Lifelong Knowledge-Enriched Social Event Representation Learning

Prashanth Vijayaraghavan
MIT Media Lab
Cambridge, MA, USA
pralav@mit.edu

Deb Roy
MIT Media Lab
Cambridge, MA, USA
dkroy@media.mit.edu

Abstract

The ability of humans to symbolically represent social events and situations is crucial for various interactions in everyday life. Several studies in cognitive psychology have established the role of mental state attributions in effectively representing variable aspects of these social events. In the past, NLP research on learning event representations often focuses on construing syntactic and semantic information from language. However, they fail to consider the importance of pragmatic aspects and the need to consistently update new social situational information without forgetting the accumulated experiences. In this work, we propose a representation learning framework to directly address these shortcomings by integrating social commonsense knowledge with recent advancements in the space of lifelong language learning. First, we investigate methods to incorporate pragmatic aspects into our social event embeddings by leveraging social commonsense knowledge. Next, we introduce continual learning strategies that allow for incremental consolidation of new knowledge while retaining and promoting efficient usage of prior knowledge. Experimental results on event similarity, reasoning, and paraphrase detection tasks prove the efficacy of our social event embeddings.

1 Introduction

Everyday life comprises the ways in which people typically act, think, and feel on a daily basis. Our life experiences unfold naturally into temporally extended daily events. The event descriptions can be packaged in various ways depending on several factors like speaker's perspective or the related domain. Interpretation of event descriptions will be incomplete without understanding multiple entities involved in the events and even more so when the focus is primarily on "social events",

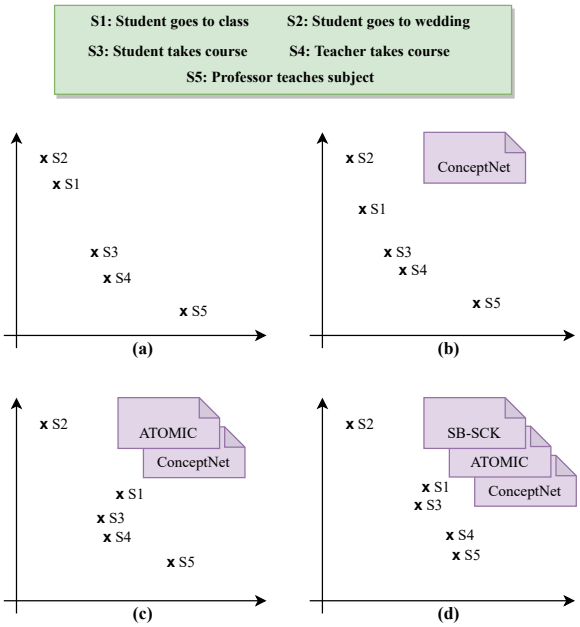

Figure 1: Illustration of functioning of our lifelong representation learning approach that produces incrementally richer social event representations. Event texts are given in the top green box. With more knowledge, social event embeddings move beyond high lexical overlap [shown in (a)] and learn to integrate semantic and pragmatic properties [shown in (b), (c)] of event texts along with social role information [shown in (d)].

i.e., events explaining social situations and interactions. Therefore, a social event representation model must capture the semantic properties from the event text description and embed salient knowledge that encompasses the implicit pragmatic abilities. Early definitions of pragmatic aspects refer to the use of language in context; comprising the verbal, paralinguistic, and non-verbal elements of language (Adams et al., 2005). Contemporary definitions have expanded beyond just communicative functions to include behavior that includes social, emotional, and communicative aspects of language (Adams et al., 2005; Parsons et al., 2017).

Moving away from the extensively studied

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3624–3635
April 19 - 23, 2021.

speech acts, we analyze characteristics that reflect how a person behaves in social situations and how social contextual aspects influence linguistic meaning. In the context of event representations, the pragmatic properties can specifically refer to the human's inferred implicit understanding of event actors' intents, beliefs, and feelings or reactions (Wood, 1976; Hopper and Naremore, 1978).

