Learning Directional Sentence-Pair Embedding for Natural Language Reasoning

Yuchen Jiang,[12] **Zhenxin Xiao,**[12] **Kai-Wei Chang,**[1]
[1]University of California, Los Angeles, [2]Zhejiang University
jyc@zju.edu.cn, alanshawzju@gmail.com, kwchang@cs.ucla.edu

Abstract

Enabling the models with the ability of reasoning and inference over text is one of the core missions of natural language understanding. Despite deep learning models have shown strong performance on various cross-sentence inference benchmarks, recent work has shown that they are leveraging spurious statistical cues rather than capturing deeper implied relations between pairs of sentences. In this paper, we show that the state-of-the-art language encoding models are especially bad at modeling **directional** relations between sentences by proposing a new evaluation task: Cause-and-Effect relation prediction task. Back by our curated Cause-and-Effect Relation dataset ($\mathcal{CER}$), we also demonstrate that a **mutual attention** mechanism can guide the model to focus on capturing directional relations between sentences when added to existing transformer-based models. Experiment results show that the proposed approach improves the performance on downstream applications, such as the abductive reasoning task.

Introduction

Reasoning over texts is regarded as a main challenge in artificial intelligence. This task aims at identifying relations (e.g., cause-and-effect) between sentences requires understanding commonsense knowledge, and reasoning over correspondence between words in both sentences.

Recently, significant progress has been made by language encoder techniques, such as ELMo (Peters et al. 2018), OpenAIs GPT (Radford 2018), and BERT (Devlin et al. 2018). These models learn to encode words and sentences into fixed-length dense vectors and they capture the association between words and phrases by training on large text corpus. They have been shown to improve a variety of downstream tasks including language inference tasks (e.g., SNLI, MNLI). However, recent work (McCoy, Pavlick, and Linzen 2019), (Niven and Kao 2019) has shown that though current models perform well on cross-sentence inference benchmarks such as MNLI (Williams, Nangia, and Bowman 2017) and SNLI, they are leveraging spurious statistical cues rather than capturing deeper associations between pairs of sentences.

In this paper, we show that current language encoder techniques are especially bad at modeling **directional** relationships between sentences. For example, consider the following sentence pair:

S_1*: PersonX fails PersonX's math class.*

S_2*: PersonX must wait another semester to graduate.*

We observe that while the state-of-the-art language encoding models (such as BERT) can identify the second sentence is the "effect" of the first one, it fails to recognize the first sentence is the "cause" of the second one. We suspect that this asymmetry of predicting power is due to the language encoder that mostly trained based on knowledge within sentences does not learn to capture relations across sentences well.

To formally evaluate the capacity of models in capturing directional relations, we collect a new Cause-and-Effect corpus $\mathcal{CER}$ that consists of 784,075 sentence pairs. The corpus serves as an evaluation task as well as a new resource for learning casual relations between sentences. We further extend a transformer-based language encoding model, BERT, by introducing a mutual attention mechanism to better capture directional relations. Results show that the proposed approach improves the accuracy of recognizing cause-and-effect relation from 80.9% to 86.7%.

We further evaluate the resulting sentence-pair embeddings in the abductive reasoning task (aNLI) (Bhagavatula et al. 2019), which is a benchmark dataset requiring cause-and-effect commonsense. Formulated as a binary-classification task, the goal is to pick the most plausible explanatory hypothesis given two observations from narrative contexts. We demonstrate that the proposed embedding model achieves better performance on aNLI compared to baseline models. Ablation study shows that both pretraining on Cause-and-Effect relation prediction task and mutual attention mechanism contribute to the gain on aNLI.

