Directional Sentence-Pair Embedding for Commonsense Causal Reasoning

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

Abstract

Enabling machines with the ability of reasoning and inference over text is one of the core missions of natural language understanding. Although deep learning models have shown strong performance on various cross-sentence inference benchmarks, recent work has shown that they tend to leverage spurious statistical cues rather than capturing deeper 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. To remedy this issue, we incorporate a mutual attention mechanism with a transformer-based model to better capture directional relations between sentences. We further curate $\mathcal{CER}$, a Cause-and-Effect Relation corpus, to facilitate the model embeds commonsense casual relations in sentence representations. Experiment results show that the proposed approach improves the performance on downstream applications, such as the abductive reasoning task.

1 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, and it requires understanding casual commonsense and capturing correspondence between sentences.

Recently, significant progress has been made by language encoder techniques, such as ELMo (Peters et al., 2018), OpenAIs GPT (Radford et al., 2018), BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). These models learn to encode words and sentences into fixed-length 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 (Bowman et al., 2015), MNLI (Williams et al., 2017)). However, recent work (McCoy et al., 2019; Niven and Kao, 2019) shows that though current models perform well on cross-sentence inference benchmarks, they leverage spurious statistical cues rather than capturing deeper associations between sentence pairs. As a result, the model fails to perform well on debiased adversarial sets.

In this paper, we extend these studies and analyze how well current language encoding techniques capture commonsense casual relations. We found that while contextualized representation models are capable of capturing associations between words, they have difficulty in capturing directional relations 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.

The language encoders (e.g., BERT, RoBERTa) are able to recognize $S2$ is the "effect" of $S1$ but fails to recognize $S1$ is the "cause" of $S2$. Besides, these models often predict S1 is the "effect" of and is the "cause" of S2 at the same time, demonstrating that they are not modeling the directional relations well. To formally evaluate the capacity of models in capturing directional relations, we collect $\mathcal{CER}$, a new Cause-and-Effect Relation corpus that consists of 784,075 sentence pairs extracted from English Wikipedia, BookCorpus and 1.46GB online novels.[1] The corpus serves as an evaluation task as well as a new resource for learning casual relations between sentences.

In addition, we extend transformer-based lan-

* Contribution during internship at UCLA.

[1] We will release the source code and dataset once the paper is accepted.

guage encoding models, BERT, and RoBERTa by introducing a mutual attention mechanism to better capture the directional relations between sentences. This approach models two directions separately and thus disentangle the directional information from co-occurrence. Results show that the proposed approach improves the accuracy of recognizing cause-and-effect relation by 6.1% on average.

We further evaluate the resulting sentence-pair embeddings on 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 baselines. Ablation study shows that both pretraining on Cause-and-Effect relation prediction task and mutual attention mechanism contribute to the performance gain.

The main contributions of this paper are:

- We analyze the inadequate capability of current language encoder techniques at modeling *directional* relations between sentences;
- We curate a new dataset $\mathcal{CER}$ specifically aiming at helping models to learn directional sentence-pair relations, which can also be used for evaluation;
- We propose a mutual attention mechanism and empirically demonstrate the effectiveness of mutual attention in distinguishing directional relations;
- We enhance the language encoders by the proposed approach and the $\mathcal{CER}$ corpus and demonstrate that they are beneficial to the downstream abductive reasoning task.

2 Related Work

2.1 Sentence Embeddings

There are many works in the line of sentence embedding learning that specifically target at constructing generic sentence representation. Skip-Thoughts (Kiros et al., 2015) and Quick-Thoughts (Logeswaran and Lee, 2018) learn unsupervised sentence embeddings by predicting the surroundings sentences of a given sentence. InferSent (Conneau et al., 2017) uses the Stanford Natural Language Inference (SNLI (Bowman et al., 2015)) Corpus (a set of 570k pairs of sentences labeled with 3 categories: neutral, contradiction and entailment) to train a classifier on top of a sentence encoder. Both sentences are encoded using the same encoder while the classifier is trained on a pair representation constructed from the two sentence embeddings. Nie et al. (2019) and Jernite et al. (2017) demonstrate that discourse-based objectives can also be leveraged to learn good sentence representations. Subramanian et al. (2018) and Cer et al. (2018) leverage a multi-tasking learning framework to learn a universal sentence embedding by switching between several tasks, including Skip-thoughts' prediction of the next/previous sentence, neural machine translation, constituency parsing, and natural language inference.

