Using Semantic Role Labeling to Combat Adversarial SNLI

Brett Szalapski
brettski@stanford.edu

Mengfan Zhang
zhangmf@stanford.edu

Miao Zhang
miaoz18@stanford.edu

Abstract

Natural language inference is a fundamental task in natural language understanding. Because of the understanding required to assess the relationship between two sentences, it can provide rich, generalized semantic representations. In this study, we implement a sentence-encoding model using recurrent neural networks. Our hypothesis was that semantic role labels, along with GloVe word embeddings, would give the sentences rich representations, making the model not only more successful on the original SNLI challenge, but also more robust to adversarial examples. However, our findings show that adding the SRL information does not improve the performance of our baseline model on either the SNLI task or the adversarial data sets.

1 Problem and Related Work

Natural language inference (NLI) is the problem of determining whether a hypothesis sentence H follows from a premise sentence P. NLI is a fundamental task in natural language understanding because of the understanding required to assess the relationship between two sentences. It has applications in many tasks, including question answering, semantic search, and automatic text summarizing. It is an ideal testing ground for theories of semantic representation, and the training for NLI tasks can provide rich, generalized semantic representations. NLI has been addressed using a variety of techniques, including symbolic logic, knowledge bases, and, in recent years, neural networks (Bowman et al., 2015). The landscape of NLI models is shown in Figure 1.

(Bowman et al., 2015) proposes a straightforward architecture of deep neural networks for NLI. In their architecture, the premise and the hypothesis are each represented by a sentence embedding vector. The two vectors are fed into a multi-layer neural network to train a classifier. It achieved an accuracy of 77.6% using LSTM networks on the SNLI corpus.

(Rocktschel et al., 2016) improves the aforementioned LSTM model by applying a neural attention model. The basic architecture is the same as (Bowman et al., 2015), which is based on sentence embeddings for the premise and the hypothesis. The key difference, which (Rocktschel et al., 2016) uses to improve performance, is that the embedding of the premise takes into consideration the alignment between the premise and the hypothesis. This attention-weighted representation of the premise improves the model performance to an accuracy of 83.5%.

One limitation of the model proposed by (Rocktschel et al., 2016) is that it reduces both the premise and the hypothesis to a single embedding vector before matching them. Thus, it uses two embedding vectors to perform sentence-level matching in the end. However, not all word or phrase-level matching results are equally important, and this model does not explicitly differentiate between good and bad matching results between the premise and the hypothesis. For example, matching of stop words is presumably less important than matching of content words. Additionally, some matching results may be particularly critical for making the final prediction. For example, a mismatch of the subjects of two sentences may be sufficient to indicate that they are not entailment, but this intuition is hard to capture if two sentence embeddings are matched in their entirety.

To address the limitations of the models proposed by (Bowman et al., 2015) and (Rocktschel et al., 2016), (Wang and Jiang, 2016) proposes a special LSTM-based architecture called match-LSTMs. Instead of using whole sentence embeddings for the premise and the hypothesis,