Understanding the pragmatic implications of social events is non-trivial for machines as they are not explicitly found in the event texts. Prior studies (Ding et al., 2014, 2015; Granroth-Wilding and Clark, 2016; Weber et al., 2018) often extract the syntactic and semantic information from the event descriptions but ignore the pragmatic aspects of language. In this work, we address this shortcoming, and aim to (a) disentangle semantic and pragmatic attributes from social event descriptions and (b) encapsulate these attributes into an embedding that can move beyond simple linguistic structures and dispel apparent ambiguities in the real sense of their context and meaning.

Towards this goal, we propose to train our models with social commonsense knowledge about events focusing specifically on intents and emotional reactions of people. Such commonsense understanding can be obtained from existing knowledge bases like ConceptNet (Speer et al., 2017), Event2Mind/ATOMIC (Sap et al., 2019a; Rashkin et al., 2018) or by collecting more noisy commonsense knowledge using data mining techniques. As new domain sources emerge, each containing different knowledge assertions, it is essential that the representation models for social events keep evolving with this growing knowledge. Since it is generally infeasible to retrain models from scratch for every new knowledge source, we consider the need to employ prominent continual learning practices (Kirkpatrick et al., 2017; Lopez-Paz and Ranzato, 2017; Asghar et al., 2018; d'Autume et al., 2019) and enable semantic and pragmatic enrichment of social event representations. This problem can be addressed from the perspective of incremental domain adaptation (Asghar et al., 2018; Wulfmeier et al., 2018), which quickly adapts to new domain knowledge without interfering with existing ones. Figure 1 presents a sample functioning scenario producing incrementally richer social event embeddings. As the model gains more knowledge from different sources, it learns to discern events based on semantic and pragmatic properties, including social roles. For example, "Student takes course", and "Teacher takes course" has significant lexical and semantic relatedness. However, the social role information changes the meaning as depicted in Figure 1(d) with the introduction of our in-house dataset (SB-SCK).

In this paper, we develop a lifelong representation learning approach for embedding social events from their free-form textual descriptions. Our model augments a growing set of knowledge obtained from various domain sources to allow for positive knowledge transfer across these domains. Our contributions are as follows:

- We propose a continual representation learning approach-that integrates both text encoding and lifelong learning techniques to aid better representation of social events.
- We adopt a domain-representative episodic memory replay strategy with text encoding techniques to effectively consolidate the expanding knowledge from several domain sources and generate a semantically & pragmatically enriched social event embedding.
- We evaluate our models primarily on four different tasks: (a) intent-emotion prediction for event texts based on an in-house *Lifelong EventRep Corpus*, (b) event similarity task using hard similarity dataset (Ding et al., 2019; Weber et al., 2018), (c) paraphrase detection using Twitter URL corpus (Lan et al., 2017), and (d) social commonsense reasoning task using SocialIQA (Sap et al., 2019b) dataset.

2 Related Work

2.1 Social Events Representation Learning

Early work in the domain of events can be traced back to modeling narrative chains. Chambers and Jurafsky (Chambers and Jurafsky, 2008, 2009) introduced models for event sequences involving coreference resolution and inferring event schemas. Similar efforts (Balasubramanian et al., 2013; Cheung et al., 2013; Jans et al., 2012) have explored the use of open-domain relations to extract event schemas but suffer from reduced predictive capabilities and increased sparsity. Recent advancements, aimed at addressing the limitations of prior works, compute distributed embeddings of events involving word embeddings, recurrent sequence models, and tensor-based composition models (Modi and

Titov, 2013; Granroth-Wilding and Clark, 2016; Pichotta and Mooney, 2016; Hu et al., 2017). Specifically, tensor-based methods have demonstrated improved performance by representing events that predict implicit arguments with event knowledge (Cheng and Erk, 2018), combine (subject, predicate, object) triples information (Weber et al., 2018) and reflect thematic fit (Tilk et al., 2016).