The Cause-and-Effect Relation prediction task proposed by us is also a classification task, which is a bit similar to the SNLI task and Nie et al. (2019)'s discourse marker prediction task. However, unlike our casual prediction task, these tasks do not try to distinguish the different relations of two sentence pairs which consist of the same two sentences but have opposite order, so *co-occurrences* instead of *implied directional relations* are learned. The $\mathcal{CER}$ dataset proposed in this work, instead, contain relations with two opposite directions; therefore it facilitates the model learn the directions better.

2.2 Commonsense Casual Reasoning

Commonsense causal reasoning is the process of capturing and understanding the causal dependencies amongst events and actions. Such events and actions can be expressed in terms, phrases or sentences in natural language text. Roemmele et al. propose the Choice Of Plausible Alternatives (COPA) evaluation task, which consists of 1,000 English-language questions that directly assess commonsense causal reasoning. Each question gives a premise and two plausible causes or effects, where the correct choice is the alternative that is more plausible than the other. However, this dataset is very small and is insufficient for training directional sentence-pair embeddings. In contrast, our curated dataset is substantially larger, which contains 7 million sentence pairs (784 times of COPA).

2.3 Transformer-based models

The Transformer introduced by Vaswani et al., which uses the self-attention mechanism, has proven to be especially effective for common natu-

Relation	Wiki	Book-Corpus	Online Novels	Total
isCauseOf	30,865	123,532	492,342	646,739
isEffectOf	9,601	43,046	57,282	109,929

Table 1: Statistics of $\mathcal{CER}$: Number of sentence-pairs with different labels.

Marker	Wiki	Book-Corpus	Online Novels	Total
because	10,014	45,526	59,533	115,073
Because	3,261	9,712	23,218	36,191
Since	5,541	7,122	83,324	95,987
so	11,194	75,768	237,419	324,381
So	2,358	38,684	75,106	116,148
Therefore	1,212	2,100	50,346	53,658
Hence	526	591	23,345	24,462
As a result	3,909	499	10,385	14,793
Consequently	1,140	353	1,516	3,009
For this purpose	65	5	40	110
To this end	191	31	41	263
Total	**39,411**	**180,391**	**564,273**	**784,075**

Table 2: Statistics of $\mathcal{CER}$: Number of sentence-pairs extracted from different datasets using different cohesive devices. Reversing/negative sampling doubles the number.

ral language processing tasks. GPT (Radford et al., 2018) uses a left-to-right Transformer while BERT (Devlin et al., 2018) builds on fully-connected Transformer networks to pretrain bidirectional (in fact all-directional) representations and improved the state-of-the-arts on a range of NLP benchmarks (Wang et al., 2018). When directly fine-tuning transformer-based models on sentence-pair tasks, segment embedding is used as a signal to distinguish different sentences when constructing input representations. However, the latest work, such as spanBERT (Joshi et al., 2019), RoBERTa (Liu et al., 2019), finds that pretraining on single segments, instead of two half-length segments with the next sentence prediction (NSP) objective, improves performance on most downstream tasks, including sentence-pair tasks. This can't help but make people think: for the sentence-pair tasks (or other multi-sentence tasks), is there a better form to characterize the relations between the sentences?

