Temporal Natural Language Inference: Evidence-based Evaluation of Temporal Text Validity

Taishi Hosokawa[1], Adam Jatowt[2], and Kazunari Sugiyama[3]

[1] Kyoto University, Japan
taishihosokawa@yahoo.co.jp,
[2] University of Innsbruck, Austria
adam.jatowt@uibk.ac.at
[3] Osaka Seikei University, Japan
zakugus@gmail.com

Abstract. It is important to learn whether text information remains valid or not for various applications including story comprehension, information retrieval, and user state tracking on microblogs and via chatbot conversations. It is also beneficial to deeply understand the story. However, this kind of inference is still difficult for computers as it requires temporal commonsense. We propose a novel task, *Temporal Natural Language Inference*, inspired by traditional natural language reasoning to determine the temporal validity of text content. The task requires inference and judgment whether an action expressed in a sentence is still ongoing or rather completed, hence, whether the sentence still remains valid, given its supplementary content. We first construct our own dataset for this task and train several machine learning models. Then we propose an effective method for learning information from an external knowledge base that gives hints on temporal commonsense knowledge. Using prepared dataset, we introduce a new machine learning model that incorporates the information from the knowledge base and demonstrate that our model outperforms state-of-the-art approaches in the proposed task.

1 Introduction

It is rather easy for humans to reason on the validity of sentences. Given a user's post: "I am taking a walk", and a subsequent post from the same user: "Ordering a cup of coffee to take away", we can guess that the person is very likely still taking a walk, and has just only stopped for a coffee during her walk. That is, the action stated in the former message is still ongoing, thus, the first sentence remains valid. On the other hand, if the subsequent post would be "I am preparing a dinner", it would be highly possible that the first message (the one about taking a walk) is no longer valid in view of this additional evidence. This kind of inference is usually smoothly done by the commonsense of humans.

Thanks to the emergence of pre-training models, computers have shown significant performance in the field of natural language understanding [59]. However, it is still a challenging task for computers to perform effective reasoning that requires commonsense knowledge [54]. As the amount of available text information is exploding nowadays, it is getting more and more important for machines to achieve much better understanding of natural language.

In this work, we propose a novel task called *Temporal Natural Language Inference* (TNLI), in which we evaluate the validity of text content using an additional related content used as evidence. Similarly to Natural Language Inference (NLI) task [55] an input is in the form of a sentence pair (we explain more on the differences of these two tasks in §2.5). We address the TNLI problem as a classification task using the following two input sentences: (1) hypothesis sentence and (2) premise sentence. The first one, the hypothesis sentence, is the one whose validity is to be judged. The second one, the premise sentence, is following the hypothesis sentence and is supposed to provide new information useful for classifying the hypothesis. The classification labels for the hypothesis sentence are as follows: SUPPORTED, INVALIDATED, and UNKNOWN. SUPPORTED means that the hypothesis is still valid after seeing the premise, INVALIDATED means that the hypothesis ceased to be valid, and otherwise it is UNKNOWN.

Considering our earlier example, if we regard “I am taking a walk” as a hypothesis and “I am preparing a dinner” as a premise, the hypothesis becomes INVALIDATED since the user has clearly concluded her earlier action. If we consider “coffee for take away” as a premise instead, the hypothesis would be SUPPORTED.[4]

The potential applications of our proposed task are as follows:

Support for story understanding & event extraction: Effective methods trained for the proposed task can lead to better understanding of stories described in text and potentially also more effective event extraction [72]. Reading comprehension of stories would be improved if one incorporates a component that can reason about action completion given the evidence provided by the following sentences. We note that this kind of knowledge is often implicit in text.

Classification & recommendation of microblog posts: Microblog posts can be valid for different lengths of time. In the era of information overload, users need to select valid messages among a large number of posts, as valid ones are typically the most relevant and important. This kind of information overload would be significantly alleviated when they could use an option to filter out invalid posts from their timelines, or the posts could be ranked by several factors including their estimated temporal validity.

User tracking & analysis: User tracking and analysis in social networks services (SNS) and chats [2, 33, 1] can be enhanced based on temporal processing of user’s posts so that the user’s current situation and action can be flagged. This could be useful for example for selecting suitable advertisements or in emergency situations like during the time of disasters to know the current state of users.

Chatbots: Chatbots and AI assistants are becoming recently increasingly popular [43]. However, their use is still quite limited as they cannot understand well users’ context such as their actions, goals and plans, and so on. To achieve much better communication with humans, it is necessary to develop user state-aware approach.

In addition to the proposal of a novel task, our second contribution is the construction of dedicated dataset for the proposed task. As we especially focus on sentences that

[4] Note that it is not always easy to determine the correct answer as the context or necessary details might be missing, and in such cases humans seem to rely on probabilistic reasoning besides the commonsense base.

describe actions and relatively dynamic states that can be relevant to the aforementioned applications, our dataset contains pairs of sentences describing concrete actions.

