End-Task Oriented Textual Entailment via Deep Explorations of Inter-Sentence Interactions

Wenpeng Yin,[1] **Hinrich Schütze,**[2] **Dan Roth**[1]
[1]Dept. of Computer Science, University of Pennsylvania, USA
[2]CIS, LMU Munich, Germany
{wenpeng,danroth}@seas.upenn.edu

Abstract

This work deals with SCITAIL, a natural entailment challenge derived from a multi-choice question answering problem. The premises and hypotheses in SCITAIL were generated with no awareness of each other, and did not specifically aim at the entailment task. This makes it more challenging than other entailment data sets and more directly useful to the end-task – question answering. We propose DEISTE (**d**eep **e**xplorations of **i**nter-**s**entence interactions for **t**extual **e**ntailment) for this entailment task. Given word-to-word interactions between the premise-hypothesis pair (P, H), DEISTE consists of: (i) a *parameter-dynamic convolution* to make important words in P and H play a dominant role in learnt representations; and (ii) a *position-aware attentive convolution* to encode the representation and position information of the aligned word pairs. Experiments show that DEISTE gets ≈5% improvement over prior state of the art and that the pretrained DEISTE on SCITAIL generalizes well on RTE-5.[1]

1 Introduction

Textual entailment (TE) is a fundamental problem in natural language understanding and has been studied intensively recently using multiple benchmarks – PASCAL RTE challenges (Dagan et al., 2006, 2013), Paragraph-Headline (Burger and Ferro, 2005), SICK (Marelli et al., 2014) and SNLI (Bowman et al., 2015). In particular, SNLI – while much easier than earlier datasets – has generated much work based on deep neural networks due to its large size. However, these benchmarks were mostly derived independently of any NLP problems.[2] Therefore, the premise-hypothesis pairs were composed under the constraint of predefined rules and the language skills of humans. As a result, while top-performing systems push forward the state-of-the-art, they do not necessarily learn to support language inferences that emerge commonly and naturally in real NLP problems.

Premise P	
Pluto rotates once on its axis every 6.39 Earth days.	0
Once per day, the earth rotates about its axis.	1
It rotates on its axis once every 60 Earth days.	0
Earth orbits Sun, and rotates once per day about axis.	1

Table 1: Examples of four premises for the hypothesis "Earth rotates on its axis once times in one day" in SCITAIL dataset. Right column (label): "1" means *entail*, "0" otherwise.

In this work, we study SCITAIL (Khot et al., 2018), an end-task oriented challenging entailment benchmark. SCITAIL is reformatted from a multi-choice question answering problem. All hypotheses H were obtained by rewriting (question, correct answer) pairs; all premises P are relevant web sentences collected by an Information retrieval (IR) method; then (P, H) pairs are annotated via crowdsourcing. Table 1 shows examples. *By this construction, a substantial performance gain on* SCITAIL *can be turned into better QA performance* (Khot et al., 2018). Khot et al. (2018) report that SCITAIL challenges neural entailment models that show outstanding performance on SNLI, e.g., Decomposable Attention Model (Parikh et al., 2016) and Enhanced LSTM (Chen et al., 2017).

We propose DEISTE for SCITAIL. Given word-to-word inter-sentence interactions between

[1] https://github.com/yinwenpeng/SciTail

[2] RTE-{5,6,7} is an exception to this rule.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 540–545
Melbourne, Australia, July 15 - 20, 2018.

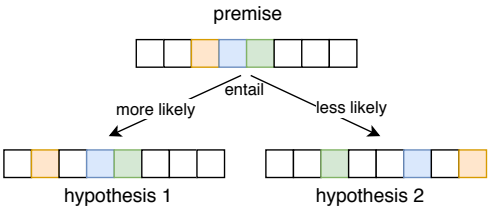

Figure 1: The motivation of considering alignment positions in TE. The same color in (premise, hypothesis) means the two words are best aligned.