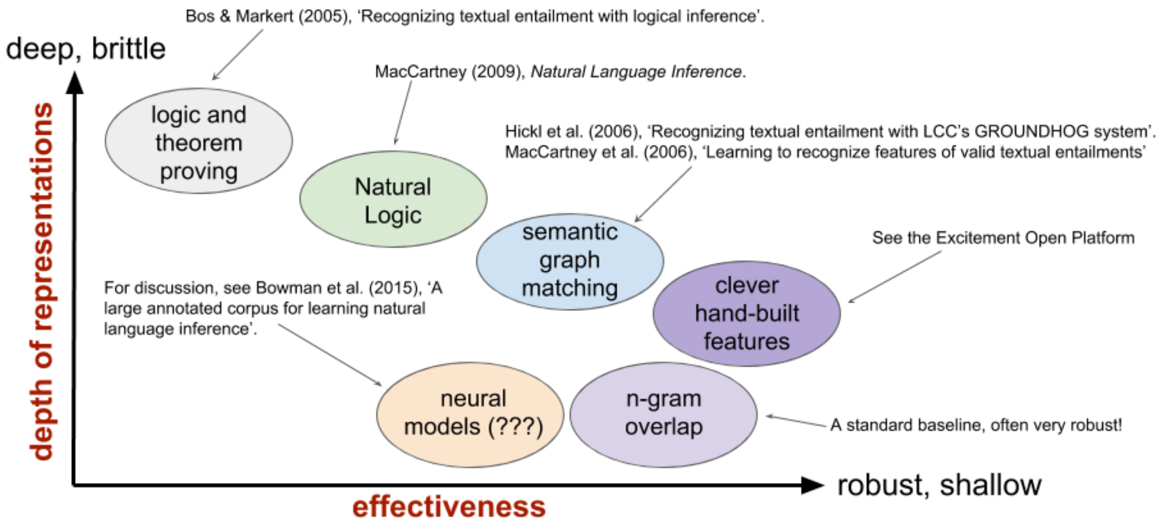
Bos & Markert (2005), 'Recognizing textual entailment with logical inference'.
deep, brittle
MacCartney (2009), *Natural Language Inference*.
logic and theorem proving
Natural Logic
Hickl et al. (2006), 'Recognizing textual entailment with LCC's GROUNDHOG system'.
MacCartney et al. (2006), 'Learning to recognize features of valid textual entailments'
semantic graph matching
See the Excitement Open Platform
clever hand-built features
For discussion, see Bowman et al. (2015), 'A large annotated corpus for learning natural language inference'.
neural models (???)
n-gram overlap
A standard baseline, often very robust!
depth of representations
effectiveness
robust, shallow

Figure 1: NLI model landscape

this model uses an LSTM to perform word-by-word matching between the hypothesis with the premise. The LSTM sequentially processes the hypothesis, matching each word in the hypothesis with an attention-weighted representation of the premise. This LSTM is able to place more emphasis on important word-level matching results. In particular, this LSTM remembers important mismatches that are critical for predicting the contradiction or the neutral relationship label. On the SNLI corpus, the match-LSTM architecture achieve an accuracy of 86.1%.

Different from (Wang and Jiang, 2016) using attention in conjunction with LSTMs, (Parikh et al., 2016) uses attention purely based on word embeddings. This model consists of feed-forward networks which operate largely independently of word order. Advantages of this model include the simple neural architecture and the way attention is used to decompose the problem into independently solvable sub-problems, facilitating parallelization. On the SNLI corpus, a new state-of-the-art was established at 86.8% accuracy, with almost an order of magnitude fewer parameters than the previous state-of-the-art, LSTMN (Cheng et al., 2016) and without relying on word-order.

The power of LSTMs and attention is well-known across a variety of tasks. However, one piece of the puzzle that most of the top results on the SNLI leaderboard share that these previous models do not have is the incorporation of pre-trained contextual word embeddings, such as ELMO or BERT. Combining these embeddings with a very deep network (Kim et al., 2018), with multitask learning (Liu et al., 2019), or with semantic knowledge (Zhang et al., 2018) leads to the best results.

Due to limited time and resources, the baseline for our NLI project is a pair of bidirectional LSTMs, one each for the premise and the hypothesis. Recurrent neural networks (RNNs) are a well-understood model for sentence encoding. They process input text sequentially and model the conditional transition between word tokens. The advantages of recursive networks include that they explicitly model the compositionality and the recursive structure of natural language, while the current recursive architecture is limited by its dependence on syntactic tree (Munkhdalai and Yu, 2017). In (Munkhdalai and Yu, 2017), a syntactic parsing-independent tree structured model, called Neural Tree Indexers (NTI), provides a middle ground between the sequential RNNs and syntactic tree-based recursive models. This model achieved the state-of-the-art performance on three different NLP tasks: natural language inference, answer sentence selection, and sentence classification. In (Chen et al., 2017), RNN-based sentence encoder equipped with intra-sentence gated-attention composition achieved the top performances on both the RepEval-2017 and the SNLI dataset.