Finally, we also develop a new machine learning model incorporating information from a knowledge graph as we believe that successful model requires external knowledge about the world. Our proposed model combines the following two encoders: the first one incorporates commonsense knowledge via pre-training and the other one is purely based on text to jointly capture and reason with the commonsense knowledge. To develop these encoders, we construct a new model for learning the embeddings of concepts in knowledge bases. In particular, we employ the ATOMIC-2020 dataset [25] as an external knowledge base.

Our main contributions can be summarized as follows:

1. We propose a novel task called, Temporal Natural Language Inference (TNLI), requiring to identify the validity of a text content given an evidence in the form of another content. We formulate it as a text classification problem.
2. We construct a dedicated dataset[5] for our proposed task that contains over 10k sentence pairs, and analyze its relation to Natural Language Inference (NLI) task and NLI's corresponding datasets.
3. We design and develop an effective new machine learning model for the proposed task which utilizes information from commonsense knowledge bases.
4. Finally, we compare our model with some state-of-the-arts in natural language inference task and discuss the results. We also test if pre-training using standard NLI datasets is effective for TNLI.

2 Related Work

2.1 Temporal Information Retrieval & Processing

Temporal Information Retrieval is a subset of Information Retrieval domain that focuses on retrieving information considering their temporal characteristics. Many tasks and approaches have been proposed so far [1, 28, 42, 9, 27, 45], including the understanding of story [23], temporal relation extraction [63, 18], question answering [24, 26], and so on. White and Awadallah [69] estimated the duration of tasks assigned by users in calendars. Takemura and Tajima [57] classified microblog posts to different lifetimes based on features specific to Twitter such as number of followers or presence of URLs. Almquist and Jatowt [3] examined the validity of sentences considering the time elapsed since their creation (more in §2.5).

2.2 Commonsense Reasoning

Implicit information that humans commonly know is addressed in, what is called, Commonsense Reasoning domain [60]. Winograd Schema Challenge [32] was one of the earliest challenges for machines in this regard, and many other challenges and approaches have also been proposed [49, 38, 21, 58, 36, 44, 34]. Temporal Commonsense

[5] The dataset and code are available at: https://tinyurl.com/T-NLI

is one of them, in which temporal challenges are addressed [78]. Zhou et al. [77] focused on comparing actions such as “going on a vacation” with others like “going for a walk” to assess which take longer, and constructed a dataset for question-answering including this kind of estimation. We further compare our task with other related ones in §2.5.

2.3 Natural Language Inference

Recently, Natural Language Understanding (NLU) by computers has attracted a lot of researchers’ attention. Natural Language Inference (NLI) or Recognizing Textual Entailment is one of NLU domains, in which computers deal with input in the form of two sentences [55], similar to our proposed task. NLI problems require to determine that a premise sentence entails, contradicts, or is neutral to a hypothesis sentence (or in some settings, entails vs. not entails). In the early stages of NLI work, Dagan et al. [15] constructed a relatively small dataset. The first largely annotated dataset was *Stanford Natural Language Inference* (SNLI) dataset [6], which was annotated through crowdsourcing. After that, many NLI datasets [16, 22], including *Multi-genre Natural Language Inference* (MNLI) [70] and Scitail [30], have been constructed. Notably, Vashishtha *et al.* [62] converted existing datasets for temporal reasoning into NLI format, pointing out that no NLI dataset focuses on temporal reasoning. Their task focuses on explicit temporal description while our task tackles implicit information. The emergence of these large scale datasets made it possible to train more complex models [55, 7]. Remarkably, state-of-the-art large-scale pre-trained models such as BERT [17] and RoBERTa [37] demonstrated significant performance on NLI datasets, and are also used to train multi-task models [14].

2.4 Incorporation of Knowledge Bases

Generally, NLU works make use of Knowledge Graphs (KG) or Knowledge Bases (KB) to improve model performance [11, 40, 73]. Especially, Commonsense Reasoning works commonly incorporate knowledge from large KBs such as ConceptNet [35, 53] and WikiData [64] in their architectures [48, 76, 75]. However, only a few works in NLI attempt to incorporate KGs into models [8, 29]. Wang et al. [66], for example, improve performance on Scitail using knowledge in ConceptNet.

2.5 Comparison with Related Tasks

Similar to NLI, our work addresses a text classification problem, in which two sentences form an input. However, we focus on neither entailment nor contradiction but the validity of sentences (see Tables 3 and 4 for comparison).

The NLI dataset constructed by Vashishtha *et al.* [62] includes temporal phenomena. However, their task addresses explicit descriptions of temporal relations such as duration and order, while we focus on implicit temporal information that is latent in sentences. Temporal Commonsense task [77] includes implicit temporal information, too. The problem that the task deals with is reasoning about event duration, ordering, and frequency in a separate manner. However, our approach requires a more comprehensive understanding of temporal phenomena through a contrastive type inference. Also, their task is posed as a question-answering problem while ours is formalized as an NLI type problem. Almquist and Jatowt [3] also worked on the validity of sentences. Unlike their work, we use premises as the additional source instead of the information on the elapsed time from sentence creation as in [3], since, in practice, in many situations, additional text is available (e.g., sequences of tweets posted by the same user, or following sentences in a story or novel). Table 1 compares our task with the most related ones.