3 $\mathcal{CER}$: Cause-and-Effect Dataset

We curate a new sentence-pair dataset for studying natural language reasoning, called $\mathcal{CER}$. The task is to predict the casual relation between two sentences. Formally, for every sentence-pair (S_1, S_2), The relation between S_1 and S_2 is denoted as $R(S_1, S_2)$. For every two sentences in $\mathcal{CER}$, there are three possible relations: *isCauseOf*, *isEffectOf* , *noRelation*. And it is worth noting that *isCauseOf* and *isEffectOf* are relations with opposite directions. That is if $(S_1, S_2, isCasueOf)$ holds, $(S_2, S_1, isResultOf)$ also holds. As shown in Table 3 for example sentence pairs corresponding to different relations (S_1, S_2, r). The model must then select the most appropriate directional relation of (S_1, S_2). To the best of our knowledge, $\mathcal{CER}$ is the first large-scale natural language casual dataset.[2]

Data Sources Our corpus consists of 784,065 sentences, derived from English Wikipedia[3], BookCorpus[4](Zhu et al., 2015) and 384 online novels (1.46GB) from Read Novel Full website[5]. These datasets are slightly different in nature and allow us to achieve broader coverage: English Wikipedia contains formal English and well-defined commonsense; BookCorpus contains daily languages and daily commonsense; online novels contain more verbal and non-formal English. Table 1 presents some statistics for each part of the dataset.

Data Construction In the following, we present an automatic process to collect a large dataset of sentence pairs which have Cause-and-Effect relationships from natural text corpora.

We first leverage a set of cohesive devices and phrases (e.g., because) which indicate cause-and-effect relationships to generate a set of templates with the help of Stanford Parser (Manning et al., 2014). We noticed that some previous work (Nie et al., 2019) which deals with cohesive devices rely on dependency parsers for extraction. However, the use of dependency parser introduces noise into the dataset, sacrificing the quality of extracted sentences. Instead of using dependency parser, we mainly rely on templates and a set of rules.

In many cases, a given template has multiple versions. For example, for "A because B." "A probably because B." is an alternative form. We collected a list of derivations for each cohesive devices to generate variations of the surface form based on regular expression and POS tagging (",[RB] because" or ",[IN] because" in this case). Similarly,

[2]Although COPA contains casual sentence pairs, the sentences in COPA are very short and this dataset only contains 1000 samples, which is insufficient for training large sentence encoders.

[3]en.wikipedia.org/wiki/Wikipedia:Database_download

[4]We crwal BookCorpus data using the script in www.github.com/soskek/bookcorpus, and get 3645 books in total from www.smashwords.com.

[5]www.readnovelfull.com

Sentence 1	Sentence 2	Relation
He had taught thousands of kids to speak basic English.	He liked teaching children abroad.	*isCauseOf*
He liked teaching children abroad.	He had taught thousands of kids to speak basic English.	*isEffectOf*
He had taught thousands of kids to speak basic English.	He knocked the door.	*noRelation*

Table 3: Examples from $\mathcal{CER}$. The first sentence pair is collected from online novels. The second sentence pair is generated by reversing direction of the first sentence pair; the third is generated by negative sampling.

we design regular expression to filter out some sentence pairs despite that they contain the targeted cohesive device. For example, "A, because of B." is excluded from the isCauseOf relation as A is not the cause of B. We also need to make sure that A is not a fragment (e.g. "In addition"). The full list of templates along with detailed extracting rules can be found in the Appendix.

Because this method is fully-automatic and it makes no assumption on source corpus, it can be applied on any natural text corpora to extract more casual sentence pairs.

After extracting sentence pairs from natural text corpora, we further augment the dataset by 1) reversing the sentence pairs to generate a dual set (e.g., *isCauseOf*$(S_1, S_2) \leftrightarrow$ *isEffectOf* (S_2, S_1)) and 2) generating negative samples by randomly selecting sentences from other sentence-pairs to construct *noRelation* class. We randomly divide the dataset into train/validation/test set with 90%, 5%, 5% split.

4 Mutual Attention as Directional Relation Encoder

In the following, we discuss how to extend a transformer-based model to better capture directional sentence relations.

