Time-aware COMET: a Commonsense Knowledge Model with Temporal Knowledge

Eiki Murata, Daisuke Kawahara
Waseda University
{eiki.1650-2951@toki., dkw@}waseda.jp

Abstract

To better handle commonsense knowledge, which is difficult to acquire in ordinary training of language models, commonsense knowledge graphs and commonsense knowledge models have been constructed. The former manually and symbolically represents commonsense, and the latter stores these graphs' knowledge in the models' parameters. However, the existing commonsense knowledge models that deal with events do not consider granularity or time axes. In this paper, we propose a time-aware commonsense knowledge model, TaCOMET. The construction of TaCOMET consists of two steps. First, we create TimeATOMIC using ChatGPT, which is a commonsense knowledge graph with time. Second, TaCOMET is built by continually finetuning an existing commonsense knowledge model on TimeATOMIC. TimeATOMIC and continual finetuning let the model make more time-aware generations with rich commonsense than the existing commonsense models. We also verify the applicability of TaCOMET on a robotic decision-making task. TaCOMET outperformed the existing commonsense knowledge model when proper times are input. Our dataset and models are available at https://github.com/nlp-waseda/TaCOMET.

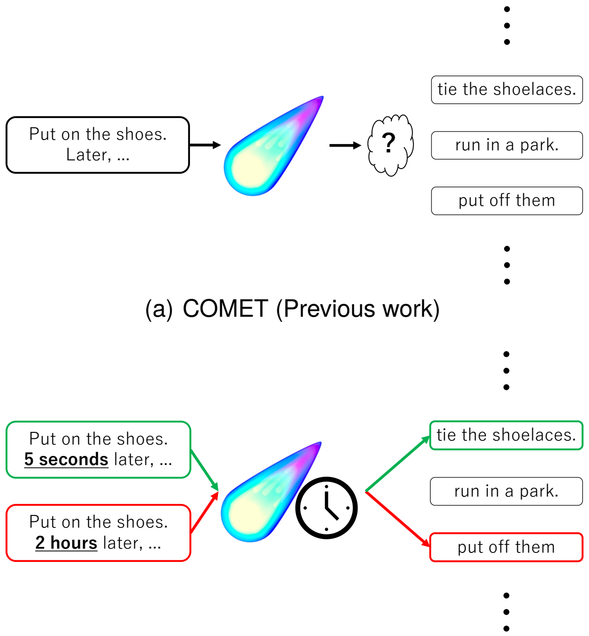
Keywords: Commonsense Knowledge, Temporal Knowledge, Knowledge Graph/Base

1. Introduction

Recent language models have acquired linguistic knowledge (Hewitt and Manning, 2019; Manning et al., 2020) and factual knowledge (Petroni et al., 2019; AlKhamissi et al., 2022) through pretraining. This process includes causal language modeling (Radford et al., 2019; Brown et al., 2020), masked language modeling (Devlin et al., 2019), and other methods (Lewis et al., 2020; Raffel et al., 2020). However, because language models learn only from the surface of the language written by human beings, they cannot deal well with commonsense knowledge, which tends to be tacit (Zhou et al., 2020b; Hwang et al., 2021).

Therefore, some symbolic commonsense knowledge graphs have been constructed, in which commonsense knowledge is mainly collected manually. Typical examples include ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). The former mainly deals with entity relations, while the latter deals with events and mental states. The latter knowledge graph, ATOMIC, includes event-to-event inferences, i.e., "what events follow an event" and "what events should have occurred before an event". Because a commonsense knowledge graph collects knowledge symbolically, its coverage is finite. Thus, commonsense knowledge models such as COMET (Bosselut et al., 2019) have also been proposed, which store the graph's knowledge in the models' parameters.

The shift from symbolic knowledge graphs to neural models has alleviated the coverage problem, but they still have problems. One of the problems is that existing commonsense knowledge models do not consider the granularity of events, even though every event has a granularity. Granularity has various aspects, such as a part-whole relation of entities or events, a cause-effect relation, the duration of an event, and the time between events (Mulkar-Mehta et al., 2011). As an example of the part-whole relation of events, an event symbolically expressed in natural language as "go to



(a) COMET (Previous work)

(b) TaCOMET (Ours)

Figure 1: Concept of our proposal. 1(a) Ambiguity exists because previous work does not address granularity or time. 1(b) Our model clarifies points in the event space by specifying the time.

LREC-COLING 2024, pages 16162–16174
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

school" is possibly interpreted with different granularity, such as (1) "visit the school building" or (2) "get an education" (Mani, 1998). When considering the events that follow "go to school", ambiguity arises in the inference depending on the interpretation of granularity, such as "go to the library" for interpretation (1) and "get a job" for interpretation (2).

However, existing commonsense knowledge graphs and models do not consider the granularity of events. Because of this, the models cannot determine which events to generate, as described above. Even in automatic evaluation, although multiple correct answers with different levels of granularity exist, only one correct answer is currently handled.