2.2 Lifelong Learning

Lifelong learning or continual learning approaches can be grouped into regularization-based, data-based, and model-based approaches. Regularization based approaches (Kirkpatrick et al., 2017; Schwarz et al., 2018; Zenke et al., 2017) minimize significant changes to the previously learned representations as we update parameters for the current task. This is usually implemented as an additional constraint to the objective function based on the sensitivity of parameters. Recent studies (Kemker and Kanan, 2017; d'Autume et al., 2019; Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019) under data-based approaches store previous task data either using a replay memory buffer or a generative model. In NLP domain, lifelong language learning approaches have investigated the use of memory replay and local adaptation techniques (d'Autume et al., 2019).

Finally, model-based approaches allow models to allocate or grow capacity (layers or features) necessary for the tasks (Rusu et al., 2016; Lee et al., 2016). More recently, (Asghar et al., 2018) augmented RNNs with a progressive memory bank leading to increased model capacity. However, the challenges related to increased architectural complexity are tackled by hybrid models as in (Sodhani et al., 2020). In this paper, we build on ideas from the hybrid models and apply them for our learning task. While previous studies (Ding et al., 2019) have attempted to incorporate commonsense knowledge, this work is one of the first efforts to integrate multi-source knowledge and address it through the lens of incremental domain adaptation.

3 Problem Formalization

Formally, we assume that our learning framework has access to streams of social commonsense knowledge data obtained from n different domains, denoted by $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$. At a particular point in time, we extract knowledge from the current domain $\mathcal{D}_i$. We produce an embedding of social events by consolidating the accumulated knowledge across the modeled domains $\mathcal{D}_{\leq i}$. Data from each domain source contains source-specific textual descriptions of social situations and their intuitive commonsense information such as intents and emotions. Training samples, drawn from a domain dataset $\mathcal{D}_i$, could contain either a significant overlap or an entirely new set of knowledge when compared with the previously processed domains $\mathcal{D}_{1:i}$. Given such a setup, we aim to generate incrementally richer social event representations using our continual learning framework.

4 Datasets

For our representation learning task, we aggregate social commonsense knowledge data from various domain sources. This knowledge contains details about pragmatic aspects like intents and emotional reactions. We create a continual learning benchmark based on these commonsense data sources[1].

4.1 Lifelong Social Events Dataset

Different domain sources of social commonsense knowledge used for training our social event representation model are explained as follows.

ATOMIC dataset consists of inferential knowledge based on 24k short events covering a diverse range of everyday events and motivations. Though each event contains nine dimensions per event, the scope of this work will be limited to intent and emotions as our inferential pragmatic dimensions.

CONCEPTNET knowledge base contains several commonsense assertions. For our purpose, we choose ConceptNet's relevant relations: /r/MotivatedByGoal, /r/CausesDesire, /r/Entails, /r/Causes, /r/HasSubevent. We convert triples in the dataset into template form.

SB-SCK Since social roles (e.g., student, mother, teacher, worker, etc.) provide additional information about the motives and emotions behind actions specified in the events (as shown in Figure 1, 2, we adopt web-based knowledge mining techniques for capturing this aspect. This dataset was collected as a part of our recent work (Vijayaraghavan and Roy, 2021) using the following steps: (a) process texts from Reddit posts containing personal narratives as in (Vijayaraghavan and

[1]The project details about future data/code releases or any updates will be available at https://pralav.github.io/lifelong_eventrep?c=10

Search-based Social Commonsense Knowledge		
Social Roles	Event Phrases	Motives
Politicians	use social media	to woo voters
Activists		to create a movement
Police		to connect with residents and solve crime
Workers	gather around table	to solve business problems
Priests		to pray to god, share wine and bread
Friends		to share a meal, conversation

SB-SCK Dataset	
#events w/ motives	103,357
#events w/ emotions	69,584
#unique social roles	586

Figure 2: **Left**: Samples from Search-based Social Commonsense Knowledge (SB-SCK) dataset with highlighted motivations for social roles, **Right**: Statistics of SB-SCK dataset.