Table 1: Comparison our work with related tasks.

Task	Task Type	Temporal	Input	Output
McTaco [77]	Question Answering	✓	source, question, answer candidates	correct answers
NLI [20, 12]	Classification		sentence pair	3 classes (entailment, contradiction, neutral)
Validity Period Estimation [3]	Classification	✓	sentence	5 classes (hours, days, weeks, months, years)
TNLI (Proposed Task)	Classification	✓	sentence pair	3 classes (SUPPORTED, INVALIDATED, UNKNOWN)

3 Task Definition

We first provide the definition of our task, in which a pair of sentences $p = (s_1, s_2)$ is given, where s_1 and s_2 are a hypothesis and a premise sentence, respectively.[6] The task is to assign one of the following three classes to s_1 based on the inference using the content of s_2:

$$c \in \{\text{SUPPORTED}, \text{INVALIDATED}, \text{UNKNOWN}\} \tag{1}$$

Here, the SUPPORTED class means that s_1 is still valid given the information in s_2. The INVALIDATED class, on the other hand, means that s_1 ceased to be valid in view of s_2. The third one, UNKNOWN class, indicates that the situation evidence is not conclusive or clear, and we cannot verify the validity of the hypothesis.

4 Proposed Method

We discuss next our proposed approach for content validity estimation task. We hypothesize that the task cannot be successfully solved without the incorporation of external knowledge given the inherent need of temporal commonsense reasoning. Therefore, we first attempt to find the useful knowledge base to provide knowledge of temporal

[6] Note that s_1 and s_2 may have temporal order: $t_{s_1} \leq t_{s_2}$, where $t_{s_{id}}$ $(id = 1, 2)$ is the creation time (or a reading order) of a sentence s_{id}. This may be for example in the case of receiving microblog posts issued by the user (or when reading next sentences of a story or a novel).

properties. We then propose a new model by combining an encoder that incorporates information from this knowledge base and a text encoder that uses only text data. The output of the encoder using the knowledge base and the text encoder are combined and used as input to the softmax classifier. Figure 1 shows the model outline.

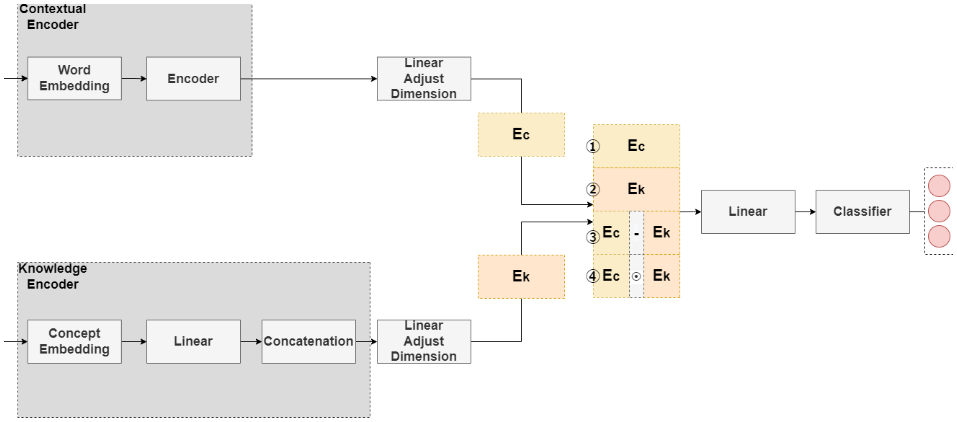

Fig. 1: Outline of our approach.

4.1 Encoding Knowledge

One of the key components of the proposed model is the knowledge encoder.

Knowledge Base We have explored different knowledge bases (KBs) for our objective including FrameNet [19], WikiHow [31], Howto100m [39], and VerbNet [52]. We concluded that ATOMIC-2020 (*An ATlas Of MachIne Commonsense*) [25] is the most suitable KB to achieve our goal thanks to its temporal commonsense knowledge and relatively large scale (1.33M commonsense knowledge tuples and 23 commonsense relations).

ATOMIC [51] is the predecessor KB of ATOMIC-2020 designed for commonsense reasoning that contains nine different if-then relations such as Cause, Effect, Intention, Reaction, and so on. Most of the entities in this KG are in the form of sentences or phrases. COMET [5] is a language model trained with ATOMIC and ConceptNet in order to generate entities that were not in the ATOMIC dataset. Then, ATOMIC-2020 adds new relations in comparison to ConceptNet and ATOMIC. The new relations include “IsAfter”, “IsBefore”, “HasSubevent”, and so on, which represent the relations between events. For example, “PersonX pays PersonY a compliment” and “PersonX will want to chat with PersonY” are sentences belonging to the if-then relation in Atomic-2020, while “PersonX bakes bread” and “PersonX needed to buy ingredients” is an example of a pair of sentences connected by the “IsAfter” relation.