To this end, we propose a time-aware commonsense knowledge model, **TaCOMET** (**T**ime-**a**ware **COMET**). We focus on the interval time between events as one of the most important aspects of event transition granularity. TaCOMET is a commonsense knowledge model that allows inference generation to be controlled by the time input.

The construction of the model consists of two steps. First, we create a dataset, **TimeATOMIC**, using ChatGPT[1] by augmenting existing knowledge graphs with time between events. Second, an existing COMET model is continually finetuned on TimeATOMIC. This simple but effective method builds a model that generates event inferences corresponding to the input time.

We conduct experiments in Japanese and English. Automatic and manual evaluations showed that the proposed model generates acceptable inferences corresponding to the input time.

The effectiveness of the constructed model was also verified in a downstream task that makes decisions for interactive robots. Considering the application of language models to decision-making in robotics (Ahn et al., 2022; Wu et al., 2023; Driess et al., 2023), commonsense inference controlled by granularity would be useful. For robots, reasoning on the scale of months or years is meaningless, but that on the scale of seconds to minutes is meaningful. Our model outperformed the model without time input, and we also observed that inputting an appropriate time for the robot improved performance.

2. Related Work

2.1. Commonsense Knowledge Graphs

Symbolic Graphs Various commonsense knowledge graphs have been constructed to deal with tacit knowledge that does not appear on the surface of the text.

WordNet (Miller, 1994) and ConceptNet (Speer et al., 2017) are the ones of linguistic or commonsense knowledge graphs, which deal with relations among entities. GenericsKB (Bhakthavatsalam et al., 2020) is a commonsense knowledge base in a natural language format, though not in a graph. A question-answering dataset based on ConceptNet was also constructed (Talmor et al., 2019).

We deal with commonsense between events in this paper. ATOMIC (Sap et al., 2019) is a commonsense knowledge graph that focuses on the relations between events. There is also ATOMIC2020 (Hwang et al., 2021), which is an extended graph of ATOMIC that integrates ConceptNet. ATOMIC stores if-then relations between events and mental states in the form of triples, such as ("X makes Y's coffee", xEffect, "X gets thanked"). In this paper, the first item of the triple, the source event of inference, is called ***head***, the second item, the type of inference, is called ***relation***, and the third item, the target event of inference, is called ***tail***.

To alleviate the coverage problem of the handcrafted commonsense knowledge graphs, automatic construction methods have been proposed. For instance, ASER (Zhang et al., 2019, 2020a) is a graph that was automatically acquired from a corpus using rules. ATOMIC10x (West et al., 2022) was constructed by extending ATOMIC2020 with GPT-3 (Brown et al., 2020). Dense-ATOMIC (Shen et al., 2023) mitigated the coverage problem by predicting additional relations in existing sparse commonsense knowledge graphs.

The above symbolic graphs have the challenges of high construction cost and finite coverage.

Neural Models By training language models such as GPT-2 (Radford et al., 2019) on symbolic graphs, commonsense knowledge models have been built to store the knowledge of commonsense knowledge graphs in their parameters. They alleviate the coverage limitations of the symbolic graphs. They can also be regarded as automatic completion of the commonsense knowledge graphs. COMET (Bosselut et al., 2019) is a commonsense knowledge model trained with ATOMIC and ConceptNet, which can generate commonsense inferences for unseen events that are not included in the original knowledge graph. COMET is trained to generate a *tail* when given a *head* and a *relation*.

As extended models of COMET, there are COMET2020 (Hwang et al., 2021) and $\text{COMET}^{\text{DIS}}_{\text{TIL}}$ (West et al., 2022) trained on ATOMIC2020 and ATOMIC10x, respectively. $\text{COMET}^{\text{DIS}}_{\text{TIL}}$ can generate commonsense inferences more accurately than GPT-3 even though its base model is GPT-2.

[1] https://chat.openai.com/

There are also models for paragraph-level reasoning (Gabriel et al., 2021) and non-English studies (Ide et al., 2023; Wang et al., 2022).

Bhagavatula et al. (2020) introduced a conditional generation task for explaining given observations in natural language. They also used COMET's embeddings for the baseline model based on GPT-2 to improve its performance on commonsense inference. Zhou et al. (2023) reported that transferring COMET knowledge to a general language model improved the performance on commonsense tasks while retaining general language skills.

2.2. Granularity and Temporal Knowledge

Granularity of Events Every event has a granularity and human beings judge it unconsciously. Mulkar-Mehta et al. (2011) modeled the granularity of events and phrases. They identified the following three elements of granularity.

1. Part-whole relationships between entities.
2. Part-whole relationships between events.
3. Causal relationships between the fine and coarse granularities.

They used the following example to illustrate these elements.

> *The San Francisco 49ers moved ahead 7-3 11 minutes into the game when William Floyd scored a two-yard touchdown run.*

The player and the team are in a part-whole relationship between entities. The phrases "moved ahead" and "scored" are in a part-whole relationship between events, and the player's scoring causes the team's lead.

