Mitigating Reporting Bias for Semi-supervised Temporal Commonsense Inference with Probabilistic Soft Logic

Bibo Cai, Xiao Ding[*], Bowen Chen, Li Du, Ting Liu

Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{bbcai, xding, bwchen, ldu, tliu}@ir.hit.edu.cn

Abstract

Acquiring high-quality temporal common sense (TCS) knowledge from free-form text is a crucial but challenging problem for event-centric natural language understanding, due to the language reporting bias problem: people rarely report the commonly observed events but highlight the special cases. For example, one may rarely report "*I get up from bed in 1 minute*", but we can observe "*It takes me an hour to get up from bed every morning*" in text. Models directly trained upon such corpus would capture distorted TCS knowledge, which could influence the model performance. Prior work addresses this issue mainly by exploiting the interactions among temporal dimensions (e.g., *duration*, *temporal relation* between events) in a multi-task view. However, this line of work suffers the limitation of implicit, inadequate and unexplainable interactions modeling. In this paper, we propose a novel neural-logic based Soft Logic Enhanced Event Temporal Reasoning (SLEER) model for acquiring unbiased TCS knowledge, in which the complementary relationship among dimensions are explicitly represented as logic rules and modeled by t-norm fuzzy logics. SLEER can utilize logic rules to regularize its inference process. Experimental results on four intrinsic evaluation datasets and two extrinsic datasets show the efficiency of our proposed method.

Introduction

Time plays critical roles in daily life. Many natural language processing problems, including information retrieval (Ning et al. 2018; Vashishtha, Van Durme, and White 2019), summarization (Yan et al. 2011), causal inference (Noah Weber 2020) and reading comprehension (Ning et al. 2020; Zhou et al. 2021), rely on system's time understanding ability to give correct answers or predictions. Hence, recent NLP systems dedicate to incorporate temporal common sense (TCS) knowledge (e.g., *duration* and *frequency* of events) to improve time understanding ability (Zhou et al. 2021; Lin, Chambers, and Durrett 2021). While manually annotating large-scale TCS knowledge is time and labor consuming. A more practical approach is to automatically extract TCS knowledge from text with well-designed patterns (Zhou et al. 2020; Zhao, Lin, and Durrett 2021).

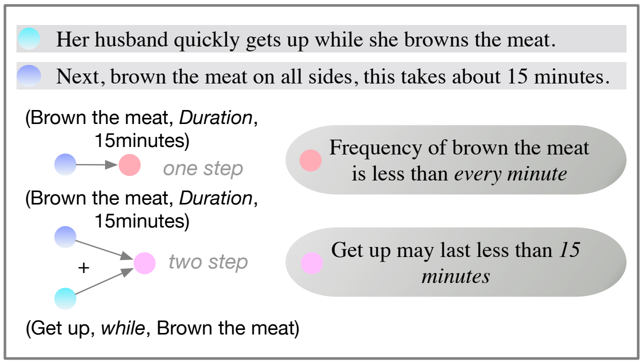
Her husband quickly gets up while she browns the meat.
Next, brown the meat on all sides, this takes about 15 minutes.
(Brown the meat, *Duration*, 15minutes)
one step
Frequency of brown the meat is less than *every minute*
(Brown the meat, *Duration*, 15minutes)
+
two step
Get up may last less than *15 minutes*
(Get up, *while*, Brown the meat)

Figure 1: Reasoning the duration of *Get up* and Frequency of *Brown the meat* with complementary relation between dimensions

However, the automatically obtained TCS can follow a distorted distribution due to the *reporting bias* phenomenon (Gordon and Van Durme 2013; Zhang et al. 2017; Shwartz and Choi 2020; Paik et al. 2021): people rarely report obvious things but sometimes highlight rarities. For example, we can hardly observe mentions of "I get up from the bed in 1 minute", but we may find "Every morning after I wake up, it takes me an hour to get up from the bed" in text. By incorporating such *biased* TCS knowledge, NLP systems can hardly achieve satisfactory performances.

To address this issue, recent studies proposed to perform TCS inference by exploiting the complementary relationship among temporal dimensions. For example, as shown in Figure 1, the second sentence illustrates the *Duration* of the event "brown the meat on all sides" is "15 minutes". Based on this evidence, we can perform one-step inference and infer that the *Frequency* of this event is at most once every few minutes. Additionally, combining the duration information with another statement indicating "get up" happens during "brown the meat" (illustrated by the first sentence), we can make another two-step inference and conclude that the duration of "get up" is less than 15 minutes. Such phenomena among temporal dimensions provides us the opportunity to estimate event's temporal attributes even if such temporal information is never expressed explicitly in text.

However, previous works utilizes such complementary re-

[*] Corresponding author

Dimension	Sentence examples	Temporal Argument	Label Set	Value
Temporal Relation Between Events				
Hierarchy	I *open* the door during fire *alarm.*	during fire alarm	During, When, While, After, Before	During
Unary Event Temporal Attributes				
Duration	He *takes* a break for 15 minutes.	for 15 minutes	Second, Minute, Hour, Day, Week, Month, Year, Decade, Century	Minute
Frequency	Amy *makes* breakfast everyday.	everyday		Day
Upper-bound	I went hiking yesterday.	yesterday		Day
Typical Time	I weak up early in the morning.	in the morning	time of a day (Morning, Afternoon, ...)	Morning
	We *went* to a bar last Friday.	last Friday	day of a week (Monday, Tuesday, ...)	Friday
	I like making snowman in winter.	in winter	season of a year (Spring, Summer, ...)	Winter
	I graduate from school in June.	in June	month of a year (January, February,...)	June

Table 1: Examples of the acquired TCS in five dimensions with cheap supervision. The temporal arguments in sentence are normalized to the *Value*, which is one of the keywords in the dimension's label set.

lationships from a regular multi-task view: they assume that by supervising model to jointly learn all kinds of temporal knowledge simultaneously, model can implicitly capture the complementary relationship to mitigate the reporting bias. The main concern for this line of work is that they have not fully capture the underlying relation betweeen dimensions and is lack of interpretability.