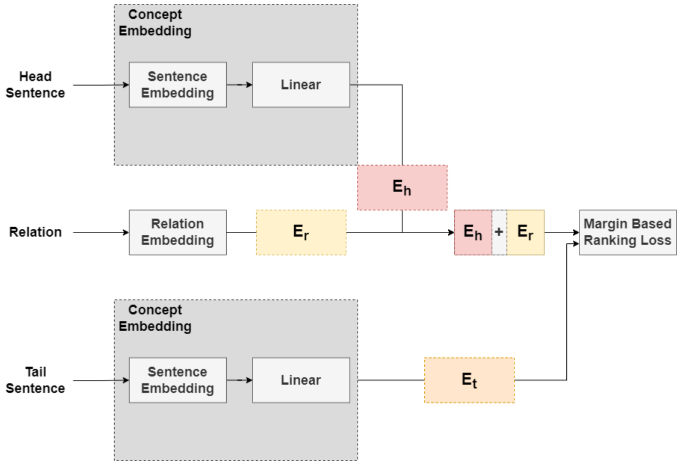

Fig. 2: TransE model for sentences.

TransE We now briefly explain TransE [4] which we adapt for the purpose of KB relation embedding. TransE is a model for learning embeddings of KBs represented in the triple form of entities and relations <head entity, relation, tail entity>. Relations are considered to be translations in the embedding space. TransE learns embedding using the following loss function, which is an operation between entities and relations as in skip-gram [41], where head entity + relation = tail entity:

$$\mathcal{L} = \sum_{(\mathbf{h},\mathbf{l},\mathbf{t})\in S} \sum_{(\mathbf{h}',\mathbf{l},\mathbf{t}')\in S'_{(\mathbf{h},\mathbf{l},\mathbf{t})}} [\gamma + d(\mathbf{h}+\mathbf{l},\mathbf{t}) - d(\mathbf{h}'+\mathbf{l},\mathbf{t}')]_{+}, \tag{2}$$

where $[x]_+$ denotes the positive part of x, γ is a margin parameter, d is the distance function, while $\mathbf{h}$, $\mathbf{l}$, and $\mathbf{t}$ are the embeddings of head entity, relation label, and tail entity, respectively. In addition, S is a set of positive examples, while S' is the set of negative examples.

TransE for Sentences Since the entities in the ATOMIC-2020 dataset are mostly in the form of short sentences, it is difficult to train with original TransE as the number of potential entities is very large, and inference for sentences that are not in the training data is not possible. To solve this problem, we adapt the TransE model for sentences. First, we compute the sentence vector corresponding to each entity in the KG using Sentence-BERT (SBERT) [50]. We then train the weights W for the sentence vectors and the relation embedding E_r using Margin Based Ranking Loss as in TransE. Here, the weights of SBERT are fixed and not trained, so that if the data is not in the training set, the embeddings of similar sentences will remain similar. Figure 2 shows the model structure. Since information related to time is crucial in our work, we only use “IsAfter” and “IsBefore” ATOMIC-2020 relations.

Other Translating Embedding Models We also explore other variants of translating embedding models: TransH [67] and ComplEx [61].

For TransH model, we adopt the same model as TransE for sentences except for an additional module for projection. To project into the hyperplane of each relation, we use relation-specific projection matrix as original TransH does.

For ComplEx model, we add a linear layer after sentence embedding so that the model has two different parallel linear layers to transform sentence embeddings, where one represents real part, and the other is for imaginary part.

Encoder with Knowledge We create an encoder for the downstream task using the ATOMIC-2020 pre-trained embeddings of the TransE model. In this encoder, the output is the concatenation of the embeddings of the hypothesis and of premise sentence.

4.2 Combined Model

The entire model for the proposed downstream task consists of text encoder, knowledge encoder, and a classification layer on top of them. Since the dimensions of the pre-trained embeddings and the output of the text encoder are not the same, each output is linearly transformed to make the dimensions equal. The two vectors obtained in this way are compared and combined, and then linearly transformed. We use the concatenation, difference, and element-wise product for combination:

$$\mathbf{H} = Linear(\mathbf{H_t}; \mathbf{H_k}; \mathbf{H_t} - \mathbf{H_k}; \mathbf{H_t} \odot \mathbf{H_k}), \tag{3}$$

where $\mathbf{H_t}$ is the output of text encoder, $\mathbf{H_k}$ is the output of knowledge encoder, and $\odot$ denotes element-wise multiplication. The obtained output is linearly transformed, and then fed into a softmax classifier to decide the validity class.

5 Dataset

5.1 Dataset Construction

To create our dataset, we need hypotheses, premises, and ground truth labels. As mentioned before, we decided to focus on sentences similar to the typical setup of NLI task. As the hypothesis sentences we used randomly selected 5,000 premise sentences from SNLI dataset[7]. These sentences were originally taken from the captions of the Flickr30k corpus [74]. We conducted clustering over all the collected sentences and sampled equal number of sentences from each cluster to maintain high variation of sentence topics. For this, we first employed BERT [17] to vectorize each word, and then vectorized the sentences based on computing the arithmetic mean of word vectors. For clustering, we employed k-means to group sentence vectors into 100 clusters. We then extracted up to 50 sentences from each cluster with uniform probability and used them as the source of the hypotheses. Premise sentences and labels were collected using crowdsourcing with

[7] SNLI dataset is licensed under CC-BY-SA 4.0.

