

# Task-Specific Attentive Pooling of Phrase Alignments Contributes to Sentence Matching

Wenpeng Yin, Hinrich Schütze

The Center for Information and Language Processing  
LMU Munich, Germany  
wenpeng@cis.lmu.de

## Abstract

This work studies comparatively two typical sentence matching tasks: textual entailment (TE) and answer selection (AS), observing that weaker phrase alignments are more critical in TE, while stronger phrase alignments deserve more attention in AS. The key to reach this observation lies in phrase detection, phrase representation, phrase alignment, and more importantly how to connect those aligned phrases of different matching degrees with the final classifier.

Prior work (i) has limitations in phrase generation and representation, or (ii) conducts alignment at word and phrase levels by handcrafted features or (iii) utilizes a single framework of alignment without considering the characteristics of specific tasks, which limits the framework’s effectiveness across tasks.

We propose an architecture based on Gated Recurrent Unit that supports (i) representation learning of phrases of *arbitrary granularity* and (ii) task-specific attentive pooling of phrase alignments between two sentences. Experimental results on TE and AS match our observation and show the effectiveness of our approach.

## 1 Introduction

How to model a pair of sentences is a critical issue in many NLP tasks, including textual entailment (Marelli et al., 2014a; Bowman et al., 2015a; Yin et al., 2016a) and answer selection (Yu et al., 2014; Yang et al., 2015; Santos et al., 2016). A key challenge common to these tasks is the lack of explicit alignment annotation between the sentences of the pair. Thus, inferring and assessing the semantic relations between words and phrases in the two sentences is a core issue.

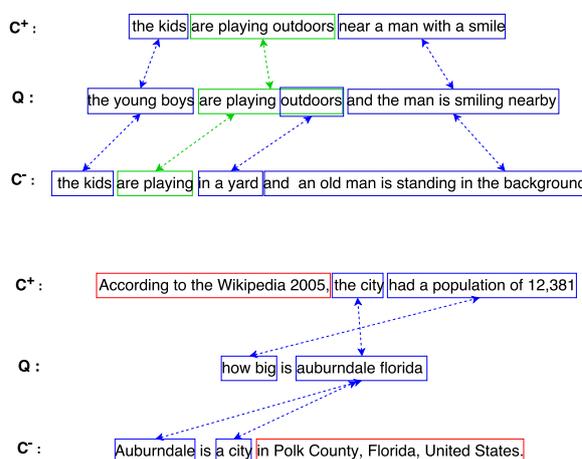


Figure 1: Alignment examples in TE (top) and AS (bottom). Green color: identical (subset) alignment; blue color: relatedness alignment; red color: unrelated alignment. Q: the first sentence in TE or the question in AS;  $C^+$ ,  $C^-$ : the correct or incorrect counterpart in the sentence pair ( $Q$ ,  $C$ ).

Figure 1 shows examples of human annotated phrase alignments. In the TE example, we try to figure out  $Q$  entails  $C^+$  (positive) or  $C^-$  (negative). As human beings, we discover the relationship of two sentences by studying the alignments between linguistic units. We see that some phrases are kept: “are playing outdoors” (between  $Q$  and  $C^+$ ), “are playing ” (between  $Q$  and  $C^-$ ). Some phrases are changed into related semantics on purpose: “the young boys” ( $Q$ )  $\rightarrow$  “the kids” ( $C^+$  &  $C^-$ ), “the man is smiling nearby” ( $Q$ )  $\rightarrow$  “near a man with a smile” ( $C^+$ ) or  $\rightarrow$  “an old man is standing in the background” ( $C^-$ ). We can see that the kept parts have stronger alignments (green color), and changed parts have weaker alignments (blue color). Here, by “strong” / “weak” we mean how semantically close the two aligned phrases are. To successfully identify the relationships of ( $Q$ ,  $C^+$ ) or ( $Q$ ,  $C^-$ ), studying the changed parts is crucial. Hence, we argue that TE should pay more attention to weaker alignments.

In AS, we try to figure out: does sentence  $C^+$  or sentence  $C^-$  answer question  $Q$ ? Roughly, the content in candidates  $C^+$  and  $C^-$  can be classified into aligned part (e.g., repeated or relevant parts) and negligible part. This differs from TE, in which it is hard to claim that some parts are negligible or play a minor role, as TE requires to make clear that each part can entail or be entailed. Hence, TE is considerably sensitive to those “unseen” parts. In contrast, *AS is more tolerant of negligible parts and less related parts*. From the AS example in Figure 1, we see that “Auburndale Florida” ( $Q$ ) can find related part “the city” ( $C^+$ ), and “Auburndale”, “a city” ( $C^-$ ); “how big” ( $Q$ ) also matches “had a population of 12,381” ( $C^+$ ) very well. And some unaligned parts exist, denoted by red color. *Hence, we argue that stronger alignments in AS deserve more attention.*

The above analysis suggests that: (i) alignments connecting two sentences can happen between phrases of arbitrary granularity; (ii) phrase alignments can have different intensities; (iii) tasks of different properties require paying different attention to alignments of different intensities.