To deal with the granularity of the commonsense knowledge graph of events, we focus on the second element and regard the granularity as the time between events. More details on the reasons for this are given in Section 3.1.

Temporal Knowledge Temporal knowledge has also been studied. Some corpora are annotated with the order between events (Pustejovsky et al., 2003; Cassidy et al., 2014; Kolomiyets et al., 2012; Reimers et al., 2016), and others map events to a time axis (Huang et al., 2016; Asakura et al., 2016). However, these time axes are divided into macroscopic temporal units, such as "days" or "years", which are unsuitable for commonsense inference.

Benchmarks to assess temporal commonsense have also been created. TempEval (Verhagen et al., 2007, 2010; UzZaman et al., 2013) are adopted in SemEval-{2007, 2010, 2013}. MC-TACO (Zhou et al., 2019) is a temporal commonsense benchmark that addresses five temporal commonsense: frequency, duration, stationarity, ordering, and typical time. TRACIE (Zhou et al., 2021), a benchmark focusing on more tacit knowledge, and TimeDial (Qin et al., 2021), which uses richer context within the dialogue, were also created.

TACOLM (Zhou et al., 2020a) is a language model that considers temporal commonsense and is superior to BERT (Devlin et al., 2019) and others in temporal tasks. Dhingra et al. (2022) dealt with time from a more macro perspective and proposed a model that deals with factual knowledge that changes over time. Unlike these models, we propose to add temporal knowledge to an event-to-event commonsense knowledge model described in Section 2.1.

3. Time-aware COMET

The existing commonsense knowledge graphs and models for events do not deal with the granularity of events. We propose TaCOMET (**T**ime-**a**ware **COMET**), a commonsense knowledge model with temporal knowledge. Handling temporal granularity alleviates the coverage problem in the *tail* event space of the existing models and also broadens its application to downstream tasks.

3.1. Task

We adopt the interval time between events as granularity. Among the three granularity elements listed in Section 2.2, we focus on the part-whole relationships between events as the granularity because of the following reasons. The part-whole relationship between entities is not appropriate as granularity because events are retained in a format that fixes the subject as X, such as "X gets X's car repaired" in ATOMIC. Since the causal relation is not an independent element of granularity but is influenced by the other two, it is also difficult to treat it as granularity. Specifically, since the commonsense knowledge graph deals with event-to-event transitions, we adopt the interval time between events as a quantitative element of the part-whole relationships between events.

In the existing commonsense knowledge model, among a triple (*head*, *relation*, *tail*) in the commonsense knowledge graph, *head* and *relation* are input, and *tail* is output. Our task is to output an appropriate *tail* given *head*, *relation*, and interval time.

To handle event transitions, we focus on the following two typical *relations*.

- **xNeed**: What X would do before the event.
- **xEffect**: What X does after the event.

Even with the same *head* and *relation*, if different interval times are given, the model is required to generate different *tails*.

3.2. Method

To build TaCOMET, we integrate temporal knowledge with an existing COMET model. Specifically, we first construct a temporal event-to-event dataset, TimeATOMIC, and then apply continual finetuning to COMET on TimeATOMIC.

To acquire TimeATOMIC as the quadruple (*head*, *relation*, time, *tail*), we feed *head*, *relation*, and time into ChatGPT to generate *tail*. By this process, multiple *tails* with different times can be obtained for the same *head*/*relation* pair.

Then, we finetune the existing COMET model continually on TimeATOMIC. COMET has been trained to generate a *tail* for a given input of *head* and *relation*. In other words, COMET maximizes $\mathrm{P}(tail|head, relation)$, while TaCOMET maximizes $\mathrm{P}(tail|head, relation, time)$. Since there are instances of the same *head*/*relation* pair with different times, time-aware generation can be achieved.

3.3. Why not ChatGPT but TaCOMET?

One might think that if the model is built on the knowledge extracted from ChatGPT, then ChatGPT could be used directly. The advantages of our method include the following three points.

Tacit Knowledge COMET, the base model of TaCOMET, is trained on commonsense knowledge graphs, which have been manually constructed using crowdsourcing and other methods. It captures tacit knowledge that does not appear on the surface of the text, which cannot be obtained simply by training on text corpora (Zhou et al., 2023). Continual training of this model integrates human-derived commonsense knowledge with the temporal knowledge gained from ChatGPT.

Soft Output Soft labels, such as the probability distribution of an output, rather than hard labels, such as a generated text, are useful for knowledge transfer, such as knowledge distillation (Hinton et al., 2015). The logits (soft labels) of LLMs whose parameters are not public are usually unavailable via API. Our proposed model can be easily used to obtain soft labels.

Model Size Our method is to add temporal knowledge to GPT-2-based COMET. The number of parameters of ChatGPT is unknown, but for reference, the number of TaCOMET's parameters is less than 1/100th of GPT-3 (175B). It reduces the resource and time costs. Related to the soft output, smaller models have higher usability in applications.