Intuitively, including information about the sentence structure, such as part of speech or semantic role labels (SRL), should improve performance on NLI challenges. Several research teams have

A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

Table 1: Examples from SNLI dataset, shown with both the selected gold labels and the full set of labels (abbreviated) from the individual annotators.

found this to be true (Zhou and Xu, 2015; Shi et al., 2016). The SRL task is generally formulated as multi-step classification subtasks in pipeline systems, consisting of predicate identification, predicate disambiguation, argument identification, and argument classification (Zhang et al., 2018). An end-to-end system for SRL using deep bi-directional recurrent network is proposed by (Zhou and Xu, 2015). Using only the original text as input, this system outperforms the previous state-of-the-art model. Additionally, this model is computationally efficient and better at handling longer sentences than traditional models (Zhou and Xu, 2015).

2 Data

2.1 SNLI Dataset

The Stanford SNLI dataset (SNLI) is a freely available collection of 570,000 human-generated English sentence pairs, manually labeled with one of three categories: entailment, contradiction, or neutral. It constitutes one of the largest, high-quality, labeled resources explicitly constructed for understanding sentence semantics. SNLI is the basis for much of the recent machine learning research in the NLI field.

There was a longstanding limitation in NLI tasks that corpora are too small for training modern data-intensive, wide-coverage models. SNLI remedies this as a new, large-scale, naturalistic corpus of sentence pairs labeled for entailment, contradiction, and independence. The differences between SNLI and many other resources are as follow: At 570,152 sentence pairs, it is two orders of magnitude larger than the next largest NLI dataset. Its sentences and labels were written by humans in a grounded, naturalistic context rather than algorithmically generated; It uses a subset of the resulting sentences on validation task to provide a reliable set of annotations over the same data and to identify areas of inferential uncertainty (Bowman et al., 2015).

Amazon Mechanical Turk was used for data collection —workers were presented with premise scene descriptions from a preexisting corpus and were asked to supply hypotheses for each of three labels: entailment, neutral, and contradiction (Bowman et al., 2015). Each pair of sentences are possible captions for the same image. If the two are labeled for entailment, it means that the second caption is consistent with the information in the first. A label of contradiction indicates that the two captions cannot possibly label the same picture. A third class of neutral allows for independent captions that might coexist (Bowman et al., 2015). Table 1 shows a set of randomly chosen examples from the SNLI dataset. Both the selected gold labels and the full set of labels (abbreviated) from the individual annotators are described. A gold label means if any one of the three labels was chosen by at least three of the five annotators, then this label will be the gold label.

2.2 Adversarial Datasets

2.2.1 Compositionality-Sensitivity Test

NLI model should understand both lexical and compositional semantics. Adversarial datasets can be used to test whether the model can sufficiently capture the compositional nature of sentences (Nie et al., 2018). Two types of adversarial datasets—SOSWAP adversaries and ADDAMOD adversaries—were used to test the compositionality-sensitivity. Two examples for the two types of adversarial data are illustrated in

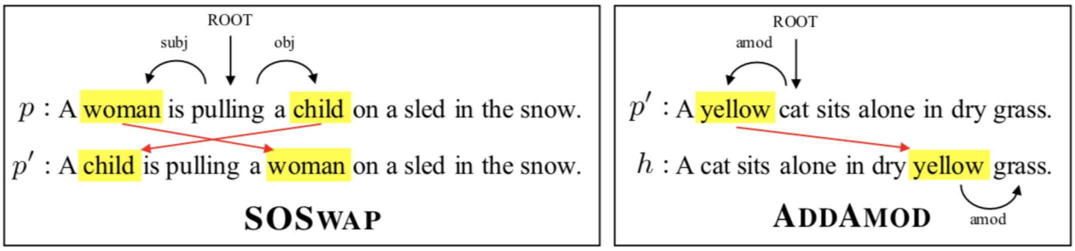
ROOT
subj
obj
p : A woman is pulling a child on a sled in the snow.
p′ : A child is pulling a woman on a sled in the snow.
SOSWAP
ROOT
amod
p′ : A yellow cat sits alone in dry grass.
h : A cat sits alone in dry yellow grass.
amod
ADDAMOD

Figure 2: Examples of SOSWAP and ADDAMOD Adversarial Data (Nie et al., 2018). On the left, the swapped subject and object are marked in yellow in p'. On the right, the added adjective modifier is marked in yellow in h.