Alignments at word level (Yih et al., 2013) or phrase level (Yao et al., 2013) both have been studied before. For example, Yih et al. (2013) make use of WordNet (Miller, 1995) and Probase (Wu et al., 2012) for identifying hyper- and hyponymy. Yao et al. (2013) use POS tags, WordNet and paraphrase database for alignment identification. Their approaches rely on manual feature design and linguistic resources. We develop a deep neural network (DNN) to learn representations of phrases of arbitrary lengths. As a result, alignments can be searched in a more automatic and exhaustive way.

DNNs have been intensively investigated in sentence pair classifications (Blacoe and Lapata, 2012; Socher et al., 2011; Yin and Schütze, 2015b), and attention mechanisms are also applied to individual tasks (Santos et al., 2016; Rocktäschel et al., 2016; Wang and Jiang, 2016); however, most attention-based DNNs have implicit assumption that stronger alignments deserve more attention (Yin et al., 2016a; Santos et al., 2016; Yin et al., 2016b). Our examples in Figure 1, instead, show that this assumption does not hold invariably. Weaker alignments in certain tasks such as TE can be the indicator of the final decision. Our inspiration comes from the analysis of some prior work. For TE, Yin et al. (2016a)

show that considering the pairs in which overlapping tokens are removed can give a boost. This simple trick matches our motivation that weaker alignment should be given more attention in TE. However, Yin et al. (2016a) remove overlapping tokens completely, potentially obscuring complex alignment configurations. In addition, Yin et al. (2016a) use the same attention mechanism for TE and AS, which is less optimal based on our observations.

This motivates us in this work to introduce DNNs with a flexible attention mechanism that is adaptable for specific tasks. For TE, it can make our system pay more attention to weaker alignments; for AS, it enables our system to focus on stronger alignments. We can treat the pre-processing in (Yin et al., 2016a) as a hard way, and ours as a soft way, as our phrases have more flexible lengths and the existence of overlapping phrases decreases the risk of losing important alignments. In experiments, we will show that this attention scheme is very effective for different tasks.

We make the following contributions. (i) We use GRU (Gated Recurrent Unit (Cho et al., 2014)) to learn representations for phrases of arbitrary granularity. Based on phrase representations, we can detect phrase alignments of different intensities. (ii) We propose attentive pooling to achieve flexible choice among alignments, depending on the characteristics of the task. (iii) We achieve state-of-the-art on TE task.

## 2 Related Work

**Non-DNN for sentence pair modeling.** Heilman and Smith (2010) describe tree edit models that generalize tree edit distance by allowing operations that better account for complex re-ordering phenomena and by learning from data how different edits should affect the model’s decisions about sentence relations. Wang and Manning (2010) cope with the alignment between a sentence pair by using a probabilistic model that models tree-edit operations on dependency parse trees. Their model treats alignments as structured latent variables, and offers a principled framework for incorporating complex linguistic features. Guo and Diab (2012) identify the degree of sentence similarity by modeling the missing words (words that are not in the sentence) so as to relieve the sparseness issue of sentence modeling. Yih et

al. (2013) try to improve the shallow semantic component, lexical semantics, by formulating sentence pair as a semantic matching problem with a latent word-alignment structure as in (Chang et al., 2010). More fine-grained word overlap and alignment between two sentences are explored in (Lai and Hockenmaier, 2014), in which negation, hypernym/hyponym, synonym and antonym relations are used. Yao et al. (2013) extend word-to-word alignment to phrase-to-phrase alignment by a semi-Markov CRF. Such approaches often require more computational resources. In addition, using syntactic/semantic parsing during run-time to find the best matching between structured representation of sentences is not trivial.

**DNN for sentence pair classification.** There recently has been great interest in using DNNs for classifying sentence pairs as they can reduce the burden of feature engineering.

For TE, Bowman et al. (2015b) employ recursive DNN to encode entailment on SICK (Marelli et al., 2014b). Rocktäschel et al. (2016) present an attention-based LSTM (long short-term memory, Hochreiter and Schmidhuber (1997)) for the SNLI corpus (Bowman et al., 2015a).

For AS, Yu et al. (2014) present a bigram CNN (convolutional neural network (LeCun et al., 1998)) to model question and answer candidates. Yang et al. (2015) extend this method and get state-of-the-art performance on the WikiQA dataset. Feng et al. (2015) test various setups of a bi-CNN architecture on an insurance domain QA dataset. Tan et al. (2015) explore bidirectional LSTM on the same dataset. Other sentence matching tasks such as paraphrase identification (Socher et al., 2011; Yin and Schütze, 2015a), question – Freebase fact matching (Yin et al., 2016b) etc. are also investigated.