Directional Relation Prediction

Dataset Learning directional relations are important for natural language reasoning, and Cause-and-Effect is the most common and extensive one in the real world. We curate a new sentence-pair dataset for studying natural language reasoning, called $\mathcal{CER}$. Our task is to predict the casual rela-

Dataset	Model	Origin	Test-Augment	Augment
$\mathcal{CER}$	Tr-fullly	67.4/73.5	33.6/41.2	54.5/63.3
	Tr-fullly w/ SE	73.2/84.5	46.9/61.5	71.2/80.9
	Tr-sub+InferSent	80.7/84.4	69.6/72.3	72.1/77.2
	Tr-sub&mut	82.0/86.2	**71.3/78.1**	**78.4/86.7**
ATOMIC	Tr-fullly	80.5/86.0	41.7/52.8	50.5/58.1
	Tr-fullly w/ SE	80.7/86.2	60.7/67.4	70.1/77.9
	Tr-sub+InferSent	79.1/85.6	69.8/74.2	77.7/84.5
	Tr-sub&mut	80.6/85.2	**70.2/77.8**	**82.6/86.1**

Table 1: Average F1 score and overall accuracy. **Origin** means models are both trained and test on the original dataset where the set of S_1 and the set of S_2 have no overlap. **Test-Augment** means models are trained on the original dataset but tested on augmented test dataset, where for every sentence pair (S_1, S_2) we add a new instance (S_2, S_1) by reversing the sentence order and swap the label from cause to effect and vise versa. **Augment** means models are both trained and test on augmented dataset.

tion between two sentences. Specifically, for every sentence-pair, there are three labels: *isCauseOf*, *isResultOf*, *noRelation*. The model must correctly predicts the corresponding label.[1] Except for $\mathcal{CER}$, we also extract a subset of ATOMIC (Sap et al. 2019) which contains directional relations.

Results We compared mutual attention mechanism with various different strategies to encode directional information: **Tr-fully** is an fully-connected transformer structure without encoding directional information. **Tr-fully w/ SE** is the original transformer structure which encodes directinal information using segment embeddings (concatenating segment embeddings with token & position embeddings). **Tr-sub+InferSent** construct the final feature vector, which consisting of the concatenation of two hidden states corresponding to the first token of both sentences, their difference, and their elementwise product (Conneau et al. 2017). **Tr-sub&mut** is a compositional model that combines subgraph attention and mutual attention. We observe that Tr-sub&mut performs as well or better than other approaches on both F1 score and Acc in each setting. It is worth noting that although when the set of S_1 and S_2 have no overlap (a bigraph), the performance of Tr-fully w/ SE is comparable with Tr-sub&mut, when tested on augmented test set, Tr-sub&mut outperforms Tr-fully w/ SE by a large margin, suggesting that the latter is likely to rely more on the co-occurrence of two sentences. The empirical results also show that even when trained on augmented dataset, the mutual attention mechanism continues outperforming other methods of combining directed information.

Experiments on aNLI

To evaluate whether the directional information contained in the resulting sentence pair embeddings learned on $\mathcal{CER}$ and ATOMIC are transferable, we report the result when transferred to Abductive Natural Language Inference (aNLI) by using $\mathcal{CER}$ and ATOMIC as supplementary training phase. We observe that both removing pertaining on directional relationship prediction tasks and removing mutual attention mechanism degrades performance on aNLI. The results suggests that 1) directional relations captured by models are beneficial for abductive reasoning; 2) mutual attention helps models better capture transferable directional relations.

[1]Examples and Data Construction for $\mathcal{CER}$ is shown in Appendix.

Model	BERT	BERT-InferSent	BERT-Mut
Finetuning	62.79	60.40	63.62
+ ATOMIC	63.64	62.13	66.31
+ CERP	64.66	63.89	65.23

Table 2: Accuracy on aNLI dataset. Fine-tuning on both ATOMIC and CERP provide a gain on aNLI, and BERT with mutual attention mechanism obtains the largest gain.

Conclusion

This paper studies the effects of training directional sentence pair embeddings. Future work includes a more extensive directional relation dataset and more extensive transfer tasks.

References

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, S. W.-t.; and Choi, Y. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; and Bordes, A. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

McCoy, R. T.; Pavlick, E.; and Linzen, T. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*.

Niven, T., and Kao, H.-Y. 2019. Probing neural network comprehension of natural language arguments. In *ACL*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Radford, A. 2018. Improving language understanding by generative pre-training.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.