CoCoLM: Complex Commonsense Enhanced Language Model with Discourse Relations

Changlong Yu[1*] **Hongming Zhang**[1,2*] **Yangqiu Song**[1] **Wilfred Ng**[1]
[1]HKUST, Hong Kong, China [2]Tencent AI Lab, Seattle, U.S.
{cyuaq, yqsong, wilfred}@cse.ust.hk, hongmzhang@tencent.com

Abstract

Large-scale pre-trained language models have demonstrated strong knowledge representation ability. However, recent studies suggest that even though these giant models contain rich simple commonsense knowledge (e.g., bird can fly and fish can swim.), they often struggle with complex commonsense knowledge that involves multiple eventualities (verb-centric phrases, e.g., identifying the relationship between "Jim yells at Bob" and "Bob is upset"). To address this issue, in this paper, we propose to help pre-trained language models better incorporate complex commonsense knowledge. Unlike direct fine-tuning approaches, we do not focus on a specific task and instead propose a general language model named CoCoLM. Through the careful training over a large-scale eventuality knowledge graph ASER, we successfully teach pre-trained language models (i.e., BERT and RoBERTa) rich discourse-level commonsense knowledge among eventualities. Experiments on multiple commonsense tasks that require the correct understanding of eventualities demonstrate the effectiveness of CoCoLM.

1 Introduction

Recently, large-scale pre-trained language representation models (LMs) (Devlin et al., 2019; Liu et al., 2019) have demonstrated the strong ability to discover useful linguistic properties of syntax and remember an impressive amount of knowledge with self-supervised training over a large unlabeled corpus (Petroni et al., 2019; Jiang et al., 2020). On top of that, with the help of the fine-tuning step, LMs can learn how to use the memorized knowledge for different tasks, and thus achieve outstanding performance on many downstream natural language processing (NLP) tasks.

As discussed in Verga et al. (2020), while language models have already captured rich knowledge, they often only perform well when the semantic unit is a single token while poorly when the semantic unit is more complex (e.g., a multi-token named entity or an **eventuality**, which is a linguistic term for verb-centric phrases covering *activities*, *states* and *events* (Bach, 1986; Araki and Mitamura, 2018)). For example, as shown in Table 1, if we follow LAMA (Petroni et al., 2019) to analyze the knowledge contained in BERT-large (Devlin et al., 2019) with a token prediction task, we can find out that BERT can understand that birds can fly, and a car is used for transportation, but it fails to understand the relation between "Jim yells at Bob" and relevant eventualities. An important reason behind this is that current language models heavily rely on token-level masked language models (MLMs) as the loss function, which can effectively represent and memorize token co-occurrence statistics[1] but struggle at perceiving multi-token concepts.

Query	Answer
Birds can [MASK].	fly
Cars are used for [MASK].	transport
Jim yells at Bob, [MASK] Jim is upset.	but
Jim yells at Bob, [MASK] Bob is upset.	but

Table 1: Exploring knowledge contained in pre-trained language models following LAMA (Petroni et al., 2019). Queries and prediction returned by BERT-large are presented. Semantically plausible and implausible prediction are indicated with blue and red colors.

To address this problem and equip LMs with complex and accurate human knowledge, several recent works attempt to integrate entity representations from external knowledge graphs. While those approaches have been proved effective in merging structured knowledge into the LMs, they still have two limitations when applying to eventuality representations: (1) The first line of work (Verga

[*] Equal contribution.

[1] Sinha et al. (2021) also explains the success of LMs due to distributional information. These models pretrained over sentences with shuffled word order still achieve high accuracy.

Findings of the Association for Computational Linguistics: ACL 2022, pages 1175 - 1187
May 22-27, 2022

Type	Sequences
Temporal	I had a dream. *Precedence* (**Before**) I met with you yesterday. *Succession* (**After**) There were so many matters.
Casual	I go to supermarket. *Reason* (**Because**) I have a coupon. *Result* (**So**) The price is great.
Others	You can come with me. *Alternative* (**Or**) You can stand here. *Contrast* (**But**) The situation remains unchanged.

Table 2: Examples of eventuality sequences with different types of discourse relations (highlighted with pink) and connectives (bolded). Note there may exist multi-relational eventuality pairs

et al., 2020; Shen et al., 2020; Févry et al., 2020) restricts a fixed set of named entities or concepts to be linked to KGs while the eventualities are not easily canonicalized. There are enormous eventualities, which many of them refer to similar meanings such as "Tom is upset" and "Alice is upset". (2) The second class of methods uses powerful contextualized representations to encode one-hop triplets from KGs (Bosselut et al., 2019; Yao et al., 2019) for the task of KG completion. However, it is not sufficient for tasks that require the understanding of complex discourse relations in the event sequences or chains. For example pretrained LMs on the story ending prediction task (Mostafazadeh et al., 2016) have gaps with human performance (Li et al., 2019). Besides that, different types of relations (*casual* or *temporal*) make high-order inference over eventualities difficult and challenging.