Some prior work aims to solve a general sentence matching problem. Hu et al. (2014) present two CNN architectures for paraphrasing, sentence completion (SC), tweet-response matching tasks. Yin and Schütze (2015b) propose the Multi-GranCNN architecture to model general sentence matching based on phrase matching on multiple levels of granularity. Wan et al. (2016) try to match two sentences in AS and SC by multiple sentence representations, each coming from the local representations of two LSTMs.

**Attention-based DNN for alignment.** DNNs have been successfully developed to detect align-

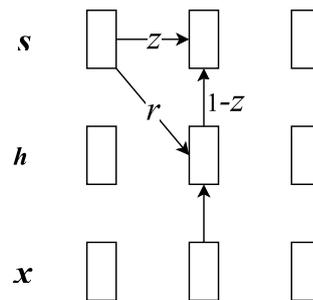


Figure 2: Gated Recurrent Unit

ments, e.g., in machine translation (Bahdanau et al., 2015; Luong et al., 2015) and text reconstruction (Li et al., 2015; Rush et al., 2015). In addition, attention-based alignment is also applied in natural language inference (e.g., Rocktäschel et al. (2016), Wang and Jiang (2016)). However, most of this work aligns word-by-word. As Figure 1 shows, many sentence relations can be better identified through phrase level alignments. This is one motivation of our work.

### 3 Model

This section first gives a brief introduction of GRU and how it performs phrase representation learning, then describes the different attentive poolings for phrase alignments w.r.t TE and AS tasks.

#### 3.1 GRU Introduction

GRU is a simplified version of LSTM. Both are found effective in sequence modeling, as they are order-sensitive and can capture long-range context. The tradeoffs between GRU and its competitor LSTM have not been fully explored yet. According to empirical evaluations in (Chung et al., 2014; Jozefowicz et al., 2015), there is not a clear winner. In many tasks both architectures yield comparable performance and tuning hyper-parameters like layer size is probably more important than picking the ideal architecture. GRU have fewer parameters and thus may train a bit faster or need less data to generalize. Hence, we use GRU, as shown in Figure 2, to model text:

$$\mathbf{z} = \sigma(\mathbf{x}_t \mathbf{U}^z + \mathbf{s}_{t-1} \mathbf{W}^z) \quad (1)$$

$$\mathbf{r} = \sigma(\mathbf{x}_t \mathbf{U}^r + \mathbf{s}_{t-1} \mathbf{W}^r) \quad (2)$$

$$\mathbf{h}_t = \tanh(\mathbf{x}_t \mathbf{U}^h + (\mathbf{s}_{t-1} \circ \mathbf{r}) \mathbf{W}^h) \quad (3)$$

$$\mathbf{s}_t = (1 - \mathbf{z}) \circ \mathbf{h}_t + \mathbf{z} \circ \mathbf{s}_{t-1} \quad (4)$$

$x$  is the input sentence with token  $\mathbf{x}_t \in \mathbb{R}^d$  at position  $t$ ,  $\mathbf{s}_t \in \mathbb{R}^h$  is the hidden state at  $t$ , supposed to

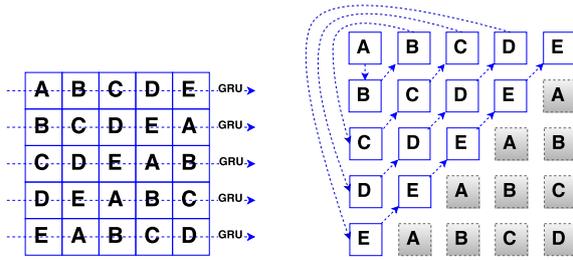


Figure 3: Phrase representation learning by GRU (left), sentence reformatting (right)

encode the history  $x_1, \dots, x_{t-1}$ .  $\mathbf{z}$  and  $\mathbf{r}$  are two gates. All  $\mathbf{U} \in \mathbb{R}^{d \times h}$ ,  $\mathbf{W} \in \mathbb{R}^{h \times h}$  are parameters in GRU.

### 3.2 Representation Learning for Phrases

For a general sentence  $s$  with five consecutive words: ABCDE, with each word represented by a word embedding of dimensionality  $d$ , we first create four fake sentences,  $s^1$ : “BCDEA”,  $s^2$ : “CDEAB”,  $s^3$ : “DEABC” and  $s^4$ : “EABCD”, then put them in a matrix (Figure 3, left).

We run GRUs on each row of this matrix in parallel. As GRU is able to encode the whole sequence up to current position, this step generates representations for any consecutive phrases in original sentence  $s$ . For example, the GRU hidden state at position “E” at coordinates (1,5) (i.e., 1st row, 5th column) denotes the representation of the phrase “ABCDE” which in fact is  $s$  itself, the hidden state at “E” (2,4) denotes the representation of phrase “BCDE”, . . . , the hidden state of “E” (5,1) denotes phrase representation of “E” itself. Hence, for each token, we can learn the representations for all phrases ending with this token. Finally, all phrases of any lengths in  $s$  can get a representation vector. GRUs in those rows are set to share weights so that all phrase representations are comparable in the same space.