Table 2: Average sentence length in our dataset.

	Average	Variance
hypothesis	11.4	19.4
premise	8.9	10.8
invalidated	8.4	8.6
supported	9.3	10.7
unknown	8.9	12.7

the Amazon Mechanical Turk[8]. For each hypothesis, we asked two crowdworkers to create a sentence corresponding to each label. To avoid sentences that are simply copied or modified with minimum effort, we accepted sentences only when 40% or more words were not overlapped with the corresponding hypothesis. Otherwise, crowdworkers could, for example, simply change one word in order to claim the added trivial sentence is of a given class (e.g., SUPPORTED class). Since the dataset should involve non-explicit temporal information and we wanted to make sure that the workers carefully consider it, we also asked for providing a description of the estimated time during which the hypotheses sentences could have been realistically valid, although this information was not included in the final dataset. In total, about 400 workers participated in the dataset creation.

Since we found out some spamming and dishonest activity, we later manually verified the validity of all the obtained data, corrected the grammar and words, as well as we manually filtered poor-quality, noisy, offensive, or too-personal sentences. For example, removed sentences included instances in which a single word was substituted with a different one that has the same or similar meaning or different case such as replacing "mother" with "MOM." We removed in total 19,341 pairs of sentences.

5.2 Dataset Statistics

The final dataset includes 10,659 sentence pairs. Since the previous research has pointed out that the number of words in sentences in some NLI datasets varies significantly depending on their labels [30], we examined the average number of words in our dataset. Table 2 shows the statistics indicating that the average number of words in our dataset does not change significantly for different labels. The variance tends to be higher, however, for the UNKNOWN class.

Note that the number of sentence pairs belonging to each class is the same (3,553). Table 3 shows some examples of the generated data, while, for contrast, we also show example sentences of NLI task in Table 4.

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

Table 3: Example sentences of TNLI task in our dataset.

Hypothesis	Label	Premise
A woman in blue rain boots is eating a sandwich outside.	INVALIDATED	She takes off her boots in her house.
A small Asian street band plays in a city park.	SUPPORTED	Their performance pulls a large crowd as they used new tunes and songs today.
A man jumping a rail on his skateboard.	UNKNOWN	His favorite food is pizza.

Table 4: Example sentences of NLI task from SNLI dataset (borrowed from SNLI website: https://nlp.stanford.edu/projects/snli/).

Hypothesis	Label	Premise
A man is driving down a lonely road.	Contradiction	A black race car starts up in front of a crowd of people.
Some men are playing a sport.	Entailment	A soccer game with multiple males playing.
A happy woman in a fairy costume holds umbrella.	Neutral	A smiling costumed woman is holding an umbrella.

6 Experiments

6.1 Experimental Settings

In our experiments, we perform 5-fold non-nested cross-validation for all the compared models. The batch size is 16, and the learning rate is determined by the performance on the validation fold chosen from the training folds among 0.005, 0.0005, 0.00005. For all the models, the optimal value was 0.00005. We evaluate our approach with accuracy, the percentage of correct answers, which is the most relevant metric and widely used for NLI task. We compare our proposed approach with the following four models: BERT (bert-base-uncased) [17], Siamese Network [10], SBERT [50] Embeddings with Feedforward Network, and Self-Explaining Model [56]. Except for ones including the self-explaining model, we train the models with cross-entropy loss.

The architecture of the Siamese Network is similar to that of Bowman *et al.* [6] using the 8B version of GloVe [47] for word embedding and multiple layers of tanh. For SBERT with Feedforward Network, each layer has 500 dimensions and ReLU activation and dropout. The number of hidden layers on the top of SBERT equals to 3, and the output layer is a softmax classifier. The dropout rates are 75%.

The Self-Explaining [56] is a model with an attention-like Self-Explaining layer on top of a text encoder (RoBERTa-base [37]), and it achieves state-of-the-art results on SNLI. The Self-Explaining layer consists of three layers: Span Infor Collecting (SIC) layer, Interpretation layer, and an output layer.

For testing the proposed architecture, we experimented with the two types of contextual encoders: Siamese Network and Self-Explaining model. The dimensionality of the entity embedding was 256, and the combined embedding was linearly transformed to 128 to match the dimensionality of each encoder. We trained the entity embeddings of Bordes *et al.* [4] with a learning rate of 0.001.

We implemented the models using PyTorch [46] with HuggingFace's transformers [71] and we conducted our experiments on a machine equipped with GPU.