Roy, 2021), (b) extract propositions from text using OpenIE tools, (c) perform a web search for plausible intents and emotions by attaching purpose clauses (Palmer et al., 2005) and feelings lexical units from Framenet (Baker et al., 1998) and (d) finally, remove the poorly extracted facts using a simple classifier trained on some seed commonsense knowledge. Figure 2(Left) shows samples from this dataset indicating how the same action could have different social role related motivations. We refer to this as Search-based Social Commonsense Knowledge (SB-SCK) data. Figure 2(Right) presents the data statistics.

For data from each of the above domain sources, we sample free-form event text, its paraphrase, intent, emotional reactions, and negative samples of paraphrases, intents, and emotional reactions. Based on the annotated labels for motivation (Maslow's) and emotional reactions (Plutchik) in STORYCOMMONSENSE data, we run a simple K-Means clustering on the open text intent data. We identify five disjoint clusters on each of the three domains and map them to those categories. For the purpose of our lifelong learning problem, we divide each domain data into two sets (3 clusters and 2 clusters) and consider them as different subdomains. Therefore, this results in 6 tasks in our continual learning setup. We refer to this dataset as *Lifelong EventRep Corpus*.

4.2 Paraphrase Datasets

We use random samples of parallel texts from paraphrase datasets like PARANMT-50M corpus(Wieting and Gimpel, 2017) and Quora Question Pair dataset [2]. These paraphrase datasets are primarily used for pretraining our model. We also produce paraphrases of free-form event texts in our dataset using a back-translation approach (Iyyer et al., 2018). We used pretrained English↔German translation models for this purpose.

[2]https://www.kaggle.com/c/quora-question-pairs/data

5 Framework

Our goal is to learn distributed representations of social events by incorporating pragmatic aspects of the language beyond shallow event semantics. Moving away from conventional supervised multi-task classification based lifelong learning approaches, we focus on a lifelong representation learning approach that enables us to adapt and sequentially learn a social event embedding model. The motivation for a lifelong learning framework is that the growing knowledge obtained from various domain sources can effectively guide the modeling of complex social events. This involves systematically updating the model by consolidating this expanding knowledge to produce richer embeddings without forgetting previously accumulated knowledge. In this section, we will explain various components of our modeling framework.

5.1 Social Event Representation

Given an input event text description, the core idea is to first encode the free-form event text and decompose the ensuing representation into pragmatic (implied emotions and intents) and non-pragmatic (syntactic and semantic information) components. Eventually, we combine these decomposed representations to obtain an overall event representation and apply it in different downstream tasks.

5.1.1 Encoder

The input to our model is a free-form event text description from i^{th} domain, $x_j^{(i)} \in \mathcal{D}_i$. This free-form event text contains a sequence of tokens, $x_j^{(i)} = [w_1, w_2, ..., w_L]$, where each token $w_{(.)}$ is obtained from an input vocabulary $\mathcal{V}$. The model encodes the input event text $x_j^{(i)} \in \mathcal{R}^{L \times d_X}$ in multiple steps. First, we construct a context-dependent token embedding using a context embedding function $\mathcal{G} : \mathcal{R}^{L \times d_X} \mapsto \mathcal{R}^{L \times d_H}$, where d_X and d_H refer to the embedding and hidden layer dimensions respectively. Following this encoding step, we incorporate pooling or projection function, $\mathcal{G}_{p\bar{p}}$: $\mathcal{R}^{L \times d_H} \mapsto \mathcal{R}^{3 \times d_H}$, that transform event text from context-dependent embedding space into pragmatic and semantic space. More specifically, we produce latent vectors for intents (h_I), reactions (h_R) and non-pragmatic (h_N) information. Finally, we com-