Now, we reformat sentence “ABCDE” into  $s^* = \text{“(A) (B) (AB) (C) (BC) (ABC) (D) (CD) (BCD) (ABCD) (E) (DE) (CDE) (BCDE) (ABCDE)”}$ , as shown by arrows in Figure 3 (right), the arrow direction means phrase order. Each sequence in parentheses is a phrase (we use parentheses just for making the phrase boundaries clear). Randomly taking a phrase “CDE” as an example, its representation comes from the hidden state at “E” (3,3) in Figure 3 (left). Shaded parts are discarded. The main advantage of reformatting sentence “ABCDE” into the *new sentence*  $s^*$  is to cre-

ate phrase-level semantic units, but at the same time we maintain the order information.

Hence, the sentence “how big is Auburndale Florida” in Figure 1 will be reformatted into “(how) (big) (how big) (is) (big is) (how big is) (Auburndale) (is Auburndale) (big is Auburndale) (how big is Auburndale) (Florida) (Auburndale Florida) (is Auburndale Florida) (big is Auburndale Florida) (how big is Auburndale Florida)”. We can see that phrases are exhaustively detected and represented.

In the experiments of this work, we explore the phrases of maximal length 7 instead of arbitrary lengths.

### 3.3 Attentive Pooling

As each sentence  $s^*$  consists of a sequence of phrases, and each phrase is denoted by a representation vector generated by GRU, we can compute an *alignment matrix*  $\mathbf{A}$  between two sentences  $s_1^*$  and  $s_2^*$ , by comparing each two phrases, one from  $s_1^*$  and one from  $s_2^*$ . Let  $s_1^*$  and  $s_2^*$  also denote lengths respectively, thus  $\mathbf{A} \in \mathbb{R}^{s_1^* \times s_2^*}$ . While there are many ways of computing the entries of  $\mathbf{A}$ , we found that cosine works well in our setting.

The first step then is to detect the best alignment for each phrase by leveraging  $\mathbf{A}$ . To be concrete, for sentence  $s_1^*$ , we do row-wise max-pooling over  $\mathbf{A}$  as attention vector  $\mathbf{a}_1$ :

$$\mathbf{a}_{1,i} = \max(\mathbf{A}[i, :]) \quad (5)$$

In  $\mathbf{a}_1$ , the entry  $\mathbf{a}_{1,i}$  denotes the best alignment for  $i^{\text{th}}$  phrase in sentence  $s_1^*$ . Similarly, we can do column-wise max-pooling to generate attention vector  $\mathbf{a}_2$  for sentence  $s_2^*$ .

Now, the problem is that we need to pay most attention to the phrases aligned very well or phrases aligned badly. According to the analysis of the two examples in Figure 1, we need to pay more attention to weaker (resp. stronger) alignments in TE (resp. AS). To this end, we adopt different second step over attention vector  $\mathbf{a}_i$  ( $i = 1, 2$ ) for TE and AS.

For TE, in which weaker alignments are supposed to contribute more, we do *k-min-pooling* over  $\mathbf{a}_i$ , i.e., we only keep the  $k$  phrases which are aligned worst. For the  $(Q, C^+)$  pair in TE example of Figure 1, we expect this step is able to put most of our attention to the phrases “the kids”, “the young boys”, “near a man with a smile” and “and the man is smiling nearby” as they have rela-