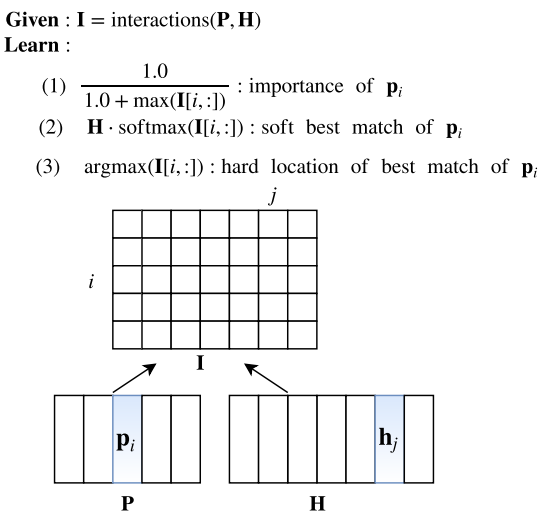

Figure 2: The basic principles of DEISTE in modeling the pair (P, H)

the pair (P, H), DEISTE pursues three deep exploration strategies of these interactions. (a) How to express the importance of a word, and let it play a dominant role in learnt representations. (b) For any word in one of (P, H), how to find the best aligned word in the other sentence, so that we know their connection is indicative of the final decision. (c) For a window of words in P or H, whether the locations of their best aligned words in the other sentence provides clues. As Figure 1 illustrates, the premise "in this incident, the cop (C) shot (S) the thief (T)" is more likely to entail the hypothesis "$\hat{C} \ldots \hat{S} \ldots \hat{T}$" than "$\hat{T} \ldots \hat{S} \ldots \hat{C}$" where $\hat{X}$ is the word that best matches X.

Our model DEISTE is implemented in convolutional neural architecture (LeCun et al., 1998). Specifically, DEISTE consists of (i) a parameter-dynamic convolution for exploration strategy (a) given above; and (ii) a position-aware attentive convolution for exploration strategies (b) and (c). In experiments, DEISTE outperforms prior top systems by ≈5%. Perhaps even more interestingly, the pretrained model over SCITAIL generalizes well on RTE-5.

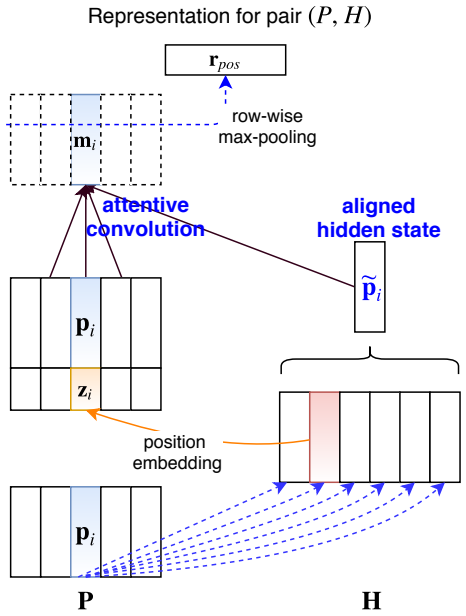

Figure 3: Position-aware attentive convolution in modeling the pair (P, H)

2 Method

To start, a sentence S ($S \in \{P, H\}$) is represented as a sequence of n_s hidden states, e.g., $\mathbf{p}_i, \mathbf{h}_i \in \mathbb{R}^d$ ($i = 1, 2, \ldots, |n_{p/h}|$), forming a feature map $\mathbf{S} \in \mathbb{R}^{d \times |n_s|}$, where d is the dimensionality of hidden states. Figure 2 depicts the basic principle of DEISTE in modeling premise-hypothesis pair (P, H) with feature maps $\mathbf{P}$ and $\mathbf{H}$, respectively.

First, $\mathbf{P}$ and $\mathbf{H}$ have fine-grained interactions $\mathbf{I} \in \mathbb{R}^{n_p \times n_h}$ by comparing any pair of ($\mathbf{p}_i$,$\mathbf{h}_j$):

$$\mathbf{I}[i, j] = \text{cosine}(\mathbf{p}_i, \mathbf{h}_j) \quad (1)$$

We now elaborate DEISTE's exploration strategies (a), (b) and (c) of the interaction results $\mathbf{I}$.