bine the latent vectors h_N, h_I, h_R using a simple feed-forward layer, $\mathcal{G}_C$: $\mathcal{R}^{3\times d_H} \mapsto \mathcal{R}^{d_H}$, to produce a powerful social event representation, h_C, capable of dispelling apparent ambiguities in the true sense of their meaning. Given positive and negative examples of intents, emotional reactions and paraphrases associated with the input event text, we learn to effectively sharpen each of these embeddings h_I, h_R and using metric learning methods.

For the sake of brevity, we drop the domain index i and the sample index j in this section. These encoding steps are summarized as:

$$H_e = [h_1, h_2, ..., h_L] = \mathcal{G}([w_1, w_2, ..., w_L]) \quad (1)$$

$$h_I, h_R, h_N = \mathcal{G}_{p\bar{p}}(H_e) \quad (2)$$

$$h_C = \mathcal{G}_C(h_I, h_R, h_N) \quad (3)$$

We denote this multi-step encoding process resulting in h_I, h_R, h_C as a function $\mathcal{G}_{event}$. Now, we experiment with the following text embedding techniques as our context embedding function ($\mathcal{G}$):

BiGRU Using bidirectional GRUs (Chung et al., 2014), we compute the context embedding of the input event text by concatenating the forward ($\overrightarrow{h_t}$) and backward hidden states ($\overleftarrow{h_t}$), $\overleftrightarrow{h_t} = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

BERT We employ BERT (Devlin et al., 2018), a multi-layer bidirectional Transformer-based encoder, as our context embedding method $\mathcal{G}$. We fine-tune a BERT model that takes attribute-augmented event text $x = [CLS]\ m\ [SEP]\ w_1, ..., w_L\ [SEP]$ as input and outputs a powerful context-dependent event representation H_e. The attribute $m \in \{xIntent, xReact, xNprag\}$ refers to special tokens for intents, reactions and non-pragmatic aspects.

In our default case, our $\mathcal{G}_{p\bar{p}}$ function is the output embedding of $[CLS]$ token associated with their respective attribute-augmented input. In cases where input event text is not augmented with attribute special tokens, we apply pooling strategies such as attentive pooling (AP) and mean (MEAN) of all context vectors obtained from the previous encoding step $\mathcal{G}$. We obtain h_I, h_R, h_N based on these techniques. Depending on the type of context embedding function, we refer our multi-step event text encoder, $\mathcal{G}_{event}$, as EVENTGRU or EVENTBERT.

5.1.2 Objective Loss

Using positive $\{u_I^p, u_R^p, u_C^p\}$ and $N-1$ negative $\{u_I^n, u_R^n, u_C^n\}$ examples of intents, emotions and paraphrases associated with the event texts, we calculate N-pair loss, $\mathcal{L}_v(h, z^p, \{z_k^n\}_{k=1}^{N-1})$, to maximize the similarity between the representation of positive examples (z_v^p) and the computed embeddings (h_v). Here, z_v^e is computed using a transformation function f_v as: $z_v^e = f_v(u_v^e)$, where $v \in \{I, R, C\}$ and $e \in \{p, n\}$. Thus, our loss function is devised as:

$$\mathcal{L}_\mathcal{T} = \beta_D \cdot (\mathcal{L}_I + \mathcal{L}_R) + \beta_E \cdot \mathcal{L}_C \quad (4)$$

where $\mathcal{L}_I, \mathcal{L}_R$ are used to learn disentangled pragmatic embeddings (intent and emotion), $\mathcal{L}_C$ is intended to jointly embed semantic and pragmatic aspects to produce an overall social event representation. β_D, β_E are loss coefficients that weigh the importance of disentanglement loss and an overall joint embedding loss. These coefficients are non-negative and they sum to 1.