In this paper, to effectively inject eventuality knowledge into pre-trained language representation models, we propose a knowledge injection framework CoCoLM, which requires no concept or eventuality linking and encodes multi-hop eventuality information as well as their discourse relations. The starting point is a large-scale eventuality knowledge graph, ASER (Zhang et al., 2020b), where the edges are discourse relations among eventualities (e.g., "being hungry" can cause "eat food" and "being hungry" often happens at the same time as "being tired"). First, we go beyond one-hop modeling (Yao et al., 2019; Bosselut et al., 2019) and carefully conduct weighted random walk over ASER to harvest multi-hop eventuality sequences connected by discourse relations (§2.1). Individual eventualities are contextualized by coherent sequences (examples in Table 2). Second, we fine-tune pretrained LMs on the sampled sequences and reformulate the masked language modeling objective to new *eventuality-level* masking to perceive the eventualities as independent semantic units (§ 2.2). In addition, two auxiliary tasks of discourse relation prediction are proposed to make implicit commonsense inferences (§ 2.3). For example, the new tasks explicitly reinforce the casual relation prediction between "I have a coupon" and "The price is great". By doing so, we successfully expose and inject fruitful high-order information between eventualities to pretrained LMs. To understand the impact of our proposed CoCoLM, we conduct experiments on three tasks that require the understanding of temporal, causal, mixed (multiple) relations respectively. The results show that our method achieves substantial improvements on the multiple-relation task while competitive performance on single-relation tasks. Extensive analyses are conducted to show the effect and contribution of all components in CoCoLM. Our main contributions are as follows:

- We propose CoCoLM, a new contextualized language model enhanced by complex commonsense knowledge from high-order discourse relations. CoCoLM is trained to predict the whole eventuality among the sequences using a large-scale eventuality KG.
- We introduce two auxiliary discourse tasks to help incorporate discourse-related knowledge into pre-trained language models, which complement the special eventuality masking strategy.
- CoCoLM achieves stronger performance than the baseline LMs on multiple datasets that require the understanding of complex commonsense knowledge about eventualities.[2]

2 Methods

The overall framework of CoCoLM is presented in Figure 1. Given a pre-trained language model, we inject complex commonsense knowledge about eventualities by adding one adaptive pre-training stage (Gururangan et al., 2020). Specifically, we first generate eventuality sequences based on carefully controlled walks over existing eventuality knowledge graphs and then use the sequences as

[2]Our code and models are available at https://github.com/HKUST-KnowComp/Co2LM.

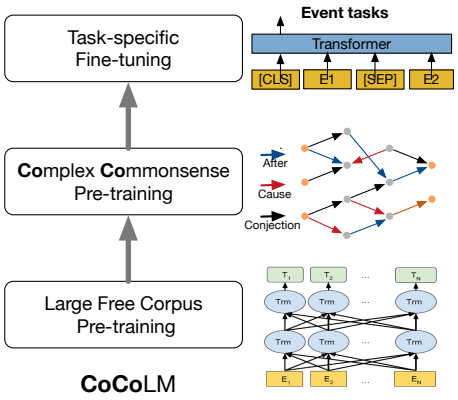

Figure 1: The overall framework of CoCoLM. On top of base pre-trained LMs, complex commonsense knowledge from the eventuality sequences is injected by fine-tuning MLMs and auxiliary discourse tasks.

the context to help LMs handle eventualities. Besides the original token-level MLM objective, we also introduce the *eventuality-level* masking strategy and several auxiliary tasks to assist the training. As the training is not task-specific, the resulting LM can be easily applied to any downstream tasks via another task-specific fine-tuning stage.

2.1 Eventuality Sequence Generation

Multi-hop path information have been shown useful and interpretable to provide extra context knowledge by connecting the concepts from ConceptNet (Speer et al., 2017) for commonsense reasoning tasks (Lin et al., 2019; Wang et al., 2020a). Similarly, we generate eventuality sequences by leveraging ASER, which uses eventualities as nodes and the discourse relations as edges. ASER extracts rich eventuality knowledge from diverse corpus, such as "being hungry" and "being tired" often happen together and people often "make a call" before they go. It contains much larger scale of discourse relations than DisSent (Nie et al., 2019). Interestingly, beyond the single edges, higher-order connections over ASER can also reflect insightful eventuality knowledge. For example, "sleep" and "go" are not likely to happen at the same time because "sleep" can be caused by "being tired" and there exist *contrast* connections between "being tired" and "go". To include higher-order knowledge into the model, we propose to take the whole graphical structure into consideration rather than single-hop edges. Motivated by DeepWalk (Perozzi et al., 2014), we randomly sample paths to simulate the overall graph structure and generate eventuality-level co-occurrence information.

Given the initial knowledge graph $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ is the eventuality set and $\mathcal{R}$ is the relation set, we conduct the weighted random walk based on the edge weights over $\mathcal{G}$ to sample eventuality paths. We denote each path as $(E_0, r_0, E_1, r_1, ..., r_{l-1}, E_l)$, where E means an eventuality, r a discourse edge connecting two eventualities, and l the numbers of eventualities along the sequence. To convert the sampled sentence into a token list, we keep all words in each event as a sentence and use representative connectives for each discourse relation to connect them (examples in the Table 2; full list in the Appendix Table 9). As ASER is automatically extracted from raw corpus, it may contain noise. To minimize the influence of the noise and improve the informativeness, the selected paths should fulfill:

1. To filter out rare eventualities, the frequency of starting eventualities has to be larger than five.

2. Other than the relations that have the transitive property (e.g., *Precedence*, *Result*), each selected path should not contain successive edges with repeated relations.