Fully-connected Transformer heads treat every token equally by nature, regardless of which sentence (segment) they belong. Different transformer-based models have different strategies of encoding segment information: BERT embeds segment information via a special token <sep> and segment embedding; RoBERTa abandons the reluctant segment embedding of BERT (Figure 1 (a)). When single sentence encoders are transferred to sentence pair tasks, a concatenation of both sentence embeddings (along with their product and difference) is usually used as the representation for the whole sentence pair (Figure 1 (b)). However, as our probing shows, both approaches are inefficient in distinguishing the directions of the relations between sentences and do not have good transferability for natural language reasoning since this kind of advanced tasks need models to be aware of the asymmetry of the relations between sentences. We propose to adopt the mutual attention mechanism to encode the inter-sentence direction information (Figure 1 (c)).

Formally, given a pair of sentences (S_1, S_2), each sentence is a sequence of words or sub-words $S_i = (x_1^i, x_2^i, ..., x_{n_i}^i)$ from vocabulary $\mathbf{V}$. The set of all tokens is denoted as X. The relation between S_1 and S_2 is denoted as $R(S_1, S_2)$ while $R(S_2, S_1) = \overline{R}(S_1, S_2)$.

Backbone Transformer Networks Given an input sentence S_i, an L-layer Transformer is used to encode the input:

$$H^l = Transformer_l(H^{l-1})$$

where $l \in [1, L]$, and $H^l = [h_1^l, h_2^l, ...]$. The hidden vector h_j^l is the l-th layer's contextualized representation of the j-th token in the input sentence.

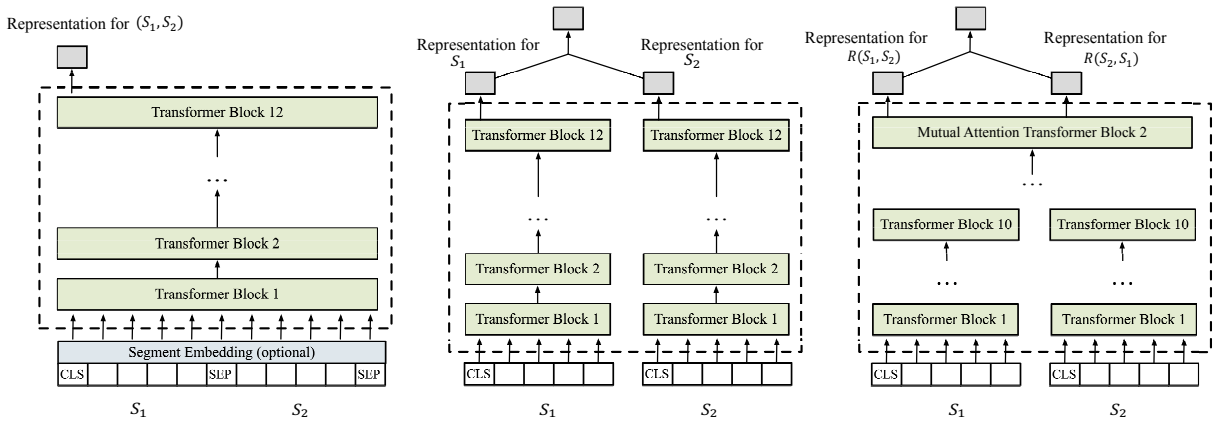
In each Transformer block, there are multiple self-attention heads used to aggregate the output vectors of the previous layer. For the l-th Transformer layer, the output of an original self-attention head A_l is computed via:

$$\begin{aligned} Q = H^{l-1}W_l^Q, K = H^{l-1}W_l^K, V = H^{l-1}W_l^V \\ A_l = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \end{aligned} \quad (1)$$

where W_l^Q, W_l^V, W_l^K are the linear transformations to get the query matrix Q, the key matrix K and the value matrix V in the l-th layer, and d_k is the dimension of the rows in K.