6.2 Experiments with NLI Pre-Training

Generally, pre-training with NLI datasets improves accuracy in many downstream tasks [13]. As there is certain degree of relevance between our proposed task and NLI, we first experimented with pre-training selected models using the NLI datasets and then fine-tuning them on the proposed task. The datasets we used are the training sets of SNLI 1.0 and MNLI 0.9, which contain 550,152 and 392,702 examples, respectively. We mapped SUPPORTED to the entailment class of NLI, INVALIDATED to contradiction, and UNKNOWN to neutral.

Table 5 shows the results obtained by NLI pre-training, indicating that the NLI data has some relevance to our proposed task and can improve accuracy as it improves the results for Siamese network. However, when we use the Self-Explaining model, which is an already pre-trained model, we did not observe any improvement. This indicates that NLI datasets include effective information, yet it can likely be learned from general corpus, especially when using RoBERTa [37].

Table 5: NLI pre-training results in Siamese Network and Self-Explaining Model.

Model	**Accuracy**
Siamese	0.715
+SNLI	0.756
+MNLI	0.757
Self-Explaining	0.873
+SNLI	0.867
+MNLI	0.535

6.3 Incorporating Common-sense Knowledge

Table 6 shows the main experimental results. Among the compared models, the Self-Explaining model achieves the best accuracy. BERT, on the other hand, gives the worst results. We also observe that training of BERT is unstable as also pointed out in [17].

The incorporation of commonsense knowledge improves the accuracy in both the Siamese and Self-Explaining cases. While it can significantly improve the performance for the case of Siamese Network, it does not help much for Self-Explaining model. Self-explaining uses RoBERTa model which has been trained on larger data and has been carefully optimized, while other models use standard BERT. This might affect the results. In general, we believe that adding commonsense data through the architecture that we have proposed is a promising direction for TNLI task that calls for exploration of more sophisticated commonsense reasoning and more extensive datasets. This conclusion is also supported by the analysis of the confusion matrices shown in Figure 3. Incorporating TransE to Siamese net helps to more correctly determine SUPPORTED and UNKNOWN classes (improvement by 28% and 6.5%, respectively), while only slightly confusing the INVALIDATED class (decrease of 1.4%).

Table 6: Results on TNLI task.

Model	Accuracy
Siamese	0.715
SBERT + FFN	0.806
BERT	0.441
Self-Explaining	0.873
Siamese+TransE	0.784
Self-Explaining+TransE	0.878

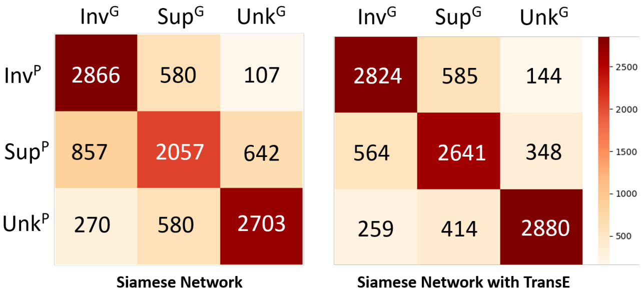

Fig. 3: Confusion matrices for TNLI prediction task of Siamese network (left) and Siamese network with TransE (right). The horizontal axis corresponds to the prediction (x^P) and vertical one to gold labels (x^G). The left (upper) blocks are INVALIDATED, middle ones are SUPPORTED, and right (bottom) ones are UNKNOWN.

6.4 Testing Different Knowledge Embedding Approaches

Finally, we explored different approaches of translating embedding models in addition to TransE. Table 7 shows the results of TransE variants combined with the Self-Explaining model. As it can be seen, the loss for pre-training does not go down in TransH [67] and ComplEx [61]. As the loss remains high, the accuracy with the proposed downstream task is also lower, indicating that the proposed architecture requires simpler way to construct the knowledge-based embeddings for TNLI because TransH and ComplEx are more complex models than TransE. Another possibility is that the other models were not supplied with sufficient knowledge to properly benefit from their more complex architectures.

Table 7: Results of TransE variants used with Self-Explaining model on TNLI task.

Model	Pre-Train Loss	Accuracy
TransE	0.19	0.878
TransH	0.48	0.868
ComplEx	1.24	0.856

7 Conclusion & Future Work

Computational processing of non-explicit temporal information in natural language still poses many challenges. In this work, we proposed a novel task for reasoning on the validity of sentences based on additional evidence and we trained a new model with an embedded knowledge base for this task. The motivation behind our idea is that humans can judge the temporal validity of a sentence using their commonsense, and our goal is to enable such reasoning ability for machines.

To achieve this goal, we first formally defined the task, constructed a dedicated dataset, and trained several baseline models. In addition, we proposed a new method of knowledge base embedding for sentences and a machine learning model that incorporates it. We believe that this work can contribute to our understanding of how to rely on knowledge bases that contain sentences as entities and of how to further improve the accuracy of TNLI task. We have also experimented with popular NLI datasets to answer a question on whether these can be useful for the proposed task.

Extending the dataset is one of our future goals. Our current work focused on sentences with relatively dynamic descriptions based on envisioned applications in microblogging. However, for more applications and training more robust models, it is necessary to construct datasets that also contain other forms of descriptions. More data, regardless of type, would be also necessary for larger-scale and less-biased training.