2.1 Parameter-dynamic convolution

Intuitively, important words should be expressed more intensively than other words in the learnt representation of a sentence. However, the importance of words within a specific sentence might not depend on the sentence itself. E.g., Yin and Schütze (2017b) found that in question-aware answer sentence selection, words well matched are more important; while in textual entailment, words hardly matched are more important.

In this work, we try to make the semantics of those important words dominate in the output representations of a convolution encoder.

Given a local window of hidden states in the feature map $\mathbf{P}$, e.g., three adjacent ones $\mathbf{p}_{i-1}$, $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, a conventional convolution learns a

higher-level representation $\mathbf{r}$ for this trigram:

$$\mathbf{r} = \tanh(\mathbf{W} \cdot [\mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1}] + \mathbf{b}) \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{d \times 3d}$ and $\mathbf{b} \in \mathbb{R}^d$.

For simplicity, we neglect the bias term $\mathbf{b}$ and split the multiplication of big matrices – $\mathbf{W} \cdot [\mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1}]$ – into three parts, then $\mathbf{r}$ can be formulated as:

$$\begin{aligned}\mathbf{r} &= \tanh(\mathbf{W}^{-1} \cdot \mathbf{p}_{i-1} \oplus \mathbf{W}^0 \cdot \mathbf{p}_i \oplus \mathbf{W}^{+1} \cdot \mathbf{p}_{i+1}) \\ &= \tanh(\hat{\mathbf{p}}_{i-1} \oplus \hat{\mathbf{p}}_i \oplus \hat{\mathbf{p}}_{i+1})\end{aligned}$$

where $\oplus$ means element-wise addition; $\mathbf{W}^{-1}$, $\mathbf{W}^0$, $\mathbf{W}^{+1} \in \mathbf{R}^{d \times d}$, and their concatenation equals to the $\mathbf{W}$ in Equation 2; $\hat{\mathbf{p}}_i$ is set to be $\mathbf{W}^0 \cdot \mathbf{p}_i$, so $\hat{\mathbf{p}}_i$ can be seen as the projection of $\mathbf{p}_i$ in a new space by parameters $\mathbf{W}^0$; finally the three projected outputs contribute equally in the addition: $\hat{\mathbf{p}}_{i-1} \oplus \hat{\mathbf{p}}_i \oplus \hat{\mathbf{p}}_{i+1}$.

The convolution encoder shares parameters $[\mathbf{W}^{-1}, \mathbf{W}^0, \mathbf{W}^{+1}]$ in all trigrams, so we cannot expect those parameters to express the importance of $\hat{\mathbf{p}}_{i-1}$, $\hat{\mathbf{p}}_i$ or $\hat{\mathbf{p}}_{i+1}$ in the output representation $\mathbf{r}$. Instead, we formulate the idea as follows:

$$\mathbf{m}_i = \tanh(\alpha_{i-1}\hat{\mathbf{p}}_{i-1} \oplus \alpha_i \hat{\mathbf{p}}_i \oplus \alpha_{i+1}\hat{\mathbf{p}}_{i+1})$$

in which the three scalars α_{i-1}, α_i and α_{i+1} indicate the importance scores of $\hat{\mathbf{p}}_{i-1}$, $\hat{\mathbf{p}}_i$ and $\hat{\mathbf{p}}_{i+1}$ respectively. In our work, we adopt:

$$\alpha_i = \frac{1.0}{1.0 + \max(\mathbf{I}[i, :])} \quad (3)$$

Since $\alpha_i \hat{\mathbf{p}}_i = \alpha_i \mathbf{W}^0 \cdot \mathbf{p}_i = \mathbf{W}^{0,i} \cdot \mathbf{p}_i$, we notice that the original shared parameter $\mathbf{W}^0$ is mapped to a dynamic parameter $\mathbf{W}^{0,i}$, which is specific to the input $\mathbf{p}_i$. We refer to this as *parameter-dynamic convolution*, which enables our system DEISTE to highlight important words in higher-level representations.

Finally, a max-pooling layer is stacked over $\{\mathbf{m}_i\}$ to get the representation for the pair (P, H), denoted as $\mathbf{r}_{dyn}$.

2.2 Position-aware attentive convolution

Our position-aware attentive convolution, shown in Figure 3, aims to encode the representations as well as the positions of the best word alignments.