To explicitly model the complementary relationship between dimensions more efficiently and explainable, we propose a neural-logic based framework, abbreviated as SLEER (Soft Logic Enhanced Event temporal Reasoning), which contains two components. The first one is a base model for event encoding and providing the primary inference result for each temporal dimension. Furthermore, we introduce a PSL (Probabilistic Soft Logic) module for regularizing the output of the base model by incorporating pre-defined logic rules. In this manner, the learned temporal distribution is compatible with both temporal mentions and the temporal logic rules, which enables the model to deduce the temporal knowledge of rarely observed events and revise the uncommon special cases from temporal mentions of other dimensions.

Experimental results[1] on four intrinsic datasets show the efficiency of our model to mitigate the reporting bias problem. Furthermore, the improvements on another two extrinsic temporal commonsense understanding task show the capability of unbiased event temporal representation.

Preliminary

Representation of Temporal Common Sense

To reason about temporal concepts of everyday events such as their duration, frequency or relative ordering, NLP systems should be equipped with rich commonsense knowledge about how the world works, especially the *temporal common sense* (TCS) knowledge.

Previous work (Zhou et al. 2020) focuses on five categories of the TCS occurring in text, namely *Event Hierarchy*, *Duration*,*Frequency*, *Typical Time* and *Duration Upper-bound*. The *Event Hierarchy* dimension describes the temporal relation between event-pairs, while the rest dimensions are unary temporal attributes of events. Here, the *duration*, *typical time*, *frequency* dimensions refers to "how long an event takes", "when an event happens" and "how often an event occurs", respectively. The *Duration Upper-bound* represents values that are upper-bounds to an event's duration but not necessarily the exact duration. For example, "did `[activity]` yesterday" indicates something happened within a day.

Following TacoLM (Zhou et al. 2020), we utilize syntactic patterns to identify and normalize the TCS from text as tuples in the form of (`event context`, `value`, `dimension`). Note, the temporal mentions are normalized to one of the keywords in pre-defined label set as `value`. For example, both "15 minutes" and "50 seconds" are normalized to "minute", which is the nearest time unit among the nine labels of duration dimension. The `event context` are described by sentences. As a result, the TCS inference can be formulated as a multi-class classification task upon each dimension's label set.

We list the extraction and normalization examples for each dimension in Table 1.

Probabilistic Soft Logic

We notice there exists rich complementary relationship between different dimensions of TCS, which can be effectively modeled with Probabilistic Soft Logic (Kimmig et al. 2012). Before diving into regularization details of PSL rules, we first introduce some concepts and notations for probabilistic soft logic, and illustrate how logic is applicable to define templates for TCS inference.

Definition 1 (**Atomic Formula**). The atom formula (also known simply as atom), denoted as l, consists of a predicate p together with its arguments. In the soft logic view, each atom take on continuous soft truth value with interval $[0, 1]$.

In this paper, each temporal dimension is served as a predicate: HRCHY (hierarchy), DUR (duration), FREQ (frequency), TYP (typical) and BND (upper-bound). They all take two arguments: event(s) and the normalized value in the corresponding label set.

Example. The atom DUR$(e_1, year)$ describes the statement that "the event e_1 lasts for years."

Definition 2 (**Complex Formula**). All other formulae obtained by composing atoms with logical connectives (e.g.,

[1] Code for reproduction: https://github.com/bibocai/SLEER

Type	Rule Template Example
D⇒F	$\text{DUR}(e_1, \mathit{year}) \Rightarrow \text{FREQ}(e_1, \mathit{decade}) \vee \text{FREQ}(e_1, \mathit{century})$
F⇒D	$\text{FREQ}(e_1, \mathit{hour}) \Rightarrow \text{DUR}(e_1, \mathit{second}) \vee \text{DUR}(e_1, \mathit{minute})$
T⇒F	$\text{TYP}(e_1, \mathit{time\,of\,day}) \Rightarrow \neg\text{FREQ}(e_1, \mathit{hour}) \vee \neg\text{FREQ}(e_1, \mathit{minute}) \vee \neg\text{FREQ}(e_1, \mathit{second})$
T⇒D	$\text{TYP}(e_1, \mathit{time\,of\,day}) \Rightarrow \text{DUR}(e_1, \mathit{second}) \vee \text{DUR}(e_1, \mathit{minute}) \vee \text{DUR}(e_1, \mathit{hour})$
B⇒D	$\text{BND}(e_1, \mathit{hour}) \Rightarrow \text{DUR}(e_1, \mathit{second}) \vee \text{DUR}(e_1, \mathit{minute}) \vee \text{DUR}(e_1, \mathit{hour})$
D⇒T	$\text{DUR}(e_1, \mathit{month}) \Rightarrow \neg\text{TYP}(e_1, \mathit{time\,of\,day})$
F⇒T	$\text{FREQ}(e_1, \mathit{month}) \Rightarrow \neg\text{TYP}(e_1, \mathit{month}) \wedge \neg\text{TYP}(e_1, \mathit{season})$
DD⇒D	$\text{HRCHY}((e_1, e_2), \mathit{during}) \wedge \text{DUR}(e_2, \mathit{minute}) \Rightarrow \text{DUR}(e_1, \mathit{second}) \vee \text{DUR}(e_1, \mathit{minute})$
WD⇒D	$\text{HRCHY}((e_1, e_2), \mathit{while}) \wedge \text{DUR}(e_2, \mathit{minute}) \Rightarrow \text{DUR}(e_1, \mathit{second}) \vee \text{DUR}(e_1, \mathit{minute})$
WT⇒D	$\text{HRCHY}((e_1, e_2), \mathit{when}) \wedge \text{TYP}(e_1, \mathit{morning}) \Rightarrow \text{TYP}(e_1, \mathit{morning})$