3. We manually improve the sampling probability of selecting sub-sequence patterns like "E_i **Condition** E_j **Reason** E_k". Since it has been proven that *if-then* rules (Sap et al., 2019) and *if-then-because* rules (Arabshahi et al., 2020) are crucial for reasoning.

2.2 Eventuality-Level Mask

Masking strategy plays a crucial role in the training of language representation models. Besides the random token-level masking strategy, many other masking strategies have been explored by previous literature such as the-whole-word masking (Devlin et al., 2019; Cui et al., 2019), entity masking (Sun et al., 2019; Shen et al., 2020) or text span masking (Joshi et al., 2020)[3]. Similarly, to effectively help the model view each eventuality as an independent semantic unit, we propose the following two masking strategies: (1) **Whole Eventuality Masking**: Similar to the whole word masking or entity masking strategies, the whole eventuality masking aims to reduce the prior biases of eventuality tokens. For example, given an eventuality sequence "I feel sleepy because I drink a cup of [MASK].",

[3]Unlike `SpanBERT`, we have discourse connectives as span boundaries and do not need the SBO objective.

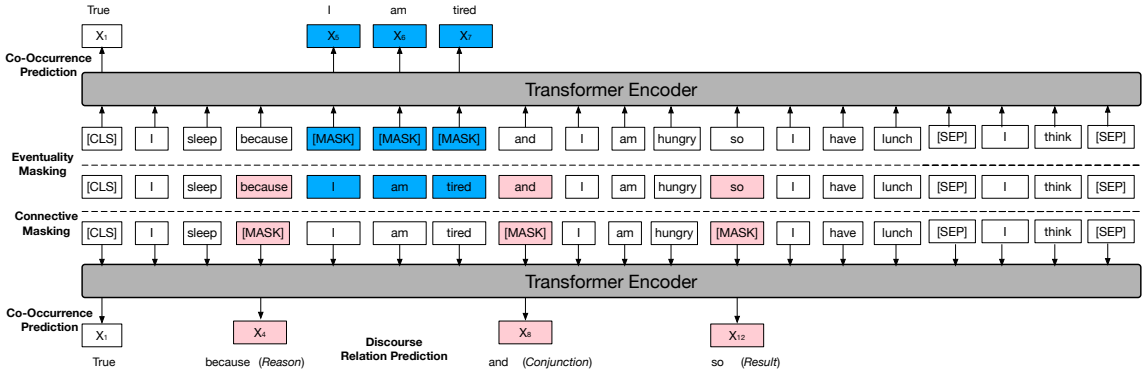

Figure 2: Illustration of CoCoLM-(Complex commonsense pre-training stage). Given an eventuality sequence, it is either masked by the whole eventuality masking (`in blue`) or discourse connective masking strategy (`in pink`). Besides the regular masked language model, the discourse relation labels are jointly predicted for masked connective tokens (on $\mathbf{x}_4$, $\mathbf{x}_8$ and $\mathbf{x}_{12}$). Co-occurrence prediction (on $\mathbf{x}_1$) is conducted for both masking strategies.

BERT would easily predict "coffee" or "tea" because of the prior knowledge of "cup of" inside the eventuality. Instead of that, masking the whole "I drink a cup of coffee" would encourage the prediction to treat each eventuality as an independent semantic unit and focus on the relations between them. For each sampled sequence, we randomly mask at most one eventuality to fulfill the masking budget, which is typically 25% of the sequence token length. (2) **Discourse Connective Masking**: Besides masking the eventualities, to effectively encode the discourse information, we also tried masking the discourse connectives.

Examples of two masking strategies are shown in Figure 2. It is worth mentioning that for each sequence, we only randomly select one type of masking strategy to guarantee that enough information is kept in the left tokens for the prediction. The formal masking objective is defined as follows. Given a tokenized sampled sequence $X = (x_1, x_2, ..., x_n)$, after masking several tokens, we pass it to a transformer encoder (Vaswani et al., 2017) and denote the resulting vector representations as $\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_n}$. The training loss $\mathcal{L}_{mlm}$ can thus be defined as:

$$\mathcal{L}_{mlm} = -\frac{1}{|\mathbf{M}|} \sum_{i \in \mathbf{M}} \log P(x_i | \mathbf{x_i}), \qquad (1)$$

where $\mathbf{M}$ means the set of masked tokens following the aforementioned masking strategies.

2.3 Auxiliary Tasks

A limitation of the MLM loss is that the prediction is over the entire vocabulary, and as a result, the model could not effectively learn the connection between eventualities and connective words. To remedy this and force the model to learn the discourse knowledge, we propose to add an additional classification layer after the last layer of transformer encoder and it feeds the output vector $\mathbf{x_i}$ of connective token x_i into a softmax layer over the set of discourse relation labels as follows.

$$P(l_i | \mathbf{x_i}) = \text{softmax}(\mathbf{x_i W} + \mathbf{b}), \qquad (2)$$

$$\mathcal{L}_{rel} = -\sum_{i \in \mathbf{M}_R} \log P(l_i = \widetilde{l}_i | \mathbf{x}_i), \qquad (3)$$

where $\mathbf{M}_R$ is the index set of masked discourse connective tokens (e.g., *because*, *and*, *so*) in Figure 2, l_i is the predicted discourse relation label, and $\widetilde{l}_i$ the label provided by ASER (10 relations in Table 9). $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters.