4. Experiments of TaCOMET

In this section, we construct TaCOMET and verify its performance. The details of dataset construction, finetuning, and evaluation are described, respectively. Experiments are conducted primarily in Japanese but also in English.

4.1. TimeATOMIC Construction

First, we construct TimeATOMIC in Japanese. We use gpt-3.5-turbo[2] through the OpenAI API (hereinafter, ChatGPT). By inputting interval time in addition to *head* and *relation* into ChatGPT and outputting *tail*, a dataset of the form (*head*, *relation*, interval time, *tail*) is constructed.

We randomly sample 2,000 events from the *heads* and *tails* for xNeed and xEffect in ATOMIC-ja (Ide et al., 2023) as events to be fed into ChatGPT as the *heads*. As described in Section 3.1, xNeed and xEffect are used as the *relations*. An interval time is created by concatenating a pair of randomly selected strings from {1, 2, 3, 4, 5}×{second(s), 0 seconds, minute(s), 0 minutes, hour(s), day(s), month(s)}.[3] Three interval times are adopted per *head* / *relation* pair. In summary, there are two *relations* for each *head* and three interval times for each *head*/*relation* pair. In other words, a total of 12,000 triples of (*head*, *relation*, interval time) are fed into ChatGPT.

Tails are obtained by few-shot generation. A few manually created examples are used for shots, and the number of shots is set to three. The template for the instruction is, "Please describe an event that happens {interval_time} after {head}."[4] In some cases, ChatGPT refused to generate unrealistic *head*/interval time pairs or sensitive *heads*, and thus simple filtering is applied.

As a result, we obtained 11,249 instances, as shown in Table 1. Of these, 10% is used as the test set and the remainder as the training set. For the same *head*/*relation*, appropriate and various *tails* were obtained depending on the input time. The entire cost of constructing this Japanese dataset was approximately $6.

[2] May 12 version.

[3] The original units are in Japanese.

[4] This is for xEffect. For xNeed, "before" is used instead of "after". The original template is in Japanese.

Relation	Time	*Tail* (Japanese)	*Tail* (English)
xNeed	10 秒 (seconds)	X が靴を履く	X puts on shoes
	2 日 (days)	X が財布にお金を入れる	X puts money in his wallet
	1 ヶ月 (month)	X が予算を立てるために貯金計画を立てる	X creates a savings plan for budgeting
xEffect	30 秒 (seconds)	X が財布を取り出す	X takes out his wallet
	30 分 (minutes)	X が自宅に帰る	X goes home
	3 時間 (hours)	X が家で料理をする	X cooks at home

Table 1: Examples of TimeATOMIC. (*Head*: "X がスーパーへ買い物に行く" ("X goes shopping at the supermarket"))

We created a same-sized dataset in English by machine translation[5].

4.2. Finetuning

We finetune COMET on TimeATOMIC. This section describes the details of finetuning.

Format As described in Section 3.2, finetuning is performed to generate the *tail* of the *head*, *relation*, and time input for each instance. We insert time while keeping the existing COMET format as much as possible.

The COMET training format is "[head] [relation] [tail]", e.g., "X goes to the supermarket xNeed X puts on shoes". We insert interval time and the suffix expression (ago or later) corresponding to *relation*. The resulting format is "[head] [relation] [time] [ago | later], [tail]". An example is "X goes to the supermarket xNeed 10 seconds ago, X puts on shoes". In the same way as COMET, loss calculation is performed only for *tails*.

Models COMET-ja (Ide et al., 2023) and COMET$^{\text{DIS}}_{\text{TIL}}$ $+$ critic$_{\text{high}}$ (West et al., 2022) are used as the base COMET models for Japanese and English, respectively. These are the COMET models trained on the ATOMIC knowledge graphs, which contain approximately 200K and 2.5M triplets, respectively. Both the models use the GPT-2-xl architecture, which has approximately 1.5B parameters.

For model size comparison, a COMET-ja model based on GPT-2-small, which has 110M parameters, is also tested. This model was also trained on the 200K ATOMIC triples. For method comparison, we also make GPT-2 models trained directly on TimeATOMIC without using COMET. To distinguish these models from TaCOMET, we call them TaCOMET$_{\text{SCRATCH}}$.

4.3. Evaluation Metrics

We conduct automatic and manual evaluations of the models' generations for the test set. Generation is performed by inputting "[head][relation] [time] [ago | later]," and outputting the following *tail*.

Automatic Evaluation We perform two kinds of automatic evaluation using BERTScore (Zhang et al., 2020b) for text similarity calculation.

The first automatic metric is the similarity between the TaCOMET generation and the reference sentence in our dataset. This metric measures the model's performance of natural language generation, and thus the higher the number, the better. It is denoted by BS$_{\text{GEN}-\text{REF}}$ and calculated as follows:

$$\text{BS}_{\text{GEN}-\text{REF}} = \frac{1}{n}\sum_{i=1}^{n} BS(g_i, c_i),$$

where n is the number of instances in the test set, $BS(s_1, s_2)$ denotes an F1 of BERTScore between s_1 and s_2, and $g_i, c_i (1 \leq i \leq n)$ denote a generation by TaCOMET and its reference sentence.