Table 2: Temporal PSL Rules

and, or) and quantifiers (e.g., for-all) are named as *complex formula*.

Definition 3 (**Logic Rule**). A logic rule r is an *implication* constructed by combining atoms with logical connectives:

$$\eta_r : f_1 \Rightarrow f_2 \tag{1}$$

where f_1 and f_2 can be either atomic formulae or complex formulae. All the logic rules defined in this paper are *unweighted*, which means they are likely to hold true all the time.

Example. A logic rule presented in the form of $\forall e$: FREQ$(e, hour)$ $\Rightarrow$ DUR$(e, second)$ $\vee$ DUR$(e, minute)$ states that "any event e which happens hourly can only lasts for seconds or few minutes." Specifically, the universally quantified rule can be instantiated with certain event and come with the the *ground rule*, which can be interpreted as a complex formula.

Definition 4 (**Truth Function**). The truth function I is a map: $\mathcal{F} \to [0, 1]$, where $\mathcal{F}$ denotes the set of training formulae, both atomic and complex (ground logic rules). I assign a soft truth value to each formula, indicating the probability that the formula holds. The larger the truth values are, the better the ground rules are satisfied.

Definition 5 (**t-norm Fuzzy Logic**). The t-norm fuzzy logics (Hájek 1998) define the truth value of a complex formula as a composition of the truth values of its constituents through logic connectives. In practice, we utilize the Łukasiewicz t-norm (Hay 1963), as we find it is more numerical stable than other kinds of t-norm logic (e.g., product t-norm) in this task. The compositions associated with logical conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$) are defined as follows:

$$\begin{aligned} I\left(f_1 \wedge f_2\right) &= \max(0, I\left(f_1\right) + I\left(f_2\right) - 1) \\ I\left(f_1 \vee f_2\right) &= \min(1, I\left(f_1\right) + I\left(f_2\right)) \\ I\left(\neg f_1\right) &= 1 - I\left(f_1\right) \end{aligned} \tag{2}$$

where f_1 and f_2 can be either atomic or complex formulae.

Method

In this paper, we propose our SLEER model which explicitly models interactions among temporal dimensions with logic rules. As shown in Figure 2, SLEER contains two components: the multi-task base model and the PSL regularization module. With the regularization of PSL rules on the outputs of base model, SLEER can make logic-coherent TCS inference, which could further benefit the mitigation of reporting bias problem.

Base Model

The base model accepts a sentence describing an event as input and provides a primary inference result (i.e., the temporal distribution upon the label set) on each dimension. It follows a multitask architecture which contains a common event encoder and five take-specific output layers, one for each temporal dimension. Specifically, we propose to adopt the pretrained language model (PLM) (Liu et al. 2019; Devlin et al. 2019) as the backbone encoder to capture the semantic features of the sequence.

For the hierarchy dimension, the input sentence is modified by replacing the five keywords in the label set (i.e., when, while, before, after, during) with the mask token. The output layer accepts the encoder's final hidden state corresponding to this token as input, and outputs a distribution upon the label set. For other dimensions, temporal mentions in the input sentences are removed in the preprocessing procedure. Encoder's last state corresponding to the target event's predicate is fed into the linear output layer to make predictions over the target dimension's label set.

Totally, we obtain five cross-entropy loss functions as $L_{hierachy}$, L_{freq}, L_{dur}, L_{bnd} and L_{typ} for the classifiers of *hierarchy*, *Frequency*, *Duration*, *Upper-bound*, *Typical* dimension, respectively. We denoted these losses collectively as *dimension loss* L_{dim}.

PSL rules Designing

As one of our contributions, we systematically summarize the common temporal PSL rules, to adequately model the interaction among different dimensions of TCS. As shown in Table 2, two types of logic rules are considered in this paper, i.e., one-step logic rules and two-step logic rules. Now we take examples to illustrate the modeling approach for each of them.

One-Step Logic Rules The one-step logic rules focus on describing the interaction among unary temporal attributes of a single event (typical time, duration, etc). For example, we can infer that the event e_1 should occur *at most* once