Mutual Attention In the mutual attention head, Q and K, V in Eq. (1) are no loner transformed from the same hidden vectors. Therefore, instead, the attention head A_l is computed via:

$$\begin{aligned} A_l(S_1, S_2) = softmax\left(\frac{Q_1 K_2^T}{\sqrt{d_k}}\right)V_2 \\ A_l(S_2, S_1) = softmax\left(\frac{Q_2 K_1^T}{\sqrt{d_k}}\right)V_1 \end{aligned} \quad (2)$$



(a) Standard BERT/RoBERTa (b) Single Sentence Encoding Models (c) Ours

Figure 1: Overview of the framework. (a) is the standard BERT/RoBERTa structure, the hidden states corresponding to the first token often used to represent the whole sentence pair. (b) is the standard single sentence representation encoding structure. (c) is the framework proposed by us: Lower transformer layers are subgraph attention layers while top layers are mutual attention layers, where tokens in S_1 and tokens in S_2 pay attention to each other rather than to all tokens.

Head Type	Complexity per head
Fully-connected Attention	$O(n^2 \cdot d)$
Mutual Attention	$O(1/2 \cdot n^2 \cdot d)$

Table 4: Mutual attention mechanism reduces half of the computational complexity compared to the fully-connected self-attention model. d is the dimension of hidden vectors and n is the number of layers.

More specifically, the aggregating rule for every hidden state in layer i is shown as following:

$$h_{mul}^{l} \leftarrow \sum_{k \in X \setminus \mathcal{N}(u)} \frac{(W_l^Q h_u^{l-1})^T W_l^K h_k^{l-1}}{\sqrt{d_k}} W_l^V h_k^{l-1} + h_u^{l-1} \tag{3}$$

where $\mathcal{N}(u)$ (the neighborhood of x_u) is the set of all the tokens in S_i ($x_u \in$S$_i$).

We set the attention heads at lower levels to be single-sentence (only do self attention within once sentence) and adopt mutual attention at top layers, since the single-sentence attention serves as intra-sentence information fusion while mutual attention focuses on inter-sentence relations. It is also worth noting that computing mutual attention instead of fully-connected attention reduces half the computational complexity (see Table 4).

5 Experiments

We first evaluate the effectiveness of mutual attention on both our curated $\mathcal{CER}$ dataset and a subset of ATOMIC (Sap et al., 2019), where the directional relations between pair of sentences is temporal precedence. Then we evaluate the generalization performance of this self-supervised method on the abductive natural language inference task by using Cause-and-Effect relation prediction as a supplementary training phase. We also conduct a set of analytical experiments to validate properties of this method.

5.1 Directional Relation Prediction

Except for the automatic curated $\mathcal{CER}$ dataset, we also extract a subset of ATOMIC which contents directional relationships. ATOMIC is a commonsense dataset which consists of 9 kinds of relations: xIntent, xNeed, xAttr, xReact, xWant, xEffect, oReact, oWant, oEffect. We only focus on the relations which have directional properties (i.e., xNeed, xEffect). The meaning of these two relations and the corresponding examples are listed in Table 5. It is not hard to tell that *xNeed*$(S_1, S_2) \leftrightarrow$ *xEffect*(S_2, S_1) (represents temporal precedence). We add the subject personX in front of sentence B[6] to avoid the false statistical features introduced by the different formats of sentence A and sentence B. Similar to our curated Cause-and-Effect relationship dataset, we also reverse the sentence pair to generate a dual set. And we randomly divide the dataset into train/validation/test sets with 80%, 10%, 10% splits.

Experiments settings We compared the proposed mutual attention mechanism with various different strategies to encode directional information:

[6]We also delete the "to" in sentence B and transform the verb to third person singular to make sentences more natural.

Relation	Sentence A	Sentence B
xNeed	[PersonX makes Person Y's coffee]	PersonX needed [to put the coffee in the filter]
xEffect	[PersonX makes Person Y's coffee]	PersonX then [gets thanked]

Table 5: The meaning of each relation and the corresponding example in ATOMIC.