The second automatic metric measures how different TaCOMET generations are when the time input changes. Referring to Paiwise-BLEU (Shen et al., 2019), we measure the average similarity among TaCOMET generations for the same *head*/*relation* but only with different times. This metric should be small because the model is required to generate differently depending on the time inputs. The metric is denoted by BS$_{\text{INNER}}$ and calculated as follows:

$$\text{BS}_{\text{INNER}} = \frac{1}{|H||R|}\sum_{(h,r)\in H\times R} \text{BS}^{\text{h,r}}_{\text{INNER}},$$

$$\text{BS}^{\text{h,r}}_{\text{INNER}} = \frac{1}{|G_{h,r}|(|G_{h,r}|-1)}\sum_{\substack{(g,g')\in G^2_{h,r} \\ g\neq g'}} BS(g, g'),$$

where H is the set of *heads*, $R = \{\text{xNeed}, \text{xEffect}\}$, and $G_{h,r}$ denotes the set containing the all generations for $h \in H$ and $r \in R$.

We use a Japanese model[6] and an English one[7] of RoBERTa (Liu et al., 2019) as the base pretrained models for BERTScore calculation. Note that it is impossible to compare the values of

[5] https://www.deepl.com/translator

[6] https://huggingface.co/nlp-waseda/roberta-base-japanese

[7] https://huggingface.co/roberta-base

Base Model	Model Type	BS$_{\text{GEN-REF}}$	BS$_{\text{INNER}}$ ↓	$\rho_{pearson}$	$\rho_{spearman}$	$\rho^{\text{log}}_{pearson}$	valid%
en-XL	TaCOMET	**0.918 ± 0.033**	**0.924 ± 0.036**	**0.052**	**0.389**	0.369	0.767
	TaCOMET$_{\text{SCRATCH}}$	0.917 ± 0.033	0.925 ± 0.036	0.019	0.387	**0.370**	**0.799**
	(COMET)	0.906 ± 0.020	0.972 ± 0.030	0.034	0.091	0.094	0.737
ja-XL	TaCOMET	**0.754 ± 0.119**	0.763 ± 0.102	**0.065**	**0.473**	**0.458**	**0.784**
	TaCOMET$_{\text{SCRATCH}}$	0.752 ± 0.118	**0.760 ± 0.103**	0.049	0.467	0.454	0.771
	(COMET)	0.569 ± 0.068	0.818 ± 0.154	0.026	0.154	0.153	0.722
ja-small	TaCOMET	**0.605 ± 0.071**	0.715 ± 0.129	**0.131**	0.450	0.438	**0.682**
	TaCOMET$_{\text{SCRATCH}}$	0.604 ± 0.072	**0.711 ± 0.132**	-0.008	**0.467**	**0.444**	0.640
	(COMET)	0.571 ± 0.063	0.873 ± 0.173	0.022	0.062	0.059	0.344

Table 2: Evaluation results of TaCOMET. Only BS$_{\text{INNER}}$ should be smaller. In the first column, "en" and "ja" denote English and Japanese, respectively.

Head	***Relation***	***Time***	**TaCOMET**	**COMET**
X が店まで走る (X runs to a store)	xNeed	30 分 (minutes)	X が運動着に着替える (X changes into jersey)	X が家を出る (X leaves home)
		4 時間 (hours)	X が運動不足になる (X feels under−exercised)	X が家を出る (X leaves home)
		2 日 (days)	X がジョギングシューズを買う (X buys jogging shoes)	X が家から出る (X leaves home)
X が店まで走る (X runs to a store)	xEffect	4 秒 (seconds)	X が息を切らす (X becomes out of breath)	X が財布を落とす (X drops the wallet)
		10 分 (minutes)	X が息を切らす (X becomes out of breath)	X が財布を忘れる (X forgets the wallet)
		4 日 (days)	X が筋肉痛になる (X becomes sore)	X が財布を落とす (X drops the wallet)
X が問題集を解く (X does the workbook)	xNeed	10 分 (minutes)	X が机の上に問題集を並べる (X lays the workbooks on the desk)	X が問題集を開く (X opens the workbook)
		3 時間 (hours)	X が勉強するための教材を整理する (X organizes study materials)	X が問題集を開く (X opens the workbook)
		4 日 (days)	X が問題集を購入する (X purchases the workbooks)	X が問題集を開く (X opens the workbook)
X が会社から パソコンを持ち帰る (X brings home a computer from the office)	xEffect	2 分 (minutes)	X が鞄を持ち上げる (X lifts the bag)	X が仕事を終える (X finishes thework)
		20 分 (minutes)	X が自宅のパソコンで仕事を始める (X starts working on the computer)	X が仕事を終える (X finishes the work)
		4 ヶ月 (months)	X がパソコンの故障に気付く (X notices a computer malfunction)	X が仕事を終える (X finishes the work)

Table 3: Comparison of generated *tails* between TaCOMET and COMET. Japanese examples are shown, and each translation is denoted in parentheses.