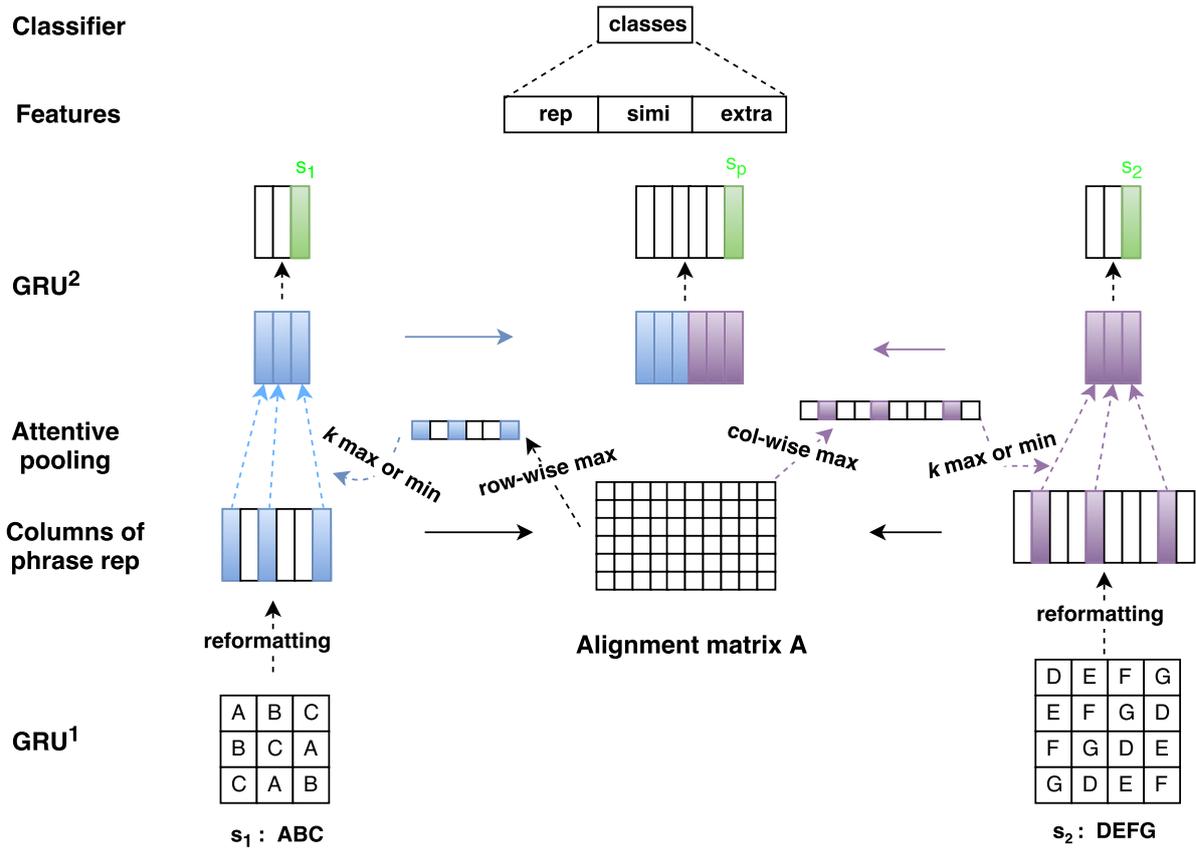


Figure 4: The whole architecture

tively weaker alignments while their relations are the indicator of the final decision.

For AS, in which stronger alignments are supposed to contribute more, we do  $k$ -max-pooling over  $\mathbf{a}_i$ , i.e., we only keep the  $k$  phrases which are aligned best. For the  $(Q, C^+)$  pair in AS example of Figure 1, we expect this  $k$ -max-pooling is able to put most of our attention to the phrases “how big” “Auburndale Florida”, “the city” and “had a population of 12,381” as they have relatively stronger alignments and their relations are the indicator of the final decision. We keep the original order of extracted phrases after  $k$ -min/max-pooling.

In summary, for TE, we first do row-wise max-pooling over alignment matrix, then do  $k$ -min-pooling over generated alignment vector; we use  $k$ -min-max-pooling to denote the whole process. In contrast, we use  $k$ -max-max-pooling for AS. We refer to this method of using two successive min or max pooling steps as *attentive pooling*.

### 3.4 The Whole Architecture

Now, we present the whole system in Figure 4. We take sentences  $s_1$  “ABC” and  $s_2$  “DEFG” as illustration. Each token, i.e., A to F, in the figure is denoted by an embedding vector, hence each sentence is represented as an order-3 tensor as input (they are depicted as rectangles just for simplicity). Based on tensor-style sentence input, we have described the phrase representation learning by GRU<sup>1</sup> in Section 3.2 and attentive pooling in Section 3.3.

Attentive pooling generates a new feature map for each sentence, as shown in Figure 4 (the third layer from the bottom), and each column representation in the feature map denotes a key phrase in this sentence that, based on our modeling assumptions, should be a good basis for the correct final decision. For instance, we expect such a feature map to contain representations of “the young boys”, “outdoors” and “and the man is smiling nearby” for the sentence  $Q$  in the TE example of Figure 1.

Now, we do another GRU<sup>2</sup> step for: 1) the new

	$d$	lr	bs	$L_2$	$div$	$k$
TE	[256,256]	.0001	1	.0006	.06	5
AS	[50,50]	.0001	1	.0006	.06	6

Table 1: Hyperparameters.  $d$ : dimensionality of hidden states in GRU layers; lr: learning rate; bs: mini-batch size;  $L_2$ :  $L_2$  normalization;  $div$ : diversity regularizer;  $k$ :  $k$ -min/max-pooling.

feature map of each sentence mentioned above, to encode all the key phrases as the sentence representation; 2) a concatenated feature map of the two new sentence feature maps, to encode all the key phrases in the two sentences sequentially as the representation of the *sentence pair*. As GRU generates a hidden state at each position, we always choose the last hidden state as the representation of the sentence or sentence pair. In Figure 4 (the fourth layer), these final GRU-generated representations for sentence  $s_1$ ,  $s_2$  and the sentence pair are depicted as green columns:  $s_1$ ,  $s_2$  and  $s_p$  respectively.

As for the input of the final classifier, it can be flexible, such as representation vectors (*rep*), similarity scores between  $s_1$  and  $s_2$  (*simi*), and extra linguistic features (*extra*). This can vary based on the specific tasks. We give details in Section 4.