5.2 Continual Learning

Given a never-ending list of social events, once-and-for-all training on a fixed dataset limits the utility of such models in real-world applications. Therefore, we draw ideas from lifelong learning literature to adapt our models to new data yet retaining prior knowledge. First, we implement a data-based approach, which is a variant of episodic memory replay (EMR) (Lopez-Paz and Ranzato, 2017; Chaudhry et al., 2019; d'Autume et al., 2019) for mitigating catastrophic forgetting while allowing for beneficial backward knowledge transfer. Next, we combine it with simple vocabulary expansion for incremental domain adaptation.

5.2.1 Domain-Representative Episodic Memory Replay (DR-EMR)

We augment our model explained in Section 5.1 with an episodic memory module to perform sparse experience replay. As we train our lifelong representation learning model, we create mixed training mini-batches ($\mathcal{B}_{dom}, \mathcal{B}_{rep}$) by drawing samples from: (a) new domain dataset, $\mathcal{B}_{dom} \subset \mathcal{D}_i$ and (b) episodic memory containing old domain samples, $\mathcal{B}_{rep} \subset \mathcal{M}$. Training our model using mixed mini-batches introduces parameter changes that alter past event data embeddings, including those stored in our episodic memory, $\mathcal{M}$. The episodic memory module is implemented as a "Read-Write" memory store containing selective event samples from the original domain dataset and their respective embeddings. The key memory operations are stated as follows:

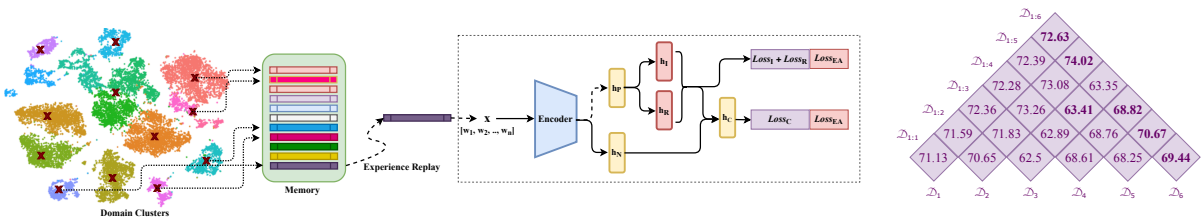

Figure 3: **Left:** Illustration of our Lifelong Social Event Representation Model, **Right:** Average accuracy score (%) of one specific permuted run across 6 domains. Table shows the dynamics of our continual learning model allowing positive backward transfer, i.e., performance increases in previous domains gradually with new knowledge.

Read Operation The read operation retrieves domain-specific random samples from our episodic memory for experience replay. These samples contain the original event training data and their representations. We choose samples from a previously trained domain at every read step assuming an overall uniform distribution over domains.

Write Operation In this work, we incorporate the following desirable characteristics of an episodic memory module: (a) $\mathcal{M}$ stores domain-representative samples that best approximate the domain's data distribution and (b) $\mathcal{M}$'s capacity is bounded or at most expands sub-linearly. To achieve this, we adopt the following strategies:

Domain-Representative Sample Selection: Past studies have explored using random writing strategy (d'Autume et al., 2019) and distribution approximation or K-Means to cluster the samples (Rebuffi et al., 2017). In our work, we perform sample selection using a CURE (Clustering Using REpresentatives) algorithm (Guha et al., 2001) to find C representative points. CURE employs hierarchical clustering that is computationally feasible by adopting random sampling and two-pass clustering. Using event representations obtained after embedding alignment transformation, we identify the domain-representative samples by computing euclidean metric-based nearest neighbors to the C representative points. We set the value of C based on the memory budget. These domain-representative samples are stored in our episodic memory with their corresponding representations.