Dataset	Model	Origin		Test-Augmented		Augmented	
		Acc	F1	Acc	F1	Acc	F1
$\mathcal{CER}$	Transformer-SE	73.2	84.5	46.9	61.5	71.2	80.9
	Transformer+InferSent	80.7	84.4	69.6	72.3	72.1	77.2
	Transformer-Mul	82.0	86.2	**71.3**	**78.1**	**78.4**	**86.7**
	BERT	87.2	91.5	62.1	69.2	82.7	90.9
	RoBERTa	88.3	93.4	61.8	68.2	84.9	94.6
	BERT-Mul	87.8	92.1	**73.7**	**80.6**	**85.2**	**95.1**
ATOMIC	Transformer-SE	80.7	86.2	60.7	67.4	70.1	77.9
	Transformer+InferSent	79.1	85.6	69.8	74.2	77.7	84.5
	Transformer-Mul	80.6	85.2	**70.2**	**77.8**	**82.6**	**86.1**

Table 6: F1 score and accuracy on $\mathcal{CER}$ and the subset of ATOMIC. **Origin** means models are both trained and test on the original dataset where one sentence can either be S_1 or S_2 once. **Test-Augmented** 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 vice versa. **Augmented** means models are both trained and test on augmented dataset.

1. **Transformer-SE**: A BERT-like transformer structure whose input consists of token embeddings, position embeddings and segment embeddings.
2. **Transformer+InferSent**: It uses the two hidden states corresponding to the first token of both sentences as sentence embeddings from the transformer encoder, then uses the two sentence embeddings to construct the final feature vector, consisting of the concatenation of two sentence vectors, their difference, and their elementwise product (Conneau et al., 2017).
3. **Transformer-Mul** is our proposed compositional model that combines single-sentence attention and mutual attention. We cascade 10 single-sentence attention layers with 2 mutual attention layers.
4. **BERT-Mul**: To test its compatibility with pretrained models, we initialize Transformer-Mul with pretrained parameters of BERT-base (Devlin et al., 2018) and compare it with BERT-base, RoBERTa-base.

Parameter Settings For models trained from scratch, we first train GloVe (Pennington et al., 2014) word embeddings on the Wikipedia and BookCorpus with 300 dimensions. We set the number of attention heads per layers $n_{head} = 5$ and $d_{model} = 300$, $d_k = d_v = d_q = 60$. We use Adam with 2×10^5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e8$. Gradient is clipped to 1. In both the training and test set, we truncate sentences with more than 128 words into 128 words. Batch size is set to 64. For BERT-Mul, we adopt 2 layers of mutual attention of the same parameters with BERT-base.

Results The evaluation results on $\mathcal{CER}$ and the subset of ATOMIC are shown in Table 6. We observe that Transformer-Mul performs better than or competitive with other approaches on both F1 score and accuracy in all the settings. It is worth noting that although in the original dataset the set of S_1 and the set of S_1 have no overlap, the performance of Transformer-SE is comparable with Transformer-Mul, when tested on the augmented test set, Transformer-Mul outperforms Transformer-SE by a large margin. We hypothesize that Transformer-SE is likely to rely more on the co-occurrence of two sentences. The empirical results also show that even when trained on the augmented dataset, the mutual attention mechanism continues outperforming other methods of combining direction information, further revealing the advantage of the mutual attention mechanism. These results indicate that by applying the mutual attention mechanism, models tend to learn sentence-pair representations which contain more information about the direction of their relations.

Score Distribution Figure 2 shows two sets of samples on the augmented $\mathcal{CER}$ corpus: 1) the samples labeled as isCauseOf and 2) the samples labeled as isEffectOf. Transformer-SE is not sensitive to direction and confuses between isCauseOf and isEffectOf. Therefore, the distributions of its predicted scores are in a saddle shape. In contrast,