## 4 Experiments

We test the proposed architectures on TE and AS benchmark datasets.

### 4.1 Common Setup

For both TE and AS, words are initialized by 300-dimensional GloVe embeddings<sup>1</sup> (Pennington et al., 2014) and not changed during training. A single randomly initialized embedding is created for all unknown words by uniform sampling from  $[-.01, .01]$ . We use ADAM (Kingma and Ba, 2015), with a first momentum coefficient of 0.9 and a second momentum coefficient of 0.999,<sup>2</sup>  $L_2$  regularization and Diversity Regularization (Xie et al., 2015). Table 1 shows the values of the hyperparameters, tuned on dev.

**Classifier.** Following Yin et al. (2016a), we use three classifiers – logistic regression in DNN, logistic regression and linear SVM with default parameters<sup>3</sup> directly on the feature vector – and report performance of the best.

<sup>1</sup>[nlp.stanford.edu/projects/glove/](http://nlp.stanford.edu/projects/glove/)

<sup>2</sup>Standard configuration recommended by Kingma and Ba

<sup>3</sup><http://scikit-learn.org/stable/> for both.

**Common Baselines.** (i) **Addition.** We sum up word embeddings element-wise to form sentence representation, then concatenate two sentence representation vectors ( $s_1^0$ ,  $s_2^0$ ) as classifier input. (ii) **A-LSTM.** The pioneering attention based LSTM system for a specific sentence pair classification task “natural language inference” (Rocktäschel et al., 2016). A-LSTM has the same dimensionality as our GRU system in terms of initialized word representations and the hidden states. (iii) **ABCNN** (Yin et al., 2016a). The state-of-the-art system in both TE and AS.

Based on the motivation in Section 1, the main hypothesis to be tested in experiments is:  $k$ -min-max-pooling is superior for TE and  $k$ -max-max-pooling is superior for AS. In addition, we would like to determine whether the second pooling step in attention pooling, i.e., the  $k$ -min/max-pooling, is more effective than a “full-pooling” in which *all* the generated phrases are forwarded into the next layer.

### 4.2 Textual Entailment

SemEval 2014 Task 1 (Marelli et al., 2014a) evaluates system predictions of textual entailment (TE) relations on sentence pairs from the SICK dataset (Marelli et al., 2014b). The three classes are entailment, contradiction and neutral. The sizes of SICK train, dev and test sets are 4439, 495 and 4906 pairs, respectively. *We choose SICK benchmark dataset so that our result is directly comparable with that of (Yin et al., 2016a), in which non-overlapping text are utilized explicitly to boost the performance. That trick inspires this work.*

Following Lai and Hockenmaier (2014), we train our final system (after fixing of hyperparameters) on train and dev (4,934 pairs). Our evaluation measure is accuracy.

#### 4.2.1 Feature Vector

The final feature vector as input of classifier contains three parts: *rep*, *simi*, *extra*.

**Rep.** Totally five vectors, three are the top sentence representation  $s_1$ ,  $s_2$  and the top sentence pair representation  $s_p$  (shown in green in Figure 4), two are  $s_1^0$ ,  $s_2^0$  from *Addition* baseline.

**Simi.** Four similarity scores, cosine similarity and euclidean distance between  $s_1$  and  $s_2$ , cosine similarity and euclidean distance between  $s_1^0$  and  $s_2^0$ . Euclidean distance  $\| \cdot \|$  is transformed into  $1/(1 + \| \cdot \|)$ .

	method	acc
SemEval Top3	(Jimenez et al., 2014)	83.1
	(Zhao et al., 2014)	83.6
	(Lai and Hockenmaier, 2014)	84.6
TrRNTN	(Bowman et al., 2015b)	76.9
Addition	no features	73.1
	plus features	79.4
A-LSTM	no features	78.0
	plus features	81.7
ABCNN	(Yin et al., 2016a)	86.2
GRU <i>k</i> -min-max ablation	– rep	86.4
	– simi	85.1
	– extra	85.5
GRU	<i>k</i> -max-max-pooling	84.9
	full-pooling	85.2
	<i>k</i> -min-max-pooling	<b>87.1*</b>

Table 2: Results on SICK. Significant improvement over both *k*-max-max-pooling and full-pooling is marked with \* (test of equal proportions,  $p < .05$ ).

**Extra.** We include the same 22 linguistic features as Yin et al. (2016a). They cover 15 machine translation metrics between the two sentences; whether or not the two sentences contain negation tokens like “no”, “not” etc; whether or not they contain synonyms, hypernyms or antonyms; two sentence lengths. See Yin et al. (2016a) for details.