Representation. Given the interaction scores in $\mathbf{I}$, the representation $\tilde{\mathbf{p}}_i$ of all soft matches for hidden state $\mathbf{p}_i$ in P is the weighted average of all hidden states in H:

$$\tilde{\mathbf{p}}_i = \sum_j \text{softmax}(\mathbf{I}[i, :])_j \cdot \mathbf{h}_j \quad (4)$$

methods	dev	test
Majority Class	50.4	60.4
Hypothesis only	66.9	65.1
Premise only	72.6	73.4
NGram model	65.0	70.6
ESIM-600D	70.5	70.6
Decomp-Att	75.4	72.3
DGEM	79.6	77.3
AttentiveConvNet	79.3	78.1
DEISTE	**82.4**	**82.1**
w/o dyn-conv	80.2	79.1
w/o representation	76.3	74.9
w/o position	82.1	81.3

Table 2: DEISTE vs. baselines on SCITAIL

Position. For a trigram $[\mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1}]$ in P, knowing where its best-matched words are located in H is helpful in TE, as discussed in Section 1.

For $\mathbf{p}_i$, we first retrieve the index x_i of the best-matched word in H by:

$$x_i = \text{argmax}(\mathbf{I}[i, :]) \quad (5)$$

then embed the concrete $\{x_i\}$ into randomly-initialized continuous space:

$$\mathbf{z}_i = \mathbf{M}[x_i] \quad (6)$$

where $\mathbf{M} \in \mathbb{R}^{n_h \times d_m}$; n_h is the length of H; d_m is the dimensionality of position embeddings.

Now, the three positions $[i-1, i, i+1]$ in P concatenate vector-wisely original hidden states $[\mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1}]$ with position embeddings $[\mathbf{z}_{i-1}, \mathbf{z}_i, \mathbf{z}_{i+1}]$, getting a new sequence of hidden states: $[\mathbf{c}_{i-1}, \mathbf{c}_i, \mathbf{c}_{i+1}]$. As a result, a position i in P has hidden state $\mathbf{c}_i$, left context $\mathbf{c}_{i-1}$, right context $\mathbf{c}_{i+1}$ and the representation of soft-aligned words in H, i.e., $\tilde{\mathbf{p}}_i$. Then a convolution works at position i in P as:

$$\mathbf{n}_i = \tanh(\mathbf{W} \cdot [\mathbf{c}_{i-1}, \mathbf{c}_i, \mathbf{c}_{i+1}, \tilde{\mathbf{p}}_i] + \mathbf{b}) \quad (7)$$

As Figure 3 shows, the position-aware attentive convolution finally stacks a standard max-pooling layer over $\{\mathbf{n}_i\}$ to get the representation for the pair (P, H), denoted as $\mathbf{r}_{pos}$.

Overall, our DEISTE will generate a representation $\mathbf{r}_{dyn}$ through the parameter-dynamic convolution, and generate a representation $\mathbf{r}_{pos}$ through the position-aware attentive convolution. Finally the concatenation – $[\mathbf{r}_{dyn}, \mathbf{r}_{pos}]$ – is fed to a logistic regression classifier.

methods	acc.
Majority Class	34.3
Premise only	33.5
Hypothesis only	68.7
ESIM	86.7
Decomp-Att	86.8
AttentiveConvNet	86.3
DEISTE$_{\text{SCITAIL}}$	84.7
DEISTE$_{tune}$	87.1
State-of-the-art	88.7

Table 3: DEISTE vs. baselines on SNLI. DEISTE$_{\text{SCITAIL}}$ has exactly the same system layout and hyperparameters as the one reported on SCITAIL in Table 2; DEISTE$_{tune}$: tune hyperparameters on SNLI dev. State-of-the-art refers to (Peters et al., 2018). Ensemble results are not considered.

3 Experiments

Dataset. SCITAIL[3] (Khot et al., 2018) is for textual entailment in binary classification: entailment or neutral. Accuracy is reported.

Training setup. All words are initialized by 300D Word2Vec (Mikolov et al., 2013) embeddings, and are fine-tuned during training. The whole system is trained by AdaGrad (Duchi et al., 2011). Other hyperparameter values include: learning rate 0.01, d_m=50 for position embeddings $\mathbf{M}$, hidden size 300, batch size 50, filter width 3.