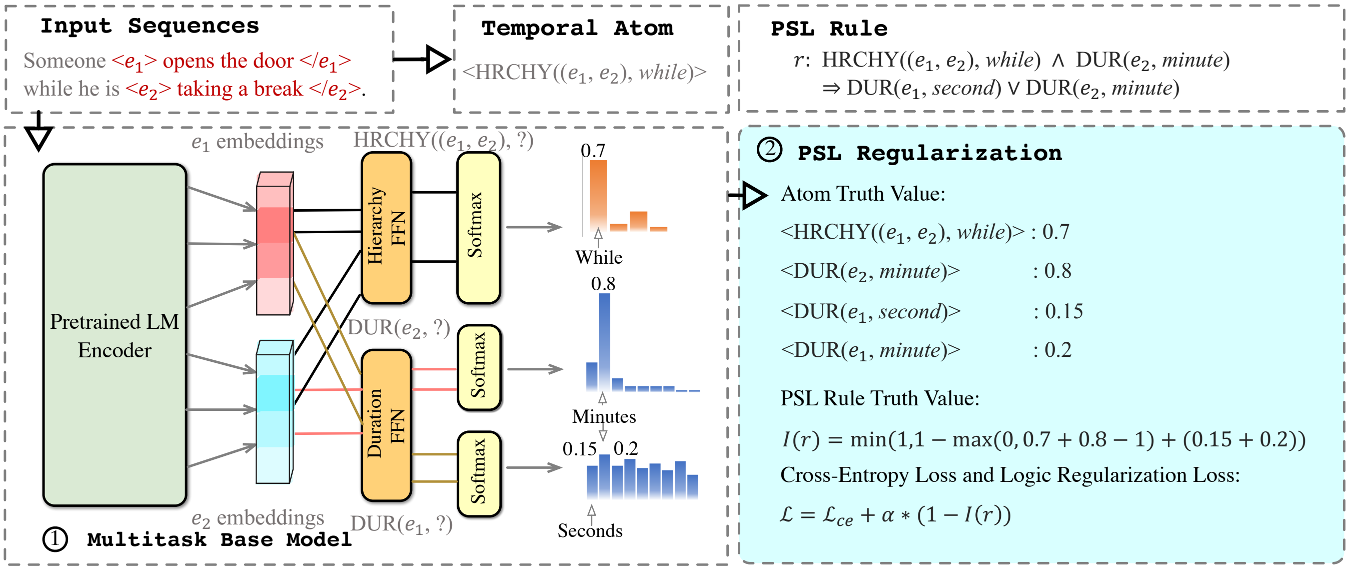
Input Sequences
Someone <e_1> opens the door </e_1>
while he is <e_2> taking a break </e_2>.
Temporal Atom
<HRCHY((e_1, e_2), while)>
PSL Rule
r: HRCHY((e_1, e_2), while) ∧ DUR(e_2, minute)
⇒ DUR(e_1, second) ∨ DUR(e_2, minute)
Pretrained LM Encoder
e_1 embeddings
e_2 embeddings
HRCHY((e_1, e_2), ?)
DUR(e_2, ?)
DUR(e_1, ?)
Hierarchy FFN
Duration FFN
Softmax
Softmax
Softmax
0.7
While
0.8
Minutes
0.15
0.2
Seconds
① Multitask Base Model
② PSL Regularization
Atom Truth Value:
<HRCHY((e_1, e_2), while)> : 0.7
<DUR(e_2, minute)> : 0.8
<DUR(e_1, second)> : 0.15
<DUR(e_1, minute)> : 0.2
PSL Rule Truth Value:
$I(r) = \min(1, 1 - \max(0, 0.7 + 0.8 - 1) + (0.15 + 0.2))$
Cross-Entropy Loss and Logic Regularization Loss:
$\mathcal{L} = \mathcal{L}_{ce} + \alpha * (1 - I(r))$

Figure 2: The architecture of SLEER. The predicted distributions of base model is regularized with pre-defined PSL rules by introducing the logic regularization loss.

every hour, if we know its duration is *hour* (i.e., e_1 lasts for hours). We formulate such logic rule as:

$$\forall e_1 : \mathrm{DUR}(e_1, hour) \Rightarrow \mathrm{FREQ}(e_1, \leq hour) \tag{3}$$

where the atom formula $\mathrm{DUR}(e_1, hour)$ denotes the Duration of event e_1 is hours, and the complex formula $\mathrm{FREQ}(e_1, \leq hour)$ denotes the frequency of event e_1 cannot be more than *once every hour*, so it can be more likely to happens *once every day*, *once every month*, etc. This complex formula can be reduced to smaller units:

$$\begin{aligned} &\mathrm{FREQ}(e_1, \leq hour) \triangleq \neg\mathrm{FREQ}(e_1, > hour) \\ &\triangleq \neg\mathrm{FREQ}(e_1, minute) \wedge \neg\mathrm{FREQ}(e_1, second) \end{aligned} \tag{4}$$

which means the statement "the frequency of the event e_1 is less than every *hour*" is equivalent to "the event e_1 can happens neither *once every minute* nor *once every second*."

Generally, a *ground* one-step logic rule is in the format as $R \triangleq l \Rightarrow f$. By applying *Łukasiewicz t-norm*, the truth value can be calculated as:

$$I(R) = \min(1, 1 - I(l) + I(f)) \tag{5}$$

where l is an atom and f can be either atomic or complex formula.

Two-Step Logic Rules The two-step logic rules describe the complementary relationship across events in the same dimension. For example, although we may not observe the duration of the event "open my door" in free-form text, we may observe the text like "Someone opens the door while he is taking a break" and "Taking a 15 minutes break in the afternoon makes you feel good". The first sentence gives the *duration inclusion* relation between the events, which indicates the duration of "open my door" is no longer than "take a break." The second sentence describes the duration of the event "take a break". By combining the two evidences, we can conclude that the duration of "open my door" should be less than minutes. This inference procedure can be induced as logic rules:

$$\begin{aligned} \forall e_1, e_2 : &\mathrm{HRCHY}((e_1, e_2), during) \wedge \mathrm{DUR}(e_2, minute) \\ &\Rightarrow \mathrm{DUR}(e_1, \leq minute) \end{aligned} \tag{6}$$

where the atom formula $\mathrm{HRCHY}((e_1, e_2), during)$ denotes the hierarchy relation between event e_1 and e_2: e_1 happens *during* event e_2 and $\mathrm{DUR}(e_2, minute)$ denotes e_2 lasts for minutes. The complex formula $\mathrm{DUR}(e_2, \leq minute)$ represents the duration of event e_2 is less than minutes, which can also be decomposed in the same manner as Eq. 4.