Replacement Policy: When the memory becomes full, we follow a simple memory replacement policy that selects an existing memory entry to delete. Specifically, we replace the q^{th} memory entry which is determined by: $q = argmax_q(softmax(\phi(x_j^{(t)}) \cdot \mathcal{M}_q)$ similar to idea proposed in (Gulcehre et al., 2018).

Inspired by Wang et al. (Wang et al., 2019), we propose a variant of the alignment model that helps overcome catastrophic forgetting by ensuring minimal distortion to previously computed representational spaces. To accomplish this, we define simple linear transformations: $\mathcal{G}_{IA}, \mathcal{G}_{RA}, \mathcal{G}_{CA}$ on the top of the multi-step encoder function $\mathcal{G}_{event}$ outputs. For simplicity, we drop the subscripts (I, R, C) and denote these transformation functions as $\mathcal{G}_A$. Given a new domain i, we initialize our multi-step event encoder $\mathcal{G}^i_{event}$ and alignment function $\mathcal{G}^i_A$ with the last trained parameters as in $\mathcal{G}^{i-1}_{event}$ and $\mathcal{G}^{i-1}_A$ respectively. The optimization is implemented in the following two steps: (a) For samples from $(\mathcal{B}_{dom}, \mathcal{B}_{rep})$, we optimize the encoder output representations to be closer to their respective ground-truth examples in this linearly transformed space. Therefore, we modify the N-pair loss function to accommodate the alignment function as: $\mathcal{L}(\mathcal{G}^i_A(\mathcal{G}^i_{event}(x)),\ \mathcal{G}^i_A(z^p),\ \{\mathcal{G}^i_A(z^n)\}_{k=1}^{N-1})$; (b) For samples from $\mathcal{B}_{rep}$, we add an extra constraint $(\mathcal{L} + \mathcal{L}_{EA})$ for each embedding component to align old and new domain embedding spaces:

$$\mathcal{L}_{EA} = ||\mathcal{G}^i_A(\mathcal{G}^i_{event}(x)) - \mathcal{G}^{i-1}_A(\mathcal{G}^{i-1}_{event}(x))||^2 \quad (5)$$

6 Training

Since our model involves metric learning, hard negative data mining is an essential step for faster convergence and improved discriminative capabilities. However, selecting too hard examples too often makes the training unstable. Therefore, we choose a hybrid negative mining technique where we choose few semi-hard negatives examples (Hermans et al., 2017) and combine it with random negative samples to effectively train our model.

In our work, we define a heuristic objective by weighing samples based on two factors: (i) word overlap or similarity in embedding space of the event text and (ii) intent and emotion free-form text or categories based on STORYCOMMONSENSE data. More specifically, given an event text as anchor and a positive intent text based on a ground

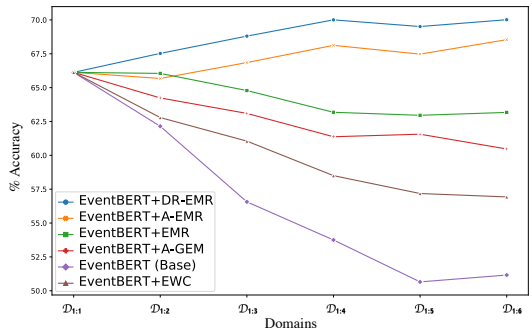

Figure 4: Average accuracy scores (%) of data after 6 permuted runs from domains that have been observed at that point during the continual learning process.

truth motivation category, we mine negative instances for intent as follows: (a) choose random text samples associated with a motivation category that is different from that of the positive example but closer in the embedding space or word overlap, (b) choose random text samples within the same motivation category but with different emotion category. We repeat this process for drawing negative instances related to emotions. For paraphrases, we consider few examples with significant word overlap while the rest are randomly chosen examples. N-pair loss helps alleviate the sensitivity of triplet loss function to the choice of hard triplets. Finally, we pre-train our model with paraphrase data and fine-tune it using the examples obtained from hard negative mining for intents, emotions and paraphrases. For our training, the learning rate is set to 0.0001, the number of training epoch is 20. By default, we use EVENTBERT as our multi-step encoder. We conduct a study by assigning different values for loss coefficients, β_D, β_E, and explain the results of the study in Section 7.1.2.