Japanese and English because the base models are different.

Human Evaluation We perform human evaluation by crowdsourcing. We asked crowdworkers to tackle the following two tasks regarding the generations of TaCOMET. Five workers per example are employed for both tasks.

The first task is to show *head*, *relation*, and *tail* (generation by TaCOMET) to the workers and ask them to answer the interval time. The mean of the five interval times obtained is used as the crowdsourced labels. We measure the correlation between the two series of interval time, one in the dataset actually entered into TaCOMET and the other in the crowdsourced labels. We calculate three human metrics: the Pearson correlation coefficient ($\rho_{pearson}$), the Spearman's rank correlation coefficient ($\rho_{spearman}$), and the Pearson correlation coefficient calculated for the series with both logarithmic transformations ($\rho^{\text{log}}_{pearson}$).

The second task is to show the same things and ask the workers to judge whether a generation is valid. The interval time is not shown, and thus only the adequacy of the inference is measured. The results obtained are aggregated by majority voting, which forms a crowdsourced label. These labels are used for calculating the ratio of valid generations out of the total inferences (valid%).

We use Yahoo! Crowdsourcing[8] and Amazon Mechanical Turk[9] as the platform for Japanese and English, respectively. Note that it is impossible to compare the scores between Japanese and English because the platforms and the nature of the crowdworkers are different between Japanese and English.

4.4. Results and Discussion

The evaluation results are shown in Table 2. For each base model, the following three settings are listed, including the two finetuned models described in Section 4.2 and a baseline.

- TaCOMET: COMET + TimeATOMIC
- TaCOMET$_{\text{SCRATCH}}$: GPT-2 + TimeATOMIC
- (COMET): Vanilla COMET without finetuning

COMET is tested in the same format as the others to verify if the ability to consider interval time is obtained by being finetuned on TimeATOMIC.

BERTScore Results BS$_{\text{GEN-REF}}$, which indicates the performance as a generation task, was improved by finetuning on TimeATOMIC. The training of TaCOMET and TaCOMET$_{\text{SCRATCH}}$ was able to adapt to TimeATOMIC.

Better results were also obtained by finetuning for BS$_{\text{INNER}}$, which shows the differences in generations at different times. By changing the input time for the same *head*/*relation*, appropriate generations can be obtained accordingly. Generated examples are shown in Table 3. We can see that COMET generates similarly for any time input, whereas TaCOMET controls generation in response to the time. These results show that finetuning on TimeATOMIC gave the models the ability to generate different *tails* according to the time.

Neither BS$_{\text{GEN-REF}}$ nor BS$_{\text{INNER}}$ showed significant differences between TaCOMET and TaCOMET$_{\text{SCRATCH}}$.

When we compare the small and XL sizes in the Japanese models, the relative score trends are the same. However, there are differences in the absolute values of the scores due to the simple difference in the number of model parameters.

Correlation Coefficient Results We discuss the correlation between the crowdsourced interval time labels and the actual interval times input into the model.

The Spearman's rank correlation coefficient was almost zero for COMET but was around 0.4 to 0.5 for the finetuned models, confirming a positive correlation. These models can not only change the generation with time but also generate the appropriate *tails* according to its time granularity. Especially for the XL models, TaCOMET generally obtained higher values than TaCOMET$_{\text{SCRATCH}}$, indicating the effectiveness of the proposed continual finetuning approach.

Regarding the Pearson correlation coefficient, little correlation was found in the intact series. For those after log transformations, a positive correlation was found for the models with finetuning. A logarithmic scale of human perception also emerged in the generations of the language models, such that the difference between one minute and five minutes is more perceptible than that in one month and five months.

For model size comparison, in the Spearman's rank correlation coefficient and the Pearson correlation coefficient after logarithmic transformations, the XL models obtained higher correlations than the small models in every model type. The ability to capture the granularity and scale of time also depends on the model size.

Acceptance Results We verify if the generations are valid when ignoring the time scale. For valid%, the XL finetuned models achieved over 75%, and the small models achieved over 60%. These models can keep the performance as the original COMET, although they have been finetuned on TimeATOMIC.

COMET could not achieve a high score due to format differences. COMET XL is relatively adaptable to the format change, but not so in the small-sized COMET.

Result Summary Finetuning on TimeATOMIC has improved the ability to change generation and to generate appropriately according to the input time. Also, the continual finetuning strategy via COMET showed effectiveness in rank correlation. We supposed that valid% should also be high because the continual finetuning via COMET provides both a sense of time and a broad range of commonsense knowledge. However, the results showed no significant difference in valid% between TaCOMET and TaCOMET$_{\text{SCRATCH}}$ for any of the base models. Since the evaluation here uses a split from a single dataset, a more open setting may use the knowledge from the first-stage training. In Section 5, we show a more open downstream experiment, where TaCOMET was found to be superior to TaCOMET$_{\text{SCRATCH}}$.