Generally, the two-step logic rules are in the form of $F \triangleq l_1 \wedge l_2 \Rightarrow f$, whose truth value can be calculated as:

$$I(F) = \min(1, 1 - I(l_1 \wedge l_2) + I(f)) \tag{7}$$

where l_1 and l_2 are atom formulae, f can be either atomic or complex. $I(l_1 \wedge l_2)$ can be calculated by applying Eq. 2.

Regularization with PSL Rules

We aim to make the predicted distributions be compatible with both temporal mentions and the temporal logic rules. To this end, the *t-norm fuzzy logic* is used to translates logical constraints into continuous almost-everywhere differentiable loss functions.

Specifically, given a training tuple `(event context, dimension, value)`, we first collect all the potential PSL rules, denoted as $R = \{r_1, r_2, ..r_n\}$, with n as the number of corresponding PSL rules. The predicted probabilities of base model are regarded as the soft truth value of the ground atom in each r_i. The truth value $I(r_i)$, which denotes the degree that the ground rules r_i are satisfied by the predicted distributions, can be derived by incorporating Eq.2, Eq.5 and Eq.7.

The larger the truth values are, the better the ground rules are satisfied. Based on this principle, we then formulate the distance of $I(r_i)$ to be true as a regularization term, to penalize the distorted distribution that violate the rule r_i. Totally, the logic loss is derived by:

$$L_{logic} = \Sigma_{i=1}^{n}(1 - I(r_i)) \quad (8)$$

We apply a weighted sum of dimension loss L_{dim} and logic loss L_{logic} to obtain the final loss L:

$$L = L_{dim} + \alpha * L_{logic} \quad (9)$$

where hyper-parameters α are employed to control the trade-off among losses. The goal is to minimize L during training.

Experiment

Dataset

We conduct experiments on two kinds of evaluation tasks: **1) The Intrinsic Evaluation** task is a temporal value recovery task, where the inputs are a sentence representation the event, an index to the event's verb, and a target dimension. The goal is to recover the temporal value of the given event in the given dimension in a *zero-shot manner*. **2) The Extrinsic Evaluation** tasks are TimeML and MCTACO tasks, which require model's implicitly understanding of typical event temporal commonsense to give right prediction. We briefly describe the datasets we used in this paper below.

Intrinsic Evaluation Datasets Each instance of intrinsic datasets describes a piece of TCS in the format as (event context, temporal value, dimension). Following (Zhou et al. 2020), we conduct intrinsic evaluation on the filtered RealNews and the filtered UDS-T datasets. As the size of the two datasets are relatively small, we transform a subset of the other two widely used temporal datasets, the train dataset of TimeML and the frequency subset of MCTACO to the target data format for further zero-shot evaluation, denoted as TimeML and MCTACO-freq. The details are described in the next paragraph. An overview of the data statistics and related temporal dimensions are shown in Table 3.

TimeML The TimeML (Saurí et al. 2006; Pan, Mulkar, and Hobbs 2006) is an ACE corpus with event duration annotated as lower and upper bounds. The task is to decide whether a given event has a duration longer or shorter than a day. For each data instance in the training data of TimeML (Gusev et al. 2011), we normalize the average of its lower bound and upper bound to the nearest time unit to construct the TimeML-train intrinsic evaluation dataset. For extrinsic evaluation, we follow the same data split with (Zhou et al. 2020).

MCTACO We also evaluate on MCTACO (Zhou et al. 2019), which is a temporal question answering dataset which requires comprehensive understanding of temporal common sense and reasoning. Note, there are five types of temporal questions in MCTACO, including three kinds of event temporal attributes: *Duration*, *Frequency*, *Typical Time*, and the other two is *Event Ordering* and *Stationary*. We only use the event temporal attributes typed QA pair for extrinsic evaluation, because our scope is focused on mitigating the reporting bias problem of temporal attributes of event, while the biases of *Event Ordering* and *Stationary* of events are less concerned about in our setting. Similar to (Yang et al. 2020), we obtain the intrinsic evaluation dataset MCTACO-freq by converting the questions to statements and normalizing the answer to nearest time unit for each *correct* question-answer pair in *frequency* dimension.

Dimension	#Typical	#Dur	#Freq
TimeML-tr	-	1664	-
MCTACO-freq	-	-	516
RealNews	200	50	50
UDS-T	-	142	-

Table 3: Intrinsic Data Statistic

Baselines

We compare our SLEER model with several baselines. Each of these models are:

BERT and RoBERTa We report the performance of pre-trained model including BERT-base (Devlin et al. 2019) and RoBERTa-base (Liu et al. 2019).

TacoLM (Zhou et al. 2020) is a temporal common sense language model based on BERT-base, which is the first work to exploit the complementary relation among temporal dimensions, but it suffers from implicit, inadequate and unexplainable interaction modeling, which may limit model's performance.