Besides the aforementioned discourse relations, ASER also provides the *Co-occurrence* relations between eventualities, which mean that two eventualities appear in the same sentence, but there are no explicit discourse markers between them. Co-occurrence information has been used for narrative event prediction (Chambers and Jurafsky, 2008) Though *Co-occurrence* relations are less informative, high frequency pairs still reflect rich knowledge about eventualities. Motivated by this, we propose another auxiliary task to help the model to learn such knowledge. Specially, given an eventuality sequence $S = (E_0, r_0, E_1, r_1, ..., r_{l-1}, E_l)$ and an eventuality E_c, we format the input[4] as "[CLS] S [SEP] E_c [SEP]". We set 50% of the time E_c to

[4]The special tokens are based on the base model, i.e., we add "[CLS]" and "[SEP]" for `BERT` models and add "<s>" and "</s>" for `RoBERTa` models. All notations in the rest of this paragraph are based on `BERT`.

be the positive *co-occurred* eventuality with one of the eventualities in the sequence while 50% of the time E_c is randomly sampled negative in ASER. Similar to the next sentence prediction in the original BERT, on top of the vector representation of token [CLS], i.e., $\mathbf{x}_{cls}$, we add another classification layer to predict whether the *Co-occurrence* relations hold or not. The training objective $\mathcal{L}_{occur}$ for binary classification is similar to $\mathcal{L}_{rel}$:

$$\mathcal{L}_{occur} = -\log P(l_i = \widetilde{l}_i, |\mathbf{x}_{cls}), \quad (4)$$

where $\widetilde{l}_i$ is the true co-occurrence label (positive or negative) for the sequence.

Merging all three losses together, we can then define the overall loss function $\mathcal{L}$ as:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_{rel} + \mathcal{L}_{occur}. \quad (5)$$

3 Experiments

3.1 Implementation Details

In this work, we use the released ASER-core version[5] extracted from multi-domain corpora, which contains over 27.6 million eventualities and 8.8 million relations. We follow the heuristic rules in Sec. 2.1 to sample eventuality sequences for pre-training. Overall we generated 4,041,572 eventuality sequences (sentences), ranging from one to five hops and the one-hop sequence means the direct (first-order) edge in the ASER. We also down-sample eventuality nodes with extremely high frequency such as *I see*. Sequence examples are listed in Table 2 and more examples as well as sequence distribution over different lengths are appended in the Appendix.

We select base and large version of `BERT` (Devlin et al., 2019), `RoBERTa` (Liu et al., 2019) as the base LM. For the continual complex commonsense pre-training phase, we use the Adam optimizer for 10 epochs with batch size 128, learning rate 1e-5 and weight decay 0.01. Considering the relative longer span of masked eventualities, we enlarge the masking proportion from 15% to 25%, which averagely add 1.7 more masked tokens in the sequences. We implemented the pretraining with Huggingface library (Wolf et al., 2020) and running CoCoLM pretraining on eight Nvidia V100 32GB GPUs took four days. Pretraining introduces two classification layers with thousands of parameters.

[5] https://github.com/HKUST-KnowComp/ASER

Dataset	Type	# Train	# Dev	# Test
ROCStories	Narrative (Multiple)	1,771	100	1,871
MATRES	Temporal	231*	25	20
COPA	Causal	400	100	500

Table 3: The statistics of evaluation datasets (See examples in A.1). The tasks are binary or multiple classification problems. Note the dataset of MATRES is split at the article level following Ballesteros et al. (2020).

3.2 Datasets and Evaluations

In this section, we introduce evaluation datasets and settings as follows:

ROCStories (Mostafazadeh et al., 2016) is widely used for story comprehension tasks such as Story Cloze Test. It contains 98,162 five-sentence coherent stories as the unlabeled training dataset, 1,872 four-sentence story contexts along with two candidate ending sentences in the dev and test datasets. The dataset split for the story ending prediction task is the same as Li et al. (2019).

MATRES (Ning et al., 2018b) is a pairwise event temporal ordering dataset, where each event pair in one document is annotated with a temporal relation (*Before, After, Equal, Vague*). It contains 13,577 event pairs extracted from 256 documents for training (25 left for dev) and 20 for testing.

COPA (Gordon et al., 2012) is a binary-choice commonsense causal reasoning task, which requires models to predict which the candidate hypothesis is the plausible effect/cause of the given premise. We follow the training/dev/test split in SuperGLUE (Wang et al., 2019).

The statistics of the three selected datasets are presented in Table 3. For fine-tuning experiments, we select the learning rate from {2e-5, 1e-5, 5e-6}, and maximize the sequence length and batch size such that they can fit into GPU memory. Fine-tuning was much faster due to fewer new parameters from classification layers.