#### 4.2.2 Results

Table 2 shows that GRU with *k*-min-max-pooling gets state-of-the-art performance on SICK and significantly outperforms *k*-max-max-pooling and full-pooling. Full-pooling has more phrase input than the combination of *k*-max-max-pooling and *k*-min-max-pooling, this might bring two problems: (i) noisy alignments increase; (ii) sentence pair representation  $s_p$  is no longer discriminative –  $s_p$  does not know its semantics comes from phrases of  $s_1$  or  $s_2$ : as different sentences have different lengths, the boundary location separating two sentences varies across pairs. However, this is crucial to determine whether  $s_1$  entails  $s_2$ .

ABCNN (Yin et al., 2016a) is based on assumptions similar to *k*-max-max-pooling: words/phrases with higher matching values should contribute more in this task. However, ABCNN gets the optimal performance by combining a reformatted SICK version in which

	method	MAP	MRR
Baselines	CNN-Cnt	0.6520	0.6652
	Addition	0.5021	0.5069
	Addition-Cnt	0.5888	0.5929
	A-LSTM	0.5321	0.5469
	A-LSTM-Cnt	0.6388	0.6529
	AP-CNN	0.6886	0.6957
GRU <i>k</i> -max-max ablation	ABCNN	0.6921	0.7127
	– rep	0.6913	0.6994
	– simi	0.6764	0.6875
GRU	– extra	0.6802	0.6899
	<i>k</i> -min-max-pooling	0.6674	0.6791
	full-pooling	0.6693	0.6785
	<i>k</i> -max-max-pooling	<b>0.7124*</b>	<b>0.7237*</b>

Table 3: Results on WikiQA. Significant improvement over both *k*-min-max-pooling and full-pooling is marked with \* ( $t$ -test,  $p < .05$ ). STOA: 74.17 (MAP)/75.88 (MRR) in (Tymoshenko et al., 2016)

overlapping tokens in two sentences are removed. This instead hints that non-overlapping units can do a big favor for this task, which is indeed the superiority of our “*k*-min-max-pooling”.

### 4.3 Answer Selection

We use WikiQA<sup>4</sup> subtask that assumes there is at least one correct answer for a question. This dataset consists of 20,360, 1130 and 2352 question-candidate pairs in train, dev and test, respectively. Following Yang et al. (2015), we truncate answers to 40 tokens and report mean average precision (MAP) and mean reciprocal rank (MRR).

Apart from the common baselines Addition, A-LSTM and ABCNN, we compare further with: (i) **CNN-Cnt** (Yang et al., 2015): combine CNN with two linguistic features “WordCnt” (the number of non-stopwords in the question that also occur in the answer) and “WgtWordCnt” (reweight the counts by the IDF values of the question words); (ii) **AP-CNN** (Santos et al., 2016).

#### 4.3.1 Feature Vector

The final feature vector in AS has the same (*rep*, *simi*, *extra*) structure as TE, except that **simi** consists of only two cosine similarity scores, and **extra** consists of four entries: two sentence lengths, WordCnt and WgtWordCnt.

<sup>4</sup><http://aka.ms/WikiQA> (Yang et al., 2015)



CNN layer to detect bigger-granular phrases, their phrases in different CNN layers cannot be aligned directly as they are in different spaces. GRU in this work uses the same weights to learn representations of arbitrary-granular phrases, hence, all phrases can share the representations in the same space and can be compared directly.

#### 4.4 Visual Analysis

In this subsection, we visualize the attention distributions over phrases, i.e.,  $a_i$  in Equation 5, of example sentences in Figure 1 (for space limit, we only show this for TE example). Figures 5(a)-5(b) respectively show the attention values of each phrase in  $(Q, C^+)$  pair in TE example in Figure 1. We can find that  $k$ -min-pooling over this distributions can indeed detect some key phrases that are supposed to determine the pair relations. Taking Figure 5(a) as an example, phrases “young boys”, phrases ending with “and”, phrases “smiling”, “is smiling”, “nearby” and a couple of phrases ending with “nearby” have lowest attention values. According to our  $k$ -min-pooling step, these phrases will be detected as key phrases. Considering further the Figure 5(b), phrases “kids”, phrases ending with “near”, and a couple of phrases ending with “smile” are detected as key phrases.

If we look at the key phrases in both sentences, we can find that the discovering of those key phrases matches our analysis in Section 1 for TE example: “kids” corresponds to “young boys”, “smiling nearby” corresponds to “near...smile”.

Another interesting phenomenon is that, taking Figure 5(b) as example, even though “are playing outdoors” can be well aligned as it appears in both sentences, nevertheless the visualization figures show that the attention values of “are playing outdoors and” in  $Q$  and “are playing outdoors near” drop dramatically. This hints that our model can get rid of some surface matching, as the key token “and” or “near” makes the semantics of “are playing outdoors and” and “are playing outdoors near” be pretty different with their sub-phrase “are playing outdoors”. This is important as “and” or “near” is crucial unit to connect the following key phrases “smiling nearby” in  $Q$  or “a smile” in  $C^+$ . If we connect those key phrases sequentially as a new fake sentence, as we did in attentive pooling layer of Figure 4, we can see that the fake sentence roughly “reconstructs” the meaning of the original sentence while it is composed of phrase-level se-