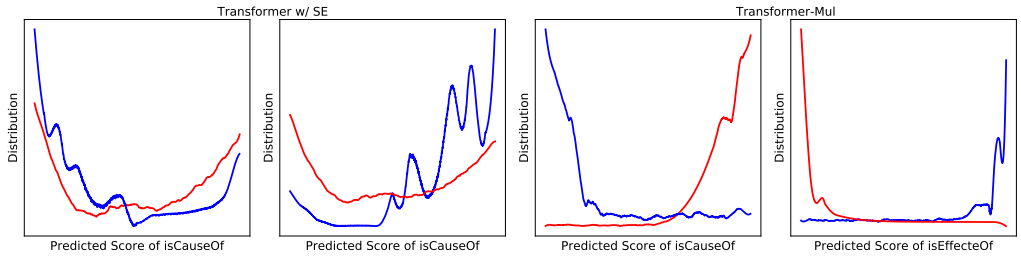


Figure 2: The distributions of predicted scores on randomly selected samples: samples labeled as isCauseOf (red line) and isEffectOf (blue curve). As shown in the figures, transformer-SE (left two figures) confuses between these two relations and assigns high scores on both categories of samples. In contrast, Transformer-Mul (right two figures) captures the directional relations better.

Transformer-Mul is able to distinguish these types of relations, so the distribution of predicted scores is more monotonous.

5.2 Experiments on aNLI

The aNLI Dataset Abduction has long been considered to be at the core of how people interpret and read between the lines in natural language. Abductive Natural Language Inference (aNLI) is a commonsense benchmark dataset designed to test an AI system's capability to apply abductive reasoning and common sense to form possible explanations for a given set of observations. Formulated as a binary-classification task, the goal is to pick the most plausible explanatory hypothesis given two observations from narrative contexts. Each instance is defined as follows: *Obs1*:The observation at time t_1; *Obs2*: The observation at time $t_2 > t_1$; *Hyp+*: A *plausible* hypothesis that explains the two observations; *Hyp-*: Am *implausible* (or less plausible) hypothesis that explains the two observations. For example:

Obs1	*Jenny was addicted to sending text messages.*
Obs2	*Jenny narrowly avoided a car accident.*
Hyp-	*Since her friend's texting and driving car accident, Jenny keeps her phone off while driving.*
Hyp+	*Jenny was looking at her phone while driving so she wasn't paying attention.*

Experiment Setup We design the following experiment to test whether a model can transfer the knowledge learned on $\mathcal{CER}$ to the downstream aNLI task. In particular, we compare the following models:

1. **BERT** & **RoBERTa**: The original BERT-base and RoBERTa-base.
2. **BERT-InferSent** & **RoBERTa-InferSent**: Two single-sentence embeddings are provided by BERT/ RoBERTa to construct the final feature vector, consisting of the concatenation of two sentence vectors along with their difference and elementwise product.
3. **BERT-Mul** & **RoBERTa-Mul**: Contextualized embeddings are provided by BERT/ RoBERTa and 8 mutual attention layers on the top are added on the top.

Implementation Details We also apply a similar transfer strategy to ATOMIC and a variation of ROC dataset, following the work by Roemmele and Gordon, in which they assume sentences from the same story have a casual relationship with each other. We implemented one-way negative sampling for all datasets in all settings. We also experimented with pretraining on three label classification task in the same format as that mentioned in previous sections. One-way negative sampling provides better results while the relative gaps between different approaches are consistent.

Results and Analysis As shown in Table 7, fine-tuning on both $\mathcal{CER}$ provide largest gains on aNLI, which indicates that $\mathcal{CER}$ is a good source for directional casual information needed to make abductive natural language inference. Models with mutual attention mechanism outperform both simple continuous fine-tuning and single sentence encoder pretraining with InferSent structure, which shows that with a mutual attention mechanism at the top, models can benefit more from causal sentence-pair pretraining task and are more transferable to natural language reasoning tasks. The results suggest that 1) directional relation information captured by models are beneficial for abductive reasoning; 2) mutual attention helps models to better capture transferable directional relation representations.

Model	no pretrain	ROC	$\mathcal{CER}$	ATOMIC	Multi-Task
BERT	63.62	62.92	68.50	60.31	62.21
BERT-InferSent	60.40	60.24	68.77	62.02	62.59
BERT-Mul	63.62	65.79	**69.57**	64.17	65.42
RoBERTa	73.85	70.05	72.18	71.69	76.93
RoBERTa-InferSent	72.13	75.98	70.98	70.51	76.20
RoBERTa-Mul	72.85	73.72	**76.70**	74.57	**77.15**

Table 7: Results on aNLI dataset. The leftmost column shows the results without pretraining on other directional relation corpora; the middle three columns shows the results of pretraining on a single dataset; and the rightmost column shows the results of pretraining on the three datasets in a multi-task manner.