Different from MATRES and COPA, solving the story ending tasks of ROCStories requires multi-type relation inferences including causal, temporal etc. Moreover, as mentioned in Sharma et al. (2018), there is a strong bias about the human-created negative endings such that the model can distinguish the positive and negative endings without seeing the first four events. Even though Sharma et al. (2018) tried to filter the annotations, the bias still cannot be fully relieved. As a result, to clearly show the effect of adding complex

Model	ROCStories		MATRES		COPA
	Accuracy	Accuracy (D)	Accuracy	F1	Accuracy
BERT-base	52.9	45.9	71.5	77.2	69.8
CoCoLM (BERT-base)	84.2	65.2	72.8	77.8	73.8
BERT-large	88.9	69.1	73.5	78.9	70.6
CoCoLM (BERT-large)	91.9	71.2	73.9	79.2	75.8
RoBERTa-base	93.3	73.2	74.0	79.2	85.4
CoCoLM (RoBERTa-base)	94.1	75.2	74.2	79.8	86.2
RoBERTa-large	97.4	88.1	75.2	81.0	90.6
CoCoLM (RoBERTa-large)	**97.9**	**89.4**	**75.5**	**81.6**	**91.3**

Table 4: Evaluation results on three commonsense task (top scores in boldface). We report the accuracy of ROCStories dataset under normal supervised setting and debiased (D) setting mentioned in the §3.2.

knowledge about events into the LM, besides the most widely used supervised setting, we also report the performance of a *debiased setting*, where the model randomly selects events from other stories as the negative ending during the training. The debiased setting is indicated with "D". Following previous works, we report accuracy for the ROCStories, MATRES and COPA tasks. For MATRES, we also report F1 scores by considering the task as general relation extraction and treating the label of *vague* as *no relation* (Ning et al., 2019). All models are trained until convergence and the best model on the dev set is selected to be evaluated.

4 Experimental Results

The results are presented in Table 4, from which we can see that CoCoLM consistently outperforms all the baselines on all three commonsense tasks, especially on the debiased setting of ROCStories.

Besides that, we can make the following observations. First, the improvement of our model is more significant on ROCStories than COPA and MATRES, which is mainly because multiple relation combinations in the eventuality sequences bring high-order information and thus help complex reasoning. Second, CoCoLM achieves significant improvement on lower-capacity LMs trained on small corpora. For example, CoCoLM brings up to 59.2% improvement over BERT-base on the ROCStories dataset. Third, compared with the original supervised setting, the debiased setting is more challenging for all models, which helps verify our assumptions that previous models might benefit from the bias. Here the debiased setting should be more fair for comparison. When we dig into the MATRES dataset, event pairs (typically verb pairs) are associated with the context, which some of could be easily inferred from the local context with obvious clues (Ballesteros et al., 2020) while the others may need external commonsense knowledge that can be memorized by the language models. As a comparison, both the ROCStories and COPA do not have any extra context, and thus require the pre-trained LMs to know the essential knowledge to solve those problems.

Method	Accuracy	Δ	Accuracy (D)	Δ (D)
CoCoLM	91.9	-	71.2	-
w/o *occur loss*	91.3	-0.6	70.4	-0.8
w/o *eventuality mask*	91.1	-0.8	70.2	-1.0
w/o *rel loss*	90.5	-1.4	69.6	-1.6
w/o *occur* & *rel losses*	90.3	-1.6	69.3	-1.9
w token-level *mlm* only	89.2	-2.7	69.2	-2.0

Table 5: Ablation study on ROCStories test set by removing different model components. *occur* and *rel* are discourse relation and co-occurrence loss respectively.

In the rest of this section, we conduct extensive experiments and case studies to demonstrate the contribution of different components. In all following analysis experiments, we use BERT-large as the base language model and ROCStories as the evaluation dataset.

4.1 Ablation Study

We conduct an ablation study in Table 5 via LM pretraining with different settings and then fine-tuning. We can see that all components contribute to the final success of our model, especially the *Relation loss*. This result again verified that discourse connective prediction is a much more challenging pretraining task as shown in Malmi et al. (2018). CoCoLM is optimized to memorize high-order discourse knowledge that strongly correlates with downstream tasks and thus brings

Resource	Accuracy	Δ	Accuracy (D)	Δ (D)
ASER (M)	91.9	-	71.2	-
ASER (S)	85.4	-6.5	67.5	-3.7
ATOMIC (S)	88.2	-3.7	68.2	- 3.0

Table 6: Effect of different event knowledge resources. "M" and "S" pertain to "multi-hop" and "single-hop".

more performance boost. Besides, when replacing the whole eventuality masking with random token masking, we can observe 0.8% (1.0%) accuracy drop, which indicates the usefulness of eventuality-level masking. The relative better improvement of *Co-occurrence loss* suggests our previous assumption that even though compared with other discourse relations (e.g., *Before* and *Cause*), the *Co-occurrence* relations have relatively weaker semantic, it still can help models to better understand events due to its large scale.

To further verify the effectiveness of proposed methods, we compare with the baseline that the `BERT-large` models are fine-tuned with only token-level MLM objective on syntactic eventuality sequences. The performance dropped close to finetuning over original `BERT-large`, which shows that the gains of CoCoLM are not simply from the MLM objective and the new proposed objectives as well as masking strategies contributed largely.