Baselines. (i) Decomposable Attention Model (Decomp-Att) (Parikh et al., 2016): Develop attention mechanisms to decompose the problem into subproblems to solve in parallel. (ii) Enhanced LSTM (ESIM) (Chen et al., 2017): Enhance LSTM by encoding syntax and semantics from parsing information. (iii) Ngram Overlap: An overlap baseline, considering unigrams, one-skip bigrams and one-skip trigrams. (iv) DGEM (Khot et al., 2018): A decomposed graph entailment model, the current state-of-the-art. (v) AttentiveConvNet (Yin and Schütze, 2017a): Our top-performing textual entailment system on SNLI dataset, equipped with RNN-style attention mechanism in convolution.[4]

In addition, to check if SCITAIL can be easily resolved by features from only premises or hypotheses (like the problem of SNLI shown by Gururangan et al. (2018)), we put a vanilla CNN (convolution&max-pooling) over merely hypothesis or premise to derive the pair label.

[3]Please refer to (Khot et al., 2018) for more details.

[4]https://github.com/yinwenpeng/Attentive_Convolution

Table 2 presents **results** on SCITAIL. (i) Our model DEISTE has a big improvement ($\sim$ 5%) over DGEM, the best baseline. Interestingly, AttentiveConvNet performs very competitively, surpassing DGEM by 0.8% on test. These two results show the effectiveness of attentive convolution. DEISTE, equipped with a parameter-dynamic convolution and a more advanced position-aware attentive convolution, clearly gets a big plus. (ii) The ablation shows that all three aspects we explore from the inter-sentence interactions contribute; "position" encoding is less important than "dyn-conv" and "representation". Without "representation", the system performs much worse. This is in line with the result of AttentiveConvNet baseline.

To further study the systems and datasets, Table 3 gives performance of DEISTE and baselines on SNLI. We see that DEISTE gets competitive performance on SNLI.

Comparing Tables 2 and 3, the baselines "hypothesis only" and "premise only" show analogous while different phenomena between SCITAIL and SNLI. On one hand, both SNLI and SCITAIL can get a relatively high number by looking at only one of {premise, hypothesis} – "premise only" gets 73.4% accuracy on SCITAIL, even higher than two more complicated baselines (ESIM-600D and Decomp-Att), and "hypothesis only" gets 68.7% accuracy on SNLI which is more than 30% higher than the "majority" and "premise only" baselines. Notice the contrast: SNLI "prefers" hypothesis, SCITAIL "prefers" premise. For SNLI, this is not surprising as the crowd-workers tend to construct the hypotheses in SNLI by some regular rules (Gururangan et al., 2018). The phenomenon in SCITAIL is left to explore in future work.

Error Analysis. Table 4 gives examples of errors. We explain them as follows.

Language conventions: The pair #1 uses dash "–" to indicate a definition sentence for "Front"; The pair #2 has "A (or B)" to denote the equivalence between A and B. This challenge is expected to be handled by rules.

Ambiguity: The pair #3 looks like having a similar challenge with the pair #2. We guess the annotators treat "$\cdots$ a vertebral column or backbone" and "$\cdots$ the backbone (or vertebral column)" as the same convention, which may be debatable.

Complex discourse relation: The premise in

#	(Premise P, Hypothesis H) Pair	G/P	Challenge
1	(P) Front – The boundary between two different air masses. (H) In weather terms, the boundary between two air masses is called front.	1/0	language conventions
2	(P) ... the notochord forms the backbone (or vertebral column). (H) Backbone is another name for the vertebral column.	1/0	language conventions
3	(P) ··· animals with a vertebral column or backbone and animals without one. (H) Backbone is another name for the vertebral column.	1/0	ambiguity
4	(P) Heterotrophs get energy and carbon from living plants or animals (consumers) or from dead organic matter (decomposers). (H) Mushrooms get their energy from decomposing dead organisms.	0/1	discourse relation
5	(P) Ethane is a simple hydrocarbon, a molecule made of two carbon and six hydrogen atoms. (H) Hydrocarbons are made of one carbon and four hydrogen atoms.	0/1	discourse relation
6	(P) ... the SI unit... for force is the Newton (N) and is defined as (kg·m/s^{-2}). (H) Newton (N) is the SI unit for weight.	0/1	beyond text