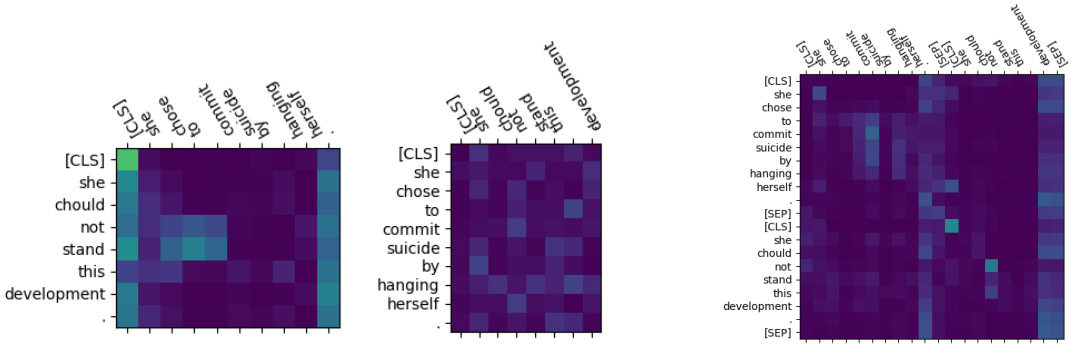


Figure 3: The figure on the left is the mutual attention head in Transformer-Mul. The figure on the right is the fully-connected attention head in Transformer-Full. The sentence pair is "she could not stand this development." vs. "she chose to commit suicide by hanging herself."

6 Case Study

We observe that the attention maps in the mutual attention heads are highly asymmetric: There exist significant differences between the attention maps of $S_1 \rightarrow S_2$ and the attention maps of $S_2 \rightarrow S_1$ for a large amount of sentence pairs. To formally verify our observation, we conduct a comparison experiment between Transformer-Mul model and another Transformer+Full model, which has the same parameters with Transformer-Mul. The only difference is the top two layers in Transformer-Full are fully-connected attention heads rather than mutual attention heads.

We randomly select 100 sentence pairs from $\mathcal{CER}$ and conduct a paired t-test between the attention maps of $S_2 \rightarrow S_1$ and the attention maps of $S_1 \rightarrow S_2$ (10 heads per sentence pair). The same t-test also conducted on Transformer-sub&full. Figure 3 shows a pair of attention maps as an example. For the fully-connected heads, the two blocks in the opposition angle are chosen to be paired since they represent attention weights for the other sentence. The means of t-scores of Transformer and Transformer-Mul are 2.01 and 9.09, respectively.

Although the differences in both models are statistically significant, Transformer-Mul consistently gets higher t-scores, which indicates that the attention maps of two directions in mutual attention heads are statistically more different than those in fully connected heads.

Moreover, the mutual attention yields more interpretability. For the example sentence pair (S_1 ="she could not stand this development." , S_2 ="she chose to commit suicide by hanging herself."), as is shown in Figure 3, while the attention from "suicide" to "could not stand" is high, there does not exist symmetrical high attention from "could not stand" to "suicide", indicating that the mutual attention directionally encodes the casual association between phrases.

7 Conclusion

This paper studies the effects of training directional sentence pair embeddings. By curating a directional casual dataset $\mathcal{CER}$, we showed that mutual attention is a superior mechanism to capture directional relations between pairs of sentences. We also showed that the learned directional sentence pair embeddings are transferable to abductive reasoning task, outperforming existing architecture and training approaches. Future work includes developing automatic schemes to extract other directional relations, such as temporal relations and extract the commonsense knowledge from multi-modality.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott W en-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Melissa Roemmele and Andrew Gordon. 2018. An encoder-decoder approach to predicting causal relations in stories. In *Proceedings of the First Workshop on Storytelling*, pages 50–59, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: an atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.