4.2 Effect of Different Knowledge Resources

To access the effect of high-order ASER commonsense knowledge, we compare with the performance of directly integrating single-hop edges from ASER. We decompose the multi-hop sequences into single-hop edges and keep the comparable size of single-hop edges with multi-hop ones. The results are shown in Table 6, from which we can see that there is still a notable gap between multi-hop and single-hop knowledge injection at the comparable size. Hence multi-hop knowledge is crucial for LMs to understand eventualities. Besides ASER, another important event knowledge resource is ATOMIC (Sap et al., 2019), which is a crowdsourced commonsense knowledge graph that contains nine types of *if-then* casual relations between social-centric events. However, it is a bipartite graph, which symbolically random walk over ATOMIC is impossible like normal graphs. Nevertheless, we are interested in the differences of injecting human-annotated and auto-extracted triplets into LMs. Though relation types and triplet size[6] may vary from other other, Fang et al. (2021) successfully converts discourse knowledge in ASER to *if-then* knowledge in ATOMIC and shows the former might roughly cover the latter. When injecting into LMs, we can see in Table 6 that ATOMIC can outperform the single-hop version of ASER since ATOMIC is cleaner with human annotations. We leave how to combine ASER and other event knowledge resources (Mostafazadeh et al., 2020; Hwang et al., 2020) to get more high-quality multi-hop event knowledge as our future work.

4.3 Effects of Knowledge Retrieval

We also study the effect of other knowledge injection methods, for example simple retrieving relevant nodes from ASER. We use the BM25 algorithm to retrieve Top 5 relevant nodes for each event in the ROCStories. Following Petroni et al. (2020), retrieved nodes are appended at the end of story context and separated by the `[SEP]` token. The results show no obvious improvements over baselines. The reason might be that single event nodes could not provide more information and all the tasks require relational knowledge. Advanced integration methods with retrieved knowledge like Lv et al. (2020) and Guu et al. (2020) are worthy to be deeply explored in the future.

4.4 Probing Experiments

We present one case study from the probing analysis experiment in Table 7 to further investigate the discourse-aware nature of our proposed language models. Motivated by Petroni et al. (2019), we put a `[MASK]` token between two events and try to ask the model to predict the connective. Take the case from COPA dataset as an example, connectives predicted by CoCoLM clearly show the *effect* relation between two events. However predictions from the baseline models reveal weaker (*temporal, conjunction*) or wrong (*contrast*) relations. Similar observations could be drawn from another two datasets. Like Table 1, we also sample 300 high-frequency pairs from ASER to predict connectives. The P@1 for CoCoLM (`BERT-large`) has 15.2% improvement over `BERT-large`. These observations show that CoCoLM manages to memorize richer discourse knowledge about daily events (Note that connective probing analysis does not mean strong correlations with downstream task performance).

[6]The detailed comparison is included in the AppendixA.4

Dataset	Example	BERT [MASK]	CoCoLM [MASK]
ROC Stories	Context: Ed made beef jerky for a living. He ran the business out of his garage. One day he woke up and noticed his garage jarred open. He looked inside and noticed everything in disarray Positive Ending: Ed was delighted to see this. Negative Ending: Ed was shocked called the police for an investigation.	Context + [MASK] + Ending: **P: when, then, while** **N: and, but, so**	Context + [MASK] + Ending: **P: so, hence, therefore** **N: or, and, though**
MATRES	The last surviving member of the team which first conquered Everest in 1953 has {e_1: died} in a Derbyshire nursing home. George Lowe, 89, {e_2: died} in Ripley on Wednesday after a long-term illness, with his wife Mary by his side.	S1, +[MASK] + S2: **and, sir, Dr**	S1, +[MASK] + S2: **then, afterwards, till**
COPA	Premise: The girl found a bug in her cereal. Positive Hypothesis: She lost her appetite. Negative Hypothesis: She poured milk in the bowl.	Pre+ [MASK] + Hypo: **P: then, but, and** **N: then, next, so**	Pre + [MASK] + Hypo: **P: so, therefore, thus.** **N: but, instead, and.**

Table 7: Examples from evaluation datasets. Connectives in blue are predicted by the `BERT-large` model and ones in pink are predicted by CoCoLM (`BERT-large`). "P" and "N" represent the positive and negative candidates.

5 Related Work

Understanding Events. It is important to represent and learn the commonsense knowledge for deeply understanding the causality and correlation between events. Recently various kinds of tasks requiring multiple dimensional event knowledge are proposed such as story ending prediction (Mostafazadeh et al., 2016), event temporal ordering prediction (Ning et al., 2018a), and event casual reasoning (Gordon et al., 2012). Prior studies have incorporated external commonsense knowledge from ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019) for solving event representation (Ding et al., 2019), story generation tasks (Guan et al., 2020), KG completion (Bosselut et al., 2019). However, their event-level knowledge is sparse and incomplete due to the human-annotated acquisition, which thus limits the model capacity, especially when injecting into LMs. Zhang et al. (2020b) builds a large-scale eventuality knowledge graph, ASER, by specifying eventuality relations mined from discourse connectives. It explicitly provides structural high-order discourse information between events spanning from temporal, casual to co-occurred relations, which has been proven to be transferable to human-defined commonsense (Zhang et al., 2020a; Fang et al., 2021) and help with script learning (Lv et al., 2020). In this work, we aim at making full use of multi-dimensional high-order event knowledge in the ASER to help pretrained LMs understand events.