Table 4: Error cases of DEISTE in SCITAIL. "···": truncated text. "G/P": gold/predicted label.

	dev	test
Majority baseline	50.0	50.0
State-of-the-art	–	73.5
training data		
SNLI	47.1	46.0
SCITAIL	60.5	60.2

Table 5: Train on different TE datasets, test accuracy on two-way RTE-5. State-of-the-art refers to (Iftene and Moruz, 2009)

the pair #4 has an "or" structure. In the pair #5, "a molecule made of ···" defines the concept "Ethane" instead of the "hydrocarbon". Both cases require the model to be able to comprehend the discourse relation.

Knowledge beyond text: The main challenge in the pair #6 is to distinguish between "weight" and "force", which requires more physical knowledge that is beyond the text described here and beyond the expressivity of word embeddings.

Transfer to RTE-5. One main motivation of exploring this SCITAIL problem is that this is an end-task oriented TE task. A natural question thus is how well the trained model can be transferred to other end-task oriented TE tasks. In Table 5, we take the models pretrained on SCITAIL and SNLI and test them on RTE-5. Clearly, the model pretrained on SNLI has not learned anything useful for RTE-5 – its performance of 46.0% is even worse than the majority baseline. The model pretrained on SCITAIL, in contrast, demonstrates much more promising generalization performance: 60.2% vs. 46.0%.

4 Related Work

Learning automatically inter-sentence word-to-word interactions or alignments was first studied in recurrent neural networks (Elman, 1990). Rocktäschel et al. (2016) employ neural word-to-word attention for SNLI task. Wang and Jiang (2016) propose match-LSTM, an extension of the attention mechanism in (Rocktäschel et al., 2016), by more fine-grained matching and accumulation. Cheng et al. (2016) present a new LSTM equipped with a memory tape. Other work following this attentive matching idea includes Bilateral Multi-Perspective Matching model (Wang et al., 2017), Enhanced LSTM (Chen et al., 2016) etc.

In addition, convolutional neural networks (LeCun et al., 1998), equipped with attention mechanisms, also perform competitively in TE. Yin et al. (2016) implement the attention *in pooling phase* so that important hidden states will be pooled with higher probabilities. Yin and Schütze (2017a) further develop the attention idea in CNNs, so that a RNN-style attention mechanism is integrated into the convolution filters. This is similar with our position-aware attentive convolution. We instead explored a way to make use of position information of alignments to do reasoning.

Attention mechanisms in both RNNs and CNNs make use of sentence interactions. Our work achieves a deep exploration of those interactions, in order to guide representation learning in TE.

5 Summary

This work proposed DEISTE to deal with an end-task oriented textual entailment task – SCITAIL. Our model enables a comprehensive utilization of inter-sentence interactions. DEISTE outperforms competitive systems by big margins.

Acknowledgments. We thank all the reviewers for insightful comments. This research is supported in part by DARPA under agreement number FA8750-13-2-0008, and by a gift from Google.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*. pages 632–642.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. pages 49–54.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR* abs/1609.06038.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of ACL*. pages 1657–1668.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of EMNLP*. pages 551–561.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop*. pages 177–190.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan and Claypool.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* 12:2121–2159.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science* 14(2):179–211.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL*.

Adrian Iftene and Mihai Alex Moruz. 2009. UAIC participation at RTE5. In *Proceedings of TAC*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *Proceedings of AAAI*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*. pages 2278–2324.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*. pages 216–223.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. pages 3111–3119.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*. pages 2249–2255.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR* abs/1802.05365.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proceedings of ICLR*.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of NAACL*. pages 1442–1451.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of IJCAI*. pages 4144–4150.

Wenpeng Yin and Hinrich Schütze. 2017a. Attentive convolution. *CoRR* abs/1710.00519.

Wenpeng Yin and Hinrich Schütze. 2017b. Task-specific attentive pooling of phrase alignments contributes to sentence matching. In *Proceedings of EACL*. pages 699–709.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *TACL* 4:259–272.