Figure 2. In SOSWAP, the subject and object of a sentence are switched. In ADDAMOD, an adjective is moved from one noun to another. The semantics of the sentences are modified through perturbing the compositionality without changing any lexical features. The intuition behind the adversarial datasets is that, while the semantic difference resulting from compositional change is obvious for humans, the two input sentences will be almost identical for models that take no compositional information into consideration. Therefore, by running the model on the adversarial test sets, we can evaluate whether the model is able to consider compositional information. Additionally, there are 971 SOSWAP examples—most of which are contradictions—and 1,783 ADDAMOD examples—most of which are neutral—in this data set.

2.2.2 Generalization Ability Test

The data created by (Glockner et al., 2018) contains one additional type of adversarial example. These examples are formed by taking a premise from the SNLI data set and replacing one word with either a synonym, hyponym, antonym, or co-hyponym. The first two create an entailment example, and the latter two create a contradiction. Table 2 shows examples from the adversarial dataset, where the examples can capture various kinds of lexical knowledge. This dataset can be used to assess the lexical inference abilities of NLI systems, and it is available at `https://github.com/BIU-NLP/Breaking_NLI`.

All of the replacement words are present in the SNLI data set and in the pre-trained GLoVe embeddings used. This set consists of 7,164 contradiction examples, 982 entailment examples, and 47 neutral examples.

Premise/Hypothesis	**Label**
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Table 2: Examples from Breaking NLI dataset

3 Methodology

3.1 Sentence-encoding RNNs

SNLI is suitably large and diverse to make it possible to train neural network models that produce distributed representations of sentence meaning (Bowman et al., 2015).

Sentence embedding is used as an intermediate step in the NLI classification and producing sentence representations. First, vector representations for each of the two sentences are produced, then these two vectors are passed to a linear classifier, which predicts the label for the pair.

Our recurrent neural network classifier, depicted in Figure 3, processes the premise and hypothesis with separate RNNs and uses the concatenation of their final states as the basis for the classification decision at the top. Words are embedded using 100-dimensional GloVe embeddings and processed sequentially in a BiDirectional LSTM with a hidden dimension of 50. The premise and hypothesis final states are concatenated and passed to a softmax layer for classification. Much of code for this baseline model comes from (Potts, 2019).

The model is trained on the SNLI Training set and evaluated on the SNLI Test set, as well as the three adversarial sets described previously: ADDAMOD, SOSWAP, and BreakingNLI. Due to the

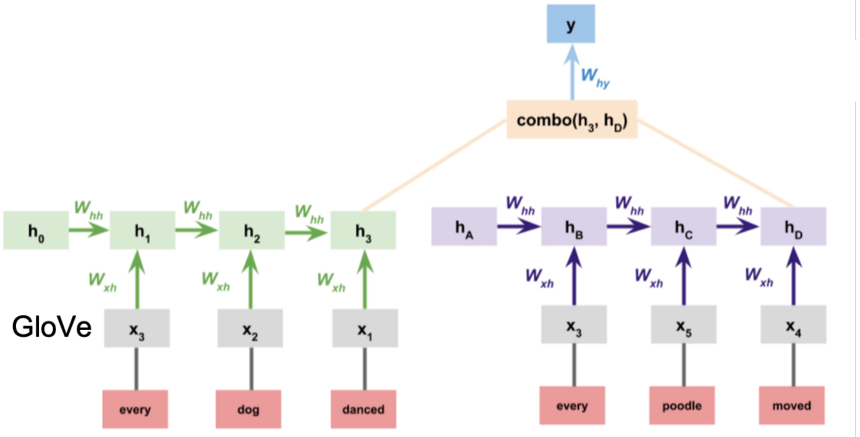

Figure 3: Architecture of the Sentence-encoding Baseline Model (Potts, 2019)