Injecting Knowledge into LMs. Though Petroni et al. (2019) shows that pre-trained LMs store factual knowledge without fine-tuning, still, LMs can not handle knowledge-intensive tasks such as open-domain question answering or commonsense reasoning. Previous works explore different ways to inject various knowledge into pre-trained LMs for downstream tasks. They mainly differ from knowledge resources, masking strategies, and training objectives. From the resource side, entity-centric KGs are infused into LMs in the form of linked entities (Zhang et al., 2019; Peters et al., 2019; Xiong et al., 2020). triplets (Yao et al., 2019; Liu et al., 2020; Wang et al., 2020b) or descriptions (Wang et al., 2021b; Yu et al., 2020). Besides that, linguistic knowledge (e.g.,synonym/hypernym relations (Lauscher et al., 2020), word-supersense (Levine et al., 2020), dependency parsing (Wang et al., 2020b), and constituent parsing (Zhou et al., 2019)) also plays a critical role to improve LMs. Simple commonsense knowledge from ConceptNet (Speer et al., 2017) is injected into LMs via linked entity-level MLMs and a new distractor loss function (Shen et al., 2020). Last but not least, domain-specific knowledge is also customized to improve relevant tasks such as mined sentiment word (Tian et al., 2020), event temporal patterns (Zhou et al., 2020), and numerical reasoning data (Geva et al., 2020). We refer readers to Safavi and Koutra (2021) for the comprehensive survey. In this work, we aim at injecting complex commonsense into pre-trained LMs with two significant difference against previous works: 1) we use the event rather than tokens as the semantic unit, and propose to use an eventuality-based masking strategy as well as two auxiliary tasks to help LMs understand events; 2) We first leverage the random walk process on a large-scale knowledge graph to include multi-hop knowledge.

6 Conclusion and Future Work

In this work, we aim at helping pre-trained language models understand complex commonsense about eventualities. Specifically, we first conduct the random walk over a large-scale eventuality-based knowledge graph to collect multi-hop event knowledge and then inject the knowledge into the pre-trained LMs with an eventuality-based mask strategy as well as two auxiliary tasks. Experiments on three downstream tasks as well as extensive analysis demonstrate the effectiveness of the proposed model. As our approach is a general solution, we believe that it can also be helpful for other tasks that require complex commonsense about events.

For future work, we would sample sub-graph structures to explore more meaningful event-centric commonsense knowledge (Wang et al., 2021a). Moreover, we will equip our models with generative abilities by finetuning powerful T5 (Raffel et al., 2020) or BART (Lewis et al., 2020) models to help narrative story completion (Ji et al., 2020), commonsense inference (Gabriel et al., 2021), event infilling tasks (Lin et al., 2021). Unified event-aware language models like Zhou et al. (2022) would be promising and interesting directions.

Acknowledgements

Yangqiu Song was supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from ITC of Hong Kong with special thanks to HKMAAC and CUSBLT, and the Jiangsu Province Science and Technology Collaboration Fund (BZ2021065). We would also like to thank Wei Wang for insightful dicussions.

References

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2020. Conversational neuro-symbolic commonsense reasoning. *arXiv preprint arXiv:2006.10022*.

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy*, 9(1):5–16.

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the EMNLP 2020*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*, pages 4762–4779.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense knowledge. In *Proceedings of EMNLP-IJCNLP*, pages 4896–4905.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Y. Song, and Bin He. 2021. Discos: Bridging the gap between discourse knowledge and commonsense knowledge. *ArXiv*, abs/2101.00154.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12857–12865.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of ACL*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of SemEval 2012*, pages 394–398, Montréal, Canada.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *TACL*, 8:93–108.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *ArXiv*, abs/2010.05953.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the EMNLP*, pages 725–736, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL*, 8:64–77.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of COLING*, pages 1371–1383, Barcelona, Spain (Online).

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of ACL*, pages 4656–4667, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the ACL*, pages 7871–7880, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2019. Story ending prediction by transferable bert. In *Proceedings of IJCAI*, pages 1800–1806. AAAI Press.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the EMNLP-IJCNLP*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2021. Conditional generation of temporally-ordered event sequences. In *Proceedings of the ACL*, pages 7142–7157, Online. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *AAAI February 7-12, 2020*, pages 2901–2908.

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*.

Shangwen Lv, Fuqing Zhu, and Songlin Hu. 2020. Integrating external event knowledge for script learning. In *Proceedings of COLING*, pages 306–315, Barcelona, Spain (Online).

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic prediction of discourse connectives. In *Proceedings of LREC 2018*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the NAACL*, pages 839–849, San Diego, California.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. Glucose: Generalized and contextualized story explanations. In *EMNLP*.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the ACL*, Florence, Italy.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of ACL*, pages 2278–2288, Melbourne, Australia.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the EMNLP-IJCNLP*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the ACL*, pages 1318–1328, Melbourne, Australia.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: online learning of social representations. In *Proceedings of KDD*, pages 701–710.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the EMNLP-IJCNLP*, pages 43–54.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP 2019*, pages 2463–2473.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the EMNLP*, pages 1053–1067, Online and Punta Cana, Dominican Republic. 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*, volume 33, pages 3027–3035.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of ACL 2018*, pages 752–757.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In *Proceedings of EMNLP*, pages 8980–8994, Online. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the EMNLP*, pages 2888–2913, Online and Punta Cana, Dominican Republic.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. pages 4444–4451.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of ACL*, pages 4067–4076, Online. Association for Computational Linguistics.