One way to achieve this would be to consider conversion methods from other datasets, as some NLI datasets such as Scitail and QNLI [68, 16, 65] have already employed them. Multi-modal datasets that include videos as well as their captions could be candidates for this. Another future direction is to extend the proposed task itself. More specifically, the timestamp of the premise sentences can be used as an additional signal to identify the validity of hypotheses sentences [3] in addition to judgments based on the content of premise sentences. This would lead to a more general task and a higher number of potential applications. It would be possible to address the cases where not only additional content is available as evidence for reasoning on the hypothesis's validity but also the time gap that elapsed from its statement is utilized (e.g., when using both the content and timestamps of user messages/posts or utterances). Therefore, the re-formulation of our task with added time information and the construction of a corresponding dataset are also in our plans. Finally, a further future work may focus on automatically generating premise sentences that would move their hypotheses into a required validity class to obtain the desired indication of action's completion or continuation.

References

1. S. Abe, M. Shirakawa, T. Nakamura, T. Hara, K. Ikeda, and K. Hoashi. Predicting the Occurrence of Life Events from User's Tweet History. In *Proceedings of the 12th IEEE International Conference on Semantic Computing (ICSC '18)*, pages 219–226, 2018.
2. F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing User Modeling on Twitter for Personalized News Recommendations. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP '11)*, pages 1–12, 2011.
3. A. Almquist and A. Jatowt. Towards Content Expiry Date Determination: Predicting Validity Periods of Sentences. In *Proceedings of the 41st European Conference on IR Research (ECIR '19)*, pages 86–101, 2019.
4. A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS '13)*, pages 2787–2795, 2013.
5. A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi. 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, July 2019. Association for Computational Linguistics.
6. S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.
7. Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada, July 2017. Association for Computational Linguistics.
8. Y. Chen, S. Huang, F. Wang, J. Cao, W. Sun, and X. Wan. Neural maximum subgraph parsing for cross-domain semantic dependency analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 562–572, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.
9. F. Cheng and Y. Miyao. Predicting event time by classifying sub-level temporal relations induced from a unified representation of time anchors. *arXiv preprint arXiv:2008.06452*, 2020.
10. D. Chicco. Siamese Neural Networks: An Overview. *Artificial Neural Networks - Third Edition*, pages 73–94, 2021.
11. P. Clark, B. Dalvi, and N. Tandon. What Happened? Leveraging VerbNet to Predict the Effects of Actions in Procedural Text. *arXiv preprint arXiv:1804.05435*, 2018.
12. C. Condoravdi, D. Crouch, V. de Paiva, R. Stolle, and D. G. Bobrow. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45, 2003.
13. A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.
14. M. Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.
15. I. Dagan, O. Glickman, and B. Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges Workshop (MLCW '05)*, pages 177–190, 2005.
16. D. Demszky, K. Guu, and P. Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

17. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, June 2019. Association for Computational Linguistics.
18. D. Dligach, T. Miller, C. Lin, S. Bethard, and G. Savova. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.
19. C. J. Fillmore and C. Baker. A Frames Approach to Semantic Analysis. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, 2010.
20. Y. Fyodorov, Y. Winter, and N. Francez. A Natural Logic Inference System. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2)*, 2000.
21. Q. Gao, S. Yang, J. Chai, and L. Vanderwende. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–945, Melbourne, Australia, July 2018. Association for Computational Linguistics.
22. M. Glockner, V. Shwartz, and Y. Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July 2018. Association for Computational Linguistics.
23. R. Han, M. Liang, B. Alhafni, and N. Peng. Contextualized Word Embeddings Enhanced Event Temporal Relation Extraction for Story Understanding. *arXiv preprint arXiv:1904.11942*, 2019.
24. S. Harabagiu and C. A. Bejan. Question Answering based on Temporal Inference. In *Proceedings of the AAAI-2005 workshop on inference for textual question answering*, pages 27–34, 2005.
25. J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 6384–6392, 2021.
26. A. Jatowt. Temporal question answering in news article collections. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, page 895. ACM, 2022.
27. A. Jatowt, É. Antoine, Y. Kawai, and T. Akiyama. Mapping Temporal Horizons: Analysis of Collective Future and Past related Attention in Twitter. In *Proceedings of the 24th international conference on World Wide Web (WWW '15)*, pages 484–494, 2015.
28. K. Kanazawa, A. Jatowt, and K. Tanaka. Improving retrieval of future-related information in text collections. In *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence (WI '11)*, pages 278–283, 2011.
29. P. Kapanipathi, V. Thost, S. S. Patel, S. Whitehead, I. Abdelaziz, A. Balakrishnan, M. Chang, K. Fadnis, C. Gunasekara, B. Makni, et al. Infusing Knowledge into the Textual Entailment Task Using Graph Convolutional Networks. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8074–8081, 2020.
30. T. Khot, A. Sabharwal, and P. Clark. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.
31. M. Koupaee and W. Y. Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