[8] https://crowdsourcing.yahoo.co.jp/

[9] https://www.mturk.com/

utterance	: あれ? もう一個足りない (Uh? Missing one more)
action	: グラスを持ってくる (To bring another glass)
viewpoint	: 1
position	: キッチン (kitchen)
pose	: 立っている (standing)
has	: グラス, お酒 (a glass, alcohol)
coffee table	: なし (None)
dining table	: お菓子, コップ, など (snacks, cups, etc.)
kitchen	: なし (None)

Figure 2: An example of the robot dataset.

5. Downstream Experiment

As an example of how the proposed model is applicable to downstream tasks, we test it on a dataset for robot decision-making. Using TaCOMET's generation probability for in-home robots' decision-making, we verify the effectiveness of adding temporal knowledge to COMET.

5.1. Dataset Description

We use the dataset created by Tanaka et al. (2024), which measures the decision-making performance of an in-home robot for ambiguous utterances. Since the utterances in this dataset are not direct instructions but rather a collection of ambiguous utterances such as monologues, commonsense reasoning is required. The dataset includes 400 examples and a list of 40 actions. Each instance consists of a user's utterance, an image from the robot's point of view, some descriptions of the image, and one correct action (Figure 2). The task is to choose the correct action from the list.

5.2. Method

We perform action classification using a score based on the generation probability of TaCOMET. This is achieved by TaCOMET's ability to generate soft outputs mentioned in Section 3.3. The score can be calculated because the model outputs logits, not only word symbols.

The template for TaCOMET is "[head] [relation] [time] [ago | later], [tail]". We use an utterance and one of the actions from the dataset as the *head* and the *tail* of TaCOMET, respectively. The log probability normalized by the action length of the *tail* part is calculated as the score of an action. This is repeated for all 40 actions in the list, and the action with the highest score is regarded as the model's prediction. That is, for an utterance **u**, a time expression **t** ("[time] [ago | later],"), the action list A, and each action $\mathbf{a} = w_1^{\mathbf{a}}...w_{|\mathbf{a}|}^{\mathbf{a}}$, we obtain the predicted action $\hat{\mathbf{a}}$ as follows:

$$\hat{\mathbf{a}} = \text{argmax}_{\mathbf{a}\in A} H(\mathbf{a}),$$

where $H(\mathbf{a}) = \frac{1}{|\mathbf{a}|}\sum_{t=1}^{|\mathbf{a}|} \log \text{P}(w_t^{\mathbf{a}}|\mathbf{u}, \text{“xEffect”}, \mathbf{t}, w_1^{\mathbf{a}}...w_{t-1}^{\mathbf{a}}).$

To adapt the input format to TaCOMET, an utterance and an action are substituted in the form "X says [utterance]"[10] and "X [action]",[10] respectively. An example of the actual input to TaCOMET is "X says he forgot to put ketchup on it xEffect 1 minute later, X brings ketchup".[10]

Given a natural interval time for the interaction between the in-home robot and the user, such as 1 second to 1 minute, the inference becomes more natural and performance improves. In contrast, for an unnatural input time, such as a day or a month, it is difficult to make a decision. For example, it is easy to infer that what to do one minute after forgetting ketchup is to bring ketchup. However, what to do a month after forgetting ketchup is not settled, nor is to bring ketchup correct.

5.3. Experimental Setup

We vary the input interval time to verify that performance increases when an appropriate time is input and decreases when an unrealistic time is input. The interval time to be assigned to [time] is one of the following: {1 second, 5 seconds, 1 minute, 1 hour, 1 day, 1 month}.

The experiment is conducted using the Japanese XL model. In the same way as Section 4, we test three models: TaCOMET, TaCOMET$_{\text{SCRATCH}}$, and COMET. As a baseline with no time input, we also test COMET with a template that removes **t** from the above format.

When the 40 actions in the list are sorted by score, the percentage of correct answers at top-1 and top-5 is calculated as Accuracy (Acc.) and Recall at 5 (R@5), respectively.

5.4. Results

Figure 3 shows the results. The finetuned models with the 1-second to 1-minute input outperformed those without time input in both Acc. and R@5. By giving the appropriate time, the models could make time-based predictions. When the input was more than one hour, the scores were equal to or less than those without time input. This unnatural

[10] The original texts are in Japanese.

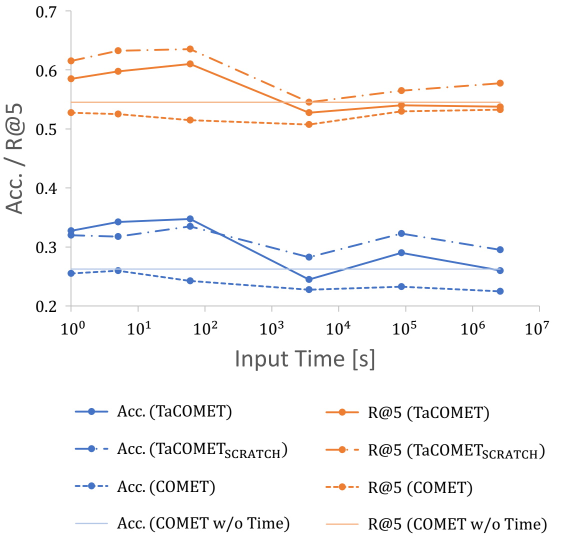


Figure 3: Robot dataset results. Acc. and R@5 denote accuracy and recall at five, respectively.

time input affected the action predictions because the finetuned models acquired the sense of time.