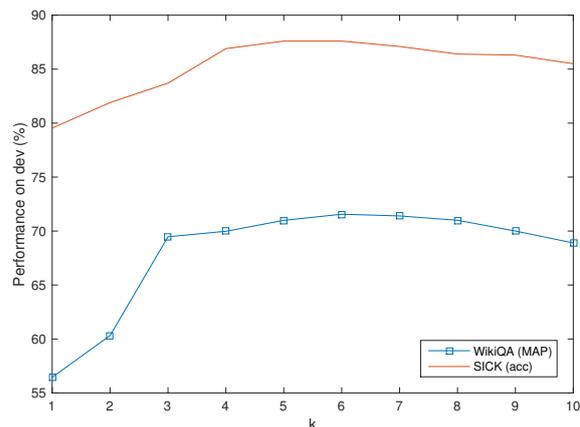


Figure 6: Effects of pooling size  $k$  (cf. Table Table 1)

mantic units now.

#### 4.5 Effects of Pooling Size $k$

The key idea of the proposed method is achieved by the  $k$ -min/max pooling. We show how the hyperparameter  $k$  influences the results by tuning on the dev sets.

In Figure 6, we can see the performance trends of changing  $k$  value between 1 and 10 in the two tasks. Roughly  $k > 4$  can give competitive results, but larger values bring performance drop.

### 5 Conclusion

In this work, we investigate the contribution of different intensities of phrase alignments for different tasks. We argue that it is not true that stronger alignments always matter more. We found TE task prefers weaker alignments while AS task prefers stronger alignments. We proposed flexible attentive poolings in GRU system to satisfy the different requirements of different tasks. Experimental results show the soundness of our argument and the effectiveness of our attention pooling based GRU systems.

As future work, we plan to investigate phrase representation learning in context and how to conduct the attentive pooling automatically regardless of the categories of the tasks.

#### Acknowledgments

We gratefully acknowledge the support of Deutsche Forschungsgemeinschaft for this work (SCHU 2246/8-2).

## References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.
- William Blacoe and Mirella Lapata. 2012. A comparison of vector-based representations for semantic composition. In *Proceedings of EMNLP-CoNLL*, pages 546–556.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642.
- Samuel R Bowman, Christopher Potts, and Christopher D Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of CVSC workshop*, pages 12–21.
- Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. 2010. Discriminative learning over constrained latent representations. In *Proceedings of NAACL-HLT*, pages 429–437.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. *Proceedings of IEEE ASRU Workshop*.
- Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of ACL*, pages 864–872.
- Michael Heilman and Noah A Smith. 2010. Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proceedings of NAACL-HLT*, pages 1011–1019.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Proceedings of NIPS*, pages 2042–2050.
- Sergio Jimenez, George Duenas, Julia Baquero, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2014. Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment. *SemEval*, pages 732–742.
- Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of ICML*, pages 2342–2350.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.
- Alice Lai and Julia Hockenmaier. 2014. Illinois-lh: A denotational and distributional approach to semantics. *SemEval*, pages 329–334.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of ACL*, pages 1106–1115.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421.
- Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *SemEval*, pages 1–8.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, pages 216–223.
- George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.
- Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of ICLR*.
- Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of EMNLP*, pages 379–389.
- Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.
- Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of NIPS*, pages 801–809.

- Ming Tan, Bing Xiang, and Bowen Zhou. 2015. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*.
- Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2016. Convolutional neural networks vs. convolution kernels: Feature engineering for answer sentence reranking. In *Proceedings of NAACL-HLT*, pages 1268–1278.
- Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of AAAI*, pages 2835–2841.
- Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of NAACL*, pages 1442–1451.
- Mengqiu Wang and Christopher D Manning. 2010. Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. In *Proceedings of Coling*, pages 1164–1172.
- Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probbase: A probabilistic taxonomy for text understanding. In *Proceedings of SIGMOD*, pages 481–492.
- Pengtao Xie, Yuntian Deng, and Eric Xing. 2015. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of EMNLP*, pages 2013–2018.
- Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. Semi-markov phrase-based monolingual alignment. In *Proceedings of EMNLP*, pages 590–600.
- Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and Andrzej Pastusiak. 2013. Question answering using enhanced lexical semantic models. In *Proceedings of ACL*, pages 1744–1753.
- Wenpeng Yin and Hinrich Schütze. 2015a. Convolutional neural network for paraphrase identification. In *Proceedings of NAACL*, pages 901–911, May–June.
- Wenpeng Yin and Hinrich Schütze. 2015b. Multi-granccnn: An architecture for general matching of text chunks on multiple levels of granularity. In *Proceedings of ACL-IJCNLP*, pages 63–73.
- Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016a. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *TACL*, 4:259–272.
- Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. 2016b. Simple question answering by attentive convolutional neural network. In *Proceedings of COLING*, pages 1746–1756.
- Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *NIPS Deep Learning Workshop*.
- Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment. *SemEval*, pages 271–277.