32. H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning (KR '12)*, pages 552–561, 2012.
33. P. Li, H. Lu, N. Kanhabua, S. Zhao, and G. Pan. Location Inference for Non-Geotagged Tweets in User Timelines. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 31(6):1150–1165, 2018.
34. B. Y. Lin, X. Chen, J. Chen, and X. Ren. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
35. H. Liu and P. Singh. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4):211–226, 2004.
36. Q. Liu, H. Jiang, Z.-H. Ling, X. Zhu, S. Wei, and Y. Hu. Combing Context and Commonsense Knowledge Through Neural Networks for Solving Winograd Schema Problems. In *Proceedings of the AAAI 2017 Spring Symposium on Computational Context: Why It's Important, What It Means, and Can It Be Computed?*, pages 315–321, 2017.
37. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.
38. Z. Luo, Y. Sha, K. Q. Zhu, S.-w. Hwang, and Z. Wang. Commonsense Causal Reasoning between Short Texts. In *Proceedings of the 15th International Conference on the Principles of Knowledge Representation and Reasoning (KR '16)*, pages 421–431, 2016.
39. A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '19)*, pages 2630–2640, 2019.
40. T. Mihaylov and A. Frank. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia, July 2018. Association for Computational Linguistics.
41. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS '13)*, pages 3111–3119, 2013.
42. A.-L. Minard, M. Speranza, E. Agirre, I. Aldabe, M. van Erp, B. Magnini, G. Rigau, and R. Urizar. SemEval-2015 task 4: TimeLine: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, Denver, Colorado, June 2015. Association for Computational Linguistics.
43. M. Mnasri. Recent advances in conversational NLP: Towards the standardization of Chatbot building. *arXiv preprint arXiv:1903.09025*, 2019.
44. N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.
45. Q. Ning, H. Wu, and D. Roth. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia, July 2018. Association for Computational Linguistics.

46. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS '19),*, pages 8026–8037. 2019.
47. J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.
48. M. E. Peters, M. Neumann, R. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
49. H. Rashkin, M. Sap, E. Allaway, N. A. Smith, and Y. Choi. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia, July 2018. Association for Computational Linguistics.
50. N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
51. M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 3027–3035, 2019.
52. K. Schuler. *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.
53. R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4444–4451, 2017.
54. S. Storks, Q. Gao, and J. Y. Chai. Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60, 2019.
55. S. Storks, Q. Gao, and J. Y. Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, 2019.
56. Z. Sun, C. Fan, Q. Han, X. Sun, Y. Meng, F. Wu, and J. Li. Self-Explaining Structures Improve NLP Models. *arXiv preprint arXiv:2012.01786*, 2020.
57. H. Takemura and K. Tajima. Tweet Classification based on Their Lifetime Duration. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pages 2367–2370, 2012.
58. A. Tamborrino, N. Pellicanò, B. Pannier, P. Voitot, and L. Naudin. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online, July 2020. Association for Computational Linguistics.
59. A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox. Natural Language Processing Advancements By Deep Learning: A Survey. *arXiv preprint arXiv:2003.01200*, 2020.
60. T. H. Trinh and Q. V. Le. A Simple Method for Commonsense Reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

61. T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard. Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33nd International Conference on Machine Learning (ICML '16)*, pages 2071–2080, 2016.
62. S. Vashishtha, A. Poliak, Y. K. Lal, B. Van Durme, and A. S. White. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online, Nov. 2020. Association for Computational Linguistics.
63. S. Vashishtha, B. Van Durme, and A. S. White. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy, July 2019. Association for Computational Linguistics.
64. D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.
65. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.
66. X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei, K. Talamadupula, and A. Fokoue. Improving Natural Language Inference Using External Knowledge in the Science Questions Domain. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 7208–7215, 2019.
67. Z. Wang, J. Zhang, J. Feng, and Z. Chen. Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14)*, pages 1112–1119, 2014.
68. A. S. White, P. Rastogi, K. Duh, and B. Van Durme. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, Taipei, Taiwan, Nov. 2017. Asian Federation of Natural Language Processing.
69. R. W. White and A. Hassan Awadallah. Task Duration Estimation. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*, pages 636–644, 2019.
70. A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
71. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*, 2019.
72. W. Xiang and B. Wang. A Survey of Event Extraction From Text. *IEEE Access*, 7:173111–173137, 2019.
73. M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online, June 2021. Association for Computational Linguistics.
74. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
75. T. Zhang, Z. Cai, C. Wang, P. Li, Y. Li, M. Qiu, C. Tang, X. He, and J. Huang. HORNET: Enriching Pre-trained Language Representations with Heterogeneous Knowledge Sources.

In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21)*, pages 2608–2617, 2021.

76. Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics.
77. B. Zhou, D. Khashabi, Q. Ning, and D. Roth. "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, Nov. 2019. Association for Computational Linguistics.
78. B. Zhou, Q. Ning, D. Khashabi, and D. Roth. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online, July 2020. Association for Computational Linguistics.