We compare TaCOMET with TaCOMET$_{\text{SCRATCH}}$ below. For Acc., TaCOMET was superior from 1 second to 1 minute, while TaCOMET was inferior for more than 1 hour, which is unnatural a time input. TaCOMET's results varied greatly depending on natural and unnatural time inputs, indicating that TaCOMET better captures time and responds to temporal input. For R@5, TaCOMET$_{\text{SCRATCH}}$ always outperformed TaCOMET. Although the performance was expected to drop for 1 hour or more, TaCOMET$_{\text{SCRATCH}}$ always remained better than the baseline without time. In contrast, TaCOMET was below the baseline in this range and was better in terms of precision.

COMET with time input underperformed the baseline without time because they could not cope with the change in the format.

As a reference, Tanaka et al. (2024) used an utterance and the corresponding images as input in a model based on RoBERTa and EfficientNet (Tan and Le, 2020). They performed supervised learning with cross-validation. They obtained Acc. and R@5 of 0.2723 and 0.5450, respectively. Our method was more accurate despite the zero-shot setting without image inputs.

6. Conclusion

We proposed a time-aware commonsense knowledge model, TaCOMET, in response to the fact that the existing commonsense knowledge graphs and models do not deal with granularity.

First, we built TimeATOMIC, a commonsense knowledge graph with time, using ChatGPT at a low cost. Based on TimeATOMIC, we then built TaCOMET by applying continual finetuning to the existing COMET model.

TimeATOMIC has improved the model's ability to change generation and to generate appropriately according to the input time. Also, our continual finetuning approach helps capture the granularity and scale of time. We observed some trends that the commonsense knowledge from the first finetuning step on ATOMIC is leveraged.

In addition, as an example of TaCOMET's applicability, we tested it on a robot dataset. We found that proper time input improves performance and has a useful time sense even for downstream tasks. This more open experiment also showed the proposed continual training was effective.

Future work includes building models that can consider context, extending *relations*, adopting other aspects of granularity, and multimodalization. TaCOMET could also be applied to more general tasks, including testing on a wider range of downstream tasks and knowledge transfer to general language models.

7. Acknowledgements

This work was conducted in collaboration with SB Intuitions Corp. and Waseda University.

8. Bibliographical References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Yasunobu Asakura, Masatsugu Hangyo, and Mamoru Komachi. 2016. Disaster analysis using user-generated weather report. In *Pro-*

ceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), pages 24–32, Osaka, Japan. The COLING 2016 Organizing Committee.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12857–12865.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The EventStatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44–54, Austin, Texas. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Tatsuya Ide, Eiki Murata, Daisuke Kawahara, Takato Yamazaki, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2023. Phalm: Building a knowledge graph from scratch by prompting humans and a language model.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea. 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Inderjeet Mani. 1998. A theory of granularity and its application to problems of polysemy and underspecification of meaning. In *International Conference on Principles of Knowledge Representation and Reasoning*.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Rutu Mulkar-Mehta, Jerry Hobbs, and Eduard Hovy. 2011. Granularity in natural language discourse. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the TimeBank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204, Berlin, Germany. 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. Dense-ATOMIC: Towards densely-connected ATOMIC with high knowledge coverage and massive multi-hop paths. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13292–13305, Toronto, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

Shohei Tanaka, Konosuke Yamasaki, Akishige Yuguchi, Seiya Kawano, Satoshi Nakamura, and Koichiro Yoshino. 2024. Do as i demand, not as i say: A dataset for developing a reflective life-support robot. *IEEE Access*, 12:11774–11784.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022. CN-AutoMIC: Distilling Chinese commonsense knowledge from pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9253–9265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020a. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2019. Aser: A large-scale eventuality knowledge graph.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 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)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020a. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1361–1371, Online. Association for Computational Linguistics.

Wangchunshu Zhou, Ronan Le Bras, and Yejin Choi. 2023. Commonsense knowledge transfer for pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5946–5960, Toronto, Canada. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020b. Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9733–9740.

9. Language Resource References

Tatsuya Ide, Eiki Murata, Daisuke Kawahara, Takato Yamazaki, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2023. Phalm: Building a knowledge graph from scratch by prompting humans and a language model.

Shohei Tanaka, Konosuke Yamasaki, Akishige Yuguchi, Seiya Kawano, Satoshi Nakamura, and Koichiro Yoshino. 2024. Do as i demand, not as i say: A dataset for developing a reflective life-support robot. *IEEE Access*, 12:11774–11784.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.