>
<td>94.5</td>
<td>88.3</td>
</tr>
<tr>
<td>MwAN (ensemble)</td>
<td>95.5</td>
<td>89.4</td>
</tr>
</tbody>
</table>
Method | Quora | SNLI | MultiNLI | SQuAD
--- | --- | --- | --- | ---
MwAN | 89.60 | 89.12 | 88.58 | 88.30 | 88.78 | 88.45 | 91.45 | 91.35

Concat attention | 87.10 | 86.54 | 88.02 | 87.75 | 76.96 | 76.83 | 90.44 | 90.17
Bilinear attention | 87.19 | 86.57 | 87.99 | 87.69 | 77.57 | 77.51 | 90.39 | 90.10
Dot attention | 89.04 | 88.34 | 88.05 | 87.74 | 78.04 | 77.83 | 90.49 | 90.30
Minus attention | 86.99 | 86.65 | 88.01 | 87.62 | 77.04 | 76.53 | 90.38 | 90.02

w/o Concat attention | 89.06 | 88.62 | 88.36 | 88.15 | 78.36 | 78.01 | 90.80 | 90.59
w/o Bilinear attention | 89.38 | 88.43 | 88.20 | 88.05 | 78.37 | 77.27 | 90.94 | 90.54
w/o Dot attention | 87.50 | 86.96 | 88.40 | 88.11 | 77.83 | 77.71 | 90.62 | 90.45
w/o Minus attention | 89.36 | 88.49 | 88.36 | 88.08 | 78.34 | 78.21 | 91.06 | 90.82

Table 5: Results for answer sentence selection on the SQuAD dataset.

5.5 Discussion

To analyze the effectiveness of each attention function, we conduct ablation tests on four datasets by using single attention function and removing any of attention functions. Since the label of test data in MultiNLI is not published, we analyze it using the Matched and Mismatched data in the development set. As illustrated in Table 5, our MwAN is obviously better than using single attention function. In addition, removing any of attention functions leads to the worse score, which shows the effectiveness of our multiway attention network.

Next, we analyze the effect of four attention functions in different categories of samples. In our adaptive combinations, we can obtain the weight of each attention function at every position in $P$ (Equation 9b). We average all weights of each attention function. Table 7 shows the result for four attention functions. We observe that each attention function has different effect on samples in different categories. The weight of the dot attention increases in positive samples (paraphrase in Quora, entailment in SNLI, and answer sentence in SQuAD) and the other functions increase in negative samples (non-paraphrase in Quora, contradiction in SNLI, and non-answer sentence in SQuAD), which indicates the dot attention is dominant in modeling similarity and other functions are good at modeling difference. Our multiway attention network combines their strengths, therefore achieves better results.

6 Conclusion

In this paper, we propose a multiway attention network that applies multiple attention functions to model the matching between a pair of sentences. We mainly focus on three tasks, namely paraphrase identification, natural language inference, and answer sentence selection. After obtaining contextual word representation based on the word embeddings. We propose using four different attention functions to match words in two sentences. Next, we aggregate the matching information from multiway attention functions. Finally, we use the attention pooling to aggregate the matching information into a fix-length vector and feed it into a classifier for the final decision. Experimental results show that our proposed multiway attention network outperforms baseline neural network models and yields state-of-the-art results on the Quora Question Pairs for the paraphrase identification, SNLI and MultiNLI for the natural language inference, and the answer sentence selection task on the SQuAD dataset.

Acknowledgments

We greatly thank Nan Yang for helpful discussions. We also thank all the anonymous reviewers for their helpful comments. Chuanqi Tan and Weifeng Lv are supported by the National Key R&D Program of China (No. 2017YFB1400200) and National Natural Science Foundation of China (No. 61421003 and 71501003).

References


[Bowman et al., 2016] R. Samuel Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, D. Christopher Manning, and Christopher Potts. A fast unified model for parsing and