Evaluation Metrics

For the intrinsic evaluation on filtered RealNews and filtered UDS-T, following (Zhou et al. 2020; Vashishtha, Van Durme, and White 2019), we employ the *distance* metric which measures the ranking difference between the system's top prediction and the gold label with respect to an ordered label set, so that "minutes" will have a distance 1 with "hours." We report the averaged number across instances. For TimeML, which can be formulated as a sequence binary classification task, we report *accuracy*. As for MCTACO (Zhou et al. 2019) , we adopt *exact match* (EM) and *F1* for evaluation. EM measures how many questions a system can correctly label all candidate answers, while F1 is more relaxed and measures the average overlap between one's predictions and the ground truth.

Experimental Settings

For fair comparison, we utilize the same syntactic pattern based TCS tuple collection method with TacoLM (Zhou et al. 2020). 4 million free-form sentences containing TCS are collected from the entire Gigaword corpus and are utilized for pretraining.

We implement SLEER with the same PLM encoder with baseline methods. During pretraining stage, we employ a

Systems	TimeML-tr	MCTACO-freq	RealNews						UDS-T
	Dur	Freq	Dur	Freq	Day	Week	Month	Season	Dur
BERT (Devlin et al. 2019)	1.72	1.10	1.33	1.68	1.75	1.53	3.78	0.87	1.77
BERT + finetune	1.48	1.49	1.19	1.57	**1.58**	1.58	3.58	0.96	1.70
TacoLM (Zhou et al. 2020)	1.61	1.28	0.75	1.17	1.72	1.19	3.42	**0.63**	**1.49**
SLEER	**1.39**	**1.06**	**0.75**	**1.03**	1.61	**1.08**	**3.17**	0.88	1.55

Table 4: Performance on intrinsic evaluations. The metric is the distance between the prediction and the gold label, smaller values indicate better performance.

learning rate of 2e-5 with 2 epochs on the whole unsupervised dataset and set the hyper-parameter α for the T⇒F, DD⇒D, T⇒D, B⇒D types of rules as 0.8 and the others are set as 0.1. Other parameters are the same as those in the pretrain language model model that the SLEER is based on.

To evaluate the correctness of temporal knowledge stored in SLEER, the trained model is evaluated the on the four intrinsic evaluation datasets in a zero-shot manner. The recovered temporal keywords are ranked in the given dimension's label set. To evaluate the capability of the unbiased event temporal representation, the trained SLEER model is further finetuned on the TimeML and MCTACO dataset. For TimeML, SLEER is finetuned with a 5e-5 learning rate and 3 epochs. For MCTACO, learning rate and epoch are set as 2e-5 and 5, respectively. Each reported number is an average from 3 runs initialized with random seeds $(30, 45, 60)$.

Intrinsic Evaluation

We perform the temporal value recovery task in a zero-shot manner to directly measure whether the model masters accurate temporal commonsense knowledge. Besides the two kinds of baselines mentioned above, we also compare SLEER with the BERT + *finetune* baseline, which is BERT finetuned on the same pre-training data used for the proposed models, but with a probability of 1 masking the temporally related keyword (i.e., all values we used in all dimensions) and 0.15 for other words.. We follow the same sequence formulation method with previous work (Zhou et al. 2020) for BERT and TacoLM. For ours, we utilize the transformer's output of the corresponding target event's predicate as event embedding and feed it to the pretrained linear output layer for temporal value prediction, no additional words are needed.

The results of intrinsic evaluation are shown in Table 4. We observe that:

(1) TacoLM is mostly better than the naive BERT based methods, which proves the advantage of modeling the joint relationship among temporal dimensions.

(2) Furthermore, by modeling the complementary relation explicitly with PSL rules, our SLEER model can outperform or be comparable with TacoLM. Note, prominent improvements are made on the two large datasets, TimeML-tr and McTaco-freq, which is 10 times larger than that in RealNews and UDS-T).

	TimeML	**MCTACO**[2]	
Model	Accuracy	EM	F1
BERT-Based Model Results			
BERT (Devlin et al. 2019)	73.7	39.6	66.7
TacoLM (Zhou et al. 2020)	81.7	40.0	67.2
SLEER (BERT)	**83.0**	**40.9**	**67.3**
RoBERTa-Based Model Results			
RoBERTa (Liu et al. 2019)	81.1	41.3	67.3
SLEER (RoBERTa)	**83.8**	**43.0**	**69.0**

Table 5: Finetuning results of SLEER on the two time-related downstream tasks

Extrinsic Evaluation

We evaluate the capability of unbiased event temporal representation by finetuning the system on downstream time-related tasks. As shown in Table 5, on the task of TimeBank Classification and MCTACO, we can observe:

(1) In the line of method implemented with BERT, our SLEER model consistently outperforms BERT and TacoLM model over all metrics. This is possibly due to the high diversity of events in TimeML, in which there exists some events whose temporal attributes may be rarely explicitly mentioned. Hence, TacoLM and BERT cannot learn the temporal knowledge of the event from free-form text, while our method can alleviate such kind of impact by fully exploit the underlying complementary relationship among temoral dimensions.

(2) The RoBERTa-base baseline implemented by ourselves outperforms the BERT-based model by a large margin. This suggests that RoBERTa can benefit from more data and compute power to store more temporal knowledge than BERT. Moreover, we can see that the SLEER implemented with RoBERTa can still make further improvement. This suggests that the improvement brought by the PSL rules does not disappear when a LM pretrained on a larger free-form text (160GB text) is used. This confirms the existence of reporting bias problem. There is a discrepancy between the reality and the knowledge learned by pretrained LM from large textual data.