overhead required for extracting the SRL tags, the SNLI training set was reduced to 267,379 examples. For the other data sets, including SNLI dev and test, all examples were pre-processed and used for evaluation.

3.2 Semantic Role Labeler

Semantic role labeling, which is a technique that has been used in state-of-the-art SNLI models (Zhang et al., 2018), encodes important grammatical aspects of a sentence that go beyond simple part-of-speech tagging or word embeddings. SRL can lead not only to top SNLI performance, but may also make a model more robust to adversarial modifications to the SNLI data set. Given a sentence, the task of semantic role labeling is dedicated to recognizing the semantic relations between the predicates and the arguments. For example, given the sentence, *Charlie sold a book to Sherry last week*, where the target verb (predicate) is *sold*, SRL yields the following outputs,

$$[_{ARG0}\ \text{Charlie}]\quad [_{V}\ \text{sold}]\quad [_{ARG1}\ \text{a book}]$$
$$[_{ARG2}\ \text{to Sherry}]\quad [_{AMTMP}\ \text{last week}]$$

where *ARG0* represents the seller (agent), *ARG1* represents the thing sold (theme), *ARG2* represents the buyer (recipient), $AM\ TMP$ is an adjunct indicating the timing of the action and V represents the predicate. (Zhang et al., 2018)

The state-of-the-art SRL module implemented by (Zhang et al., 2018) consists of an embedding layer, which includes ELMO and PIE embeddings; an 8-layer, interleaved, Bi-Directional LSTM with highway connections and dropout; and a softmax output layer which predicts SRL tags (Zhang et al., 2018). This architecture can be seen in Figure 4.

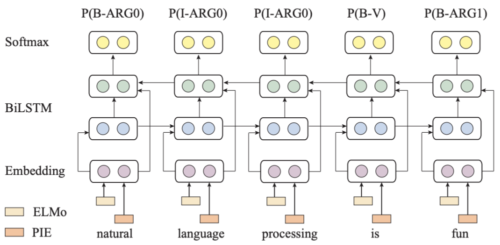

Figure 4: Architecture of the Semantic Role Labeler

In our work, we make use of the readily available SRL model from AllenNLP. Each sentence is parsed into tokens according to the AllenNLP `WordTokenizer`, and annotated with the SRL tags for each token. The words are embedded with 50-dimensional GloVe embeddings, and the SRL tags are included in a 50-dimensional embedding structure. Each word is represented as the concatenation of its GloVe and SRL embeddings. The resulting representation for each word in the input is then a 100-dimensional embedding, just as in the baseline model. See Figure 5 for the enhanced architecture.

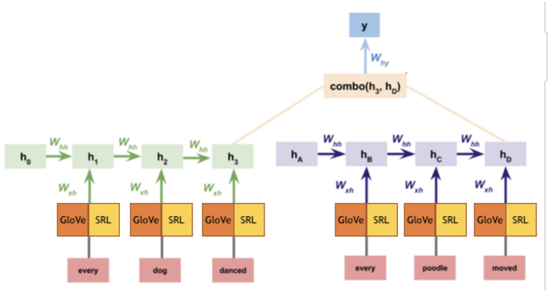

Figure 5: Architecture of the SRL-Enhanced Model

Table 3: Baseline and SRL Results on SNLI.

Category	Baseline Precision (%)	SRL Precision (%)	Baseline F1-Score	SRL F1-Score
Contradiction	0.373	0.314	0.326	0.010
Entailment	0.426	0.353	0.490	0.389
Neutral	0.467	0.334	0.429	0.426
Macro-Average	0.422	0.334	0.415	0.275

Table 4: Baseline and SRL Results on Adversarial Datasets.