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.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W. Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *CoRR*, abs/2007.00849.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020a. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online.

PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, and Xiang Ren. 2021a. Contextualized scene imagination for generative commonsense reasoning. *arXiv preprint arXiv:2112.06318*.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020b. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP*, pages 38–45, Online. Association for Computational Linguistics.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2020. Jaket: Joint pre-training of knowledge graph and language understanding. *arXiv preprint arXiv:2010.00796*.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of IJCAI*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. Aser: A large-scale eventuality knowledge graph. In *Proceedings of The Web Conference 2020*, pages 201–211.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of ACL*, pages 1441–1451.

Ben Zhou, Qiang Ning, Daniel Khashabi, and D. Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of ACL*.

Junru Zhou, Zhuosheng Zhang, and Hai Zhao. 2019. Limit-bert: Linguistic informed multi-task bert. *arXiv preprint arXiv:1910.14296*.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. *arXiv preprint arXiv:2203.02225*.

A Appendices

A.1 Examples of Evaluation Datasets

We select one example for each commonsense evaluation dataset and list in Table 8. In terms of MATRES, it has 13,577 event pairs among 256 articles with 4 temporal relations, i.e., *Before* (6,874), *After*(4,570), *Equal* (470) and *Vague* (1,656). Compared with MATRES and COPA, solving the ROCStories requires more complex commonsense knowledge to understand the whole narrative and multiple types of relations across events.

A.2 ASER Discourse Relations

In the Table 9, we list ten discourse relations as well as representative connectives (markers) used to train CoCoLM. We further categorize them into three types: "temporal", "casual" and "others". We refer the readers to original ASER papers (Zhang et al., 2020b) for detailed relation analysis.

Dataset	**Example**
ROC Stories	Context: The Mills next door had a new car. The car was stolen during the weekend. They came to my house and asked me if I knew anything. I told them I didn't, but for some reason they suspected me. Positive Ending: They called the police to come to my house. Negative Ending: They liked me a lot after that.
MATRES	Fidel Castro {e_1: invited} John Paul to {e_2: come} for a reason. *Label*: **BEFORE**
COPA	Premise: I knocked on my neighbor's door. Positive Hypothesis: My neighbor invited me in. Negative Hypothesis: My neighbor left his house.

Table 8: The examples for all commonsense evaluation datasets.

Types	Relations	Connectives
Temporal	Precedence Succession Synchronous	before after meanwhile
Casual	Reason Result Condition	because so if
Others	Conjunction Contrast Alternative Concession	and but or although

Table 9: The discourse relations as well as representative markers in the ASER knowledge graph.

A.3 Eventuality Sequences

We append more sampled eventuality sequences from random walk. Also we organize the sequences into several meta-paths (the paths with same relation patterns). Here only 2-hop and 3-hop sequences are listed and we could observe meaningful high-order connections between eventualities.

The sequence distributions with different lengths and types of relations are shown in the Figure 3. We can see that casual relations take up a small share, which again show that causal knowledge tends to be implicit and hard to acquire.

A.4 ATOMIC V.S. ASER

In this section, we summarize the nine casual/effect relations from ATOMIC (Sap et al., 2019) in the Table 2. Fang et al. (2021) shows ASER's discourse relations could be converted to causal knowl-

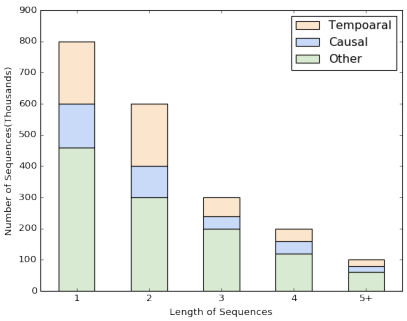

Figure 3: The distribution of lengths along with relation edges for generated eventuality sequences.

edge in the ATOMIC. Thus ASER might roughly cover the knowledge in the ATOMIC. Moreover, ASER also covers agentless events such as "the weather is good", which was partially covered by GLUCOSE (Mostafazadeh et al., 2020) However it contains noise compared with ATOMIC. CoCoLMexperiments show ATOMIC (877K edges) performs better than ASER(4.4 M - 5$\times larger$).

If-Then Types	Relations	Causal Types
If-Event-Then-State	xIntent	Cause
	xReact	Effect
	oReact	Effect
If-Event-Then-Event	xNeed	Cause
	xEffect	Effect
	xWant	Effect
	oEffect	Effect
	oWant	Effect
If-Event-Then-Persona	xAttr	Stative

Table 10: The *if-then* types and causal relations between events in the ATOMIC knowledge graph. For relations, "x" and "o" refer to PersonX and others while "xAttr", "xIntent", "xReact" mean the attribute, intent, reaction of PersonX *etc.*