[2] We use a subset of MCTACO data, including the Duration, Typical Time, Frequency dimension.

Model	TimeML-tr	McTaco-fre
BERT	1.72	1.10
BASE	1.80	1.10
+Hard Rule	1.52	1.09
+Soft Rule (SLEER)	**1.39**	**1.06**

Table 6: Performance of Ablation Study. The metric is the "*distance*" between the predicted label with gold label, which is similar to MAE, the lower the better.

Ablation Study

Moreover, three variants of our method are evaluated with the two large intrinsic datasets to demonstrate the effectiveness of regularizing the base model with soft logic:

BASE is our base model. It learns multiple dimensions of TCS in a multitask manner, without explicitly modeling relations between different dimensions.

BASE+Hard Rule injects the logical knowledge into the BASE model, but in a hard way, i.e., we augment the collected TCS data with our proposed rules, then the collected data and the augmented data are together used to train the BASE model, without utilizing probabilistic soft logic.

BASE+Soft Rule is the full SLEER model, which injects logical knowledge into the model by employing soft relaxations of Boolean formulas to explicitly model the complementary relation between dimensions.

As we can observe, pretraining with hard rules can lead to better performance over the two datasets, this proves that the logic introduced in this paper is useful. The SLEER model further provides a noticeable improvement with the help of soft rules modeled by fuzzy logic. The possible reasons are that the hard rule based augmentation is manipulated on data instance-level, which may introduce additional noise. The cross entropy loss is taken for both original data and augmented data, overlooking the intrinsic variations between them. The soft rules focus on the logic relation between origin data and augmented data, which attempt to regularize the prediction results of model by maximize the truth value of predefined logic rules. The optimization objective can be more robust and effective for the noisy data collected with cheap supervision.

Related Work

Understanding time in NLP

Time is one of the core questions in event-level language understanding and has long been studied for decades. Traditional research topic focused on the temporal expressions understanding (Vashishtha, Van Durme, and White 2019), temporal relationship understanding (Ning, Feng, and Roth 2017; Han, Ning, and Peng 2019), etc. This line of work mainly relies on the local context understanding which is not the focus of this work.

Recently, significant works have been done on the temporal common sense (TCS) reasoning. These works include but are not limited to event duration prediction (Pan, Mulkar, and Hobbs 2006; Vashishtha, Van Durme, and White 2019), scripting learning (i.e., what happens next after the certain events) (Li, Ding, and Liu 2018), event infilling (i.e., predict the implicit event in a temporally-ordered event sequence) (Lin, Chambers, and Durrett 2021; Zhou et al. 2021a) and various temporal reasoning based question answering tasks (Zhou et al. 2019; Qin et al. 2021). As human annotation on the TCS is costly, a surge of works harnesses cheap supervision methods to collect large amount of TCS data and learn reasoning model upon it. For example, (Lin, Chambers, and Durrett 2021) utilizes narrative documents corpus to automatically construct data for temporal ordering and event infilling task. (Zhou et al. 2020) jointly models three key dimensions of TCS (duration, frequency, and typical time) and the other two auxiliary dimension of TCS, the data of which is also mined from unannotated free text. Our work has two notable differences with this line of work. First, we focus on the reporting bias problem when harnessing the cheap supervision signals rather than the acquisition method. Second, we explicitly reveal the fine-grained complementary structure among different dimensions of TCS with PSL, while other works either focus on single dimension, or model the relationship among dimensions in an implicit manner.

Probabilistic Soft Logic

The soft probabilistic logic combines logic's expressive power with the ability to deal with uncertainty, which has been introduced in a variety of reasoning tasks ranging from Knowledge Base Completion (Yang, Yang, and Cohen 2017; Chen et al. 2019; Mohler, Monahan, and Tomlinson 2020) and Social Prediction (Wang et al. 2020) to Temporal Relation Extraction (Zhou et al. 2021b) and Causal Inference (Sridhar and Getoor 2016; Du et al. 2021). Most of the works inject the logic knowledge to neural networks by introducing logic-driven loss functions. (Li et al. 2019) introduce consistency-based regularization incorporating the first-order logic rules for NLI. (Asai and Hajishirzi 2020) regularize question answering systems with symmetric consistency and symmetric consistency. We are the first to leverage PSL for temporal common sense reasoning. We specially design the PSL rules to declare the complementary relation between different dimension of TCS to tackle with the reporting bias problem.

Conclusion

In summary, we explore methods to mitigate the reporting bias problem when training time understanding model with cheap supervision. With the help of predefined logic rules, our model can be more efficient to learn the complex complementary relationship between different dimensions of TCS. Extensive experimental results show that our SLEER model can improve fine-tuning performances on 4 commonly used temporal commonsense reasoning datasets. The improvements of our methods show the reporting bias problem in the free-form text should be carefully concerned when incorporating TCS acquired from large scale unsupervised text to NLP systems, which points out a promising research direction for unbiased knowledge acquisition and reasoning methods in this research area.

Acknowledgements

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the Technological Innovation "2030 Megaproject" - New Generation Artificial Intelligence of China (2018AAA0101901), and the National Natural Science Foundation of China (62176079, 61976073).

References

Asai, A.; and Hajishirzi, H. 2020. Logic-Guided Data Augmentation and Regularization for Consistent Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5642–5650. Online: Association for Computational Linguistics.

Chen, X.; Chen, M.; Shi, W.; Sun, Y.; and Zaniolo, C. 2019. Embedding uncertain knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, L.; Ding, X.; Xiong, K.; Liu, T.; and Qin, B. 2021. ExCAR: Event Graph Knowledge Enhanced Explainable Causal Reasoning. In *ACL*.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30.