Dataset	Prominent Label	Baseline Recall	SRL Recall
ADDAMOD	Neutral	0.920	0.864
SOSWAP	Contradiction	0.076	0.083
Breaking NLI	Macro-Average	0.365	0.319

In order to save time during training and iterating, we used a pre-processing script that ran each SNLI example's `sentence1` and `sentence2` through the AllenNLP SRL module and saved the tags for later use.

4 Results and Discussion

We first evaluated both the baseline LSTM and the SRL-augmented models against the SNLI task to determine whether standard SNLI performance improves as a result of SRL. These results can be seen in Table 3. Next, both systems were evaluated on the SOSWAP dev set, ADDAMOD dev set, and the Breaking NLI adversarial set. Since the ADDAMOD and SOSWAP sets consist of primarily one label, we evaluate performance of the two systems based on the recall for only Neutral and Contradiction classifications, respectively. These results are shown in Table 4.

The results show that the model with SRL embeddings actually performed worse than the model using only GloVe embeddings. However, we believe this is less due to the inability of SRL to enhance a model, and more to do with this particular architecture. One of the key drawbacks of the baseline model is that it does not include an attention mechanism. Adding attention to the model would likely allow the semantic role labels from the premise and the hypothesis to interact with each other similarly to the way in which cross-sentence word attention would. However, the SRL-enhanced embeddings would help to further cement connections or contradictions between the two sentences. For example, if a word and its synonym were in the premise and the hypothesis, and they shared a semantic role label, this cross-sentence attention would likely classify these sentences correctly. Further, if a word appeared in both the premise and the hypothesis, but it had unrelated semantic role labels, the attention mechanism may be able to differentiate between these meanings, improving upon models which overemphasize word co-occurence.

Another avenue to explore would be the embedding style. In this study, we ensured that the embedding dimension for each word was the same in both the baseline and the SRL-enhanced model. In the former, this meant using 100-dimensional GloVe embeddings, while in the latter, we used 50-dimensional GloVe embeddings and 50 dimensions for the SRL tag embedding. In hindsight, this gave our baseline model much more representational power for the words themselves, and may have put the SRL-enhanced model at a relative disadvantage.

5 Conclusion and Future Work

Although Semantic Role Labels have proven to be a useful feature in some NLI models, for our parallel, bidirectional LSTMs with GloVe word embeddings, SRL decreased performance both in the nominal SNLI evaluations, as well as in adversarial data sets. However, we believe that SRL could give the right model significant performance gains not only on the SNLI test set, but also make the model more robust to adversarial NLI examples. We recommend exploring the performance of the model in (Zhang et al., 2018), in particular against the adversarial data sets described here.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Recurrent neural network-based sentence encoder with gated attention for natural language inference. *arXiv preprint arXiv:1708.01353*.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *CoRR*, abs/1805.02266.

Seonhoon Kim, Jin-Hyuk Hong, Inho Kang, and Nojun Kwak. 2018. Semantic sentence matching with densely-connected recurrent and co-attentive information. *CoRR*, abs/1805.11360.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural tree indexers for text understanding. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 1, page 11. NIH Public Access.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of NLI models. *CoRR*, abs/1811.07033.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

Christopher Potts. 2019. cs224u. `https://github.com/cgpotts/cs224u`.

Tim Rocktschel, Edward Grefenstette, Karl Moritz Hermann, Tom Koisk, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. *ICLR*.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. 2016. Knowledge-based semantic embedding for machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2245–2254, Berlin, Germany. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. *CoRR*, abs/1512.08849.

Zhuosheng Zhang, Yuwei Wu, Zuchao Li, Shexia He, Hai Zhao, Xi Zhou, and Xiang Zhou. 2018. I know what you want: Semantic learning for text comprehension. *CoRR*, abs/1809.02794.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *ACL*.