Gusev, A.; Chambers, N.; Khilnani, D. R.; Khaitan, P.; Bethard, S.; and Jurafsky, D. 2011. Using Query Patterns to Learn the Duration of Events. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Hájek, P. 1998. *Metamathematics of fuzzy logic*, volume 4. Springer Science & Business Media.

Han, R.; Ning, Q.; and Peng, N. 2019. Joint Event and Temporal Relation Extraction with Shared Representations and Structured Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 434–444. Hong Kong, China: Association for Computational Linguistics.

Hay, L. S. 1963. Axiomatization of the infinite-valued predicate calculus1. *The Journal of Symbolic Logic*, 28(1): 77–86.

Kimmig, A.; Bach, S.; Broecheler, M.; Huang, B.; and Getoor, L. 2012. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, 1–4.

Li, T.; Gupta, V.; Mehta, M.; and Srikumar, V. 2019. A Logic-Driven Framework for Consistency of Neural Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3924–3935. Hong Kong, China: Association for Computational Linguistics.

Li, Z.; Ding, X.; and Liu, T. 2018. Constructing Narrative Event Evolutionary Graph for Script Event Prediction. In *IJCAI*, 4201–4207. ijcai.org. ISBN 978-0-9992411-2-7.

Lin, S.-T.; Chambers, N.; and Durrett, G. 2021. Conditional Generation of Temporally-ordered Event Sequences. *Proc. of ACL*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mohler, M.; Monahan, S.; and Tomlinson, M. 2020. Modeling Procedural State Changes over Time with Probabilistic Soft Logic. In *The Thirty-Third International Flairs Conference*.

Ning, Q.; Feng, Z.; and Roth, D. 2017. A Structured Learning Approach to Temporal Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1027–1037. Copenhagen, Denmark: Association for Computational Linguistics.

Ning, Q.; Wu, H.; Han, R.; Peng, N.; Gardner, M.; and Roth, D. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1158–1172. Online: Association for Computational Linguistics.

Ning, Q.; Wu, H.; Peng, H.; and Roth, D. 2018. Improving Temporal Relation Extraction with a Globally Acquired Statistical Resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 841–851. New Orleans, Louisiana: Association for Computational Linguistics.

Noah Weber, B. V. D., Rachel Rudinger. 2020. Causal Inference of Script Knowledge. In *EMNLP*.

Paik, C.; Aroca-Ouellette, S.; Roncone, A.; and Kann, K. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 823–835. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Pan, F.; Mulkar, R.; and Hobbs, J. R. 2006. Extending TimeML with Typical Durations of Events. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, 38–45. Sydney, Australia: Association for Computational Linguistics.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TimeDial: Temporal Commonsense Reasoning in Dialog. In *ACL*.

Saurí, R.; Littman, J.; Knippen, B.; Gaizauskas, R.; Setzer, A.; and Pustejovsky, J. 2006. TimeML annotation guidelines. *Version*, 1(1): 31.

Shwartz, V.; and Choi, Y. 2020. Do Neural Language Models Overcome Reporting Bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, 6863–6870. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Sridhar, D.; and Getoor, L. 2016. Joint probabilistic inference of causal structure. In *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining Workshop on Causal Discovery*.

Vashishtha, S.; Van Durme, B.; and White, A. S. 2019. Fine-Grained Temporal Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2906–2919. Florence, Italy: Association for Computational Linguistics.

Wang, R.; Tang, D.; Duan, N.; Zhong, W.; Wei, Z.; Huang, X.; Jiang, D.; and Zhou, M. 2020. Leveraging declarative knowledge in text and first-order logic for fine-grained propaganda detection. *EMNLP*.

Yan, R.; Kong, L.; Huang, C.; Wan, X.; Li, X.; and Zhang, Y. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 433–443.

Yang, F.; Yang, Z.; and Cohen, W. W. 2017. Differentiable learning of logical rules for knowledge base reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2316–2325.

Yang, Z.; Du, X.; Rush, A.; and Cardie, C. 2020. Improving Event Duration Prediction via Time-aware Pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3370–3378. Online: Association for Computational Linguistics.

Zhang, S.; Rudinger, R.; Duh, K.; and Van Durme, B. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics*, 5: 379–395.

Zhao, X.; Lin, S.-T.; and Durrett, G. 2021. Effective Distant Supervision for Temporal Relation Extraction. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, 195–203.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. “Going on a vacation” takes longer than “Going for a walk”: A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3363–3369. Hong Kong, China: Association for Computational Linguistics.

Zhou, B.; Ning, Q.; Khashabi, D.; and Roth, D. 2020. Temporal Common Sense Acquisition with Minimal Supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7579–7589. Online: Association for Computational Linguistics.

Zhou, B.; Richardson, K.; Ning, Q.; Khot, T.; Sabharwal, A.; and Roth, D. 2021. Temporal Reasoning on Implicit Events from Distant Supervision. In *NAACL*.

Zhou, B.; Richardson, K.; Ning, Q.; Khot, T.; Sabharwal, A.; and Roth, D. 2021a. Temporal Reasoning on Implicit Events from Distant Supervision. *NAACL*.

Zhou, Y.; Yan, Y.; Han, R.; Caufield, J. H.; Chang, K.-W.; Sun, Y.; Ping, P.; and Wang, W. 2021b. Clinical Temporal Relation Extraction with Probabilistic Soft Logic Regularization and Global Inference. *AAAI 2021*.