7 Experiments

We conduct several experiments to study the power of our learned embeddings. Our experiments are designed to answer the following questions:

RQ1: How well does our model perform in comparison to other continual learning approaches for intent-emotion prediction?

RQ2: To what extent do our modeling choices impact the results in predicting intents & emotions?

RQ3: Does our model outperform existing state-of-the-art methods in hard similarity task that evaluates the effectiveness of the learned embeddings?

RQ4: Do the learned embeddings demonstrate transfer capability to downstream tasks – Paraphrase detection & Social IQA reasoning?

7.1 Intent-Emotion Prediction (RQ1)

The continual learning methods evaluated using our *Lifelong EventRep Corpus* are given below.

- **Base**: We simply fine-tune our model on successive tasks from previously trained checkpoint.
- **A-GEM**: This method (Chaudhry et al., 2019) uses a constraint that enables the projected gradient to decrease the loss on older tasks. We randomly choose 2-3% samples from all the previous tasks to form a constraint.
- **EWC**: A regularization-based technique, Online-EWC (Schwarz et al., 2018), is used to address catastrophic forgetting.
- **EMR**: We use randomly stored examples for sparse experience replay.
- **A-EMR**: A variant of our model that uses random samples for experience replay with alignment constraints.
- **DR-EMR**: This is our complete model involving domain representative experience replay and alignment constraints.

7.1.1 Empirical Results

After running six permuted sequence of tasks, we calculate the mean performance on the test set of all observed task domains after time step k, given by $AvgAcc = \frac{1}{k}\sum_{i=1}^{k} Acc^{(i)}$, where $Acc^{(i)}$ is the model accuracy on the test set from the domain $\mathcal{D}_i$. Further, we also compute standard deviation to determine the importance of the order of the tasks during training. Table 1a contains the comparison of last-step $AvgAcc$ scores and standard deviation for predicting intents and emotions. Figure 4 plots the $AvgAcc$ at every step where $\mathcal{D}_{1:i}$ indicates that the model has seen data from i domains to evaluate our continual learning process.

In the absence of any lifelong learning strategies, the performance drops significantly for *Base* model while emphasizing the importance of task order as indicated by the high standard deviation value. Compared to the *Base* model, our results show some improvement in intent and emotion prediction as we introduce methods like *A-GEM* and *EWC*. However, we observe a significant performance gain with *EMR*-based techniques. Our complete model (*DR-EMR*) outperforms all the other methods, thereby demonstrating the importance of

Models	Intents		Emotions	
	AvgAcc	Std	AvgAcc	Std
EVENTBERT (Base)	51.16	10.41	62.27	9.74
EVENTBERT + A-GEM	60.47	6.75	69.44	6.12
EVENTBERT + EWC	56.93	8.86	66.89	7.55
EVENTBERT + EMR	63.17	4.20	72.04	4.66
EVENTBERT + A-EMR	68.54	2.85	73.42	3.10
EVENTBERT+ DR-EMR	**70.02**	**0.38**	**78.48**	**0.65**

(a) Lifelong EventRep Dataset

Models	% Acc.
KGEB (Ding et al., 2016)	50.09
NTN + Int+ (Ding et al., 2019)	58.83
NTN + Int + Senti (Ding et al., 2019)	64.31
EVENTBERT$_{\beta_E=0.3}$+ DR-EMR	66.19
EVENTBERT$_{\beta_E=0.5}$+ DR-EMR	**71.23**
EVENTBERT$_{\beta_E=0.7}$+ DR-EMR	69.79

(b) Hard Similarity Dataset

Table 1: Evaluation results on: (a) held-out Lifelong EventRep test set and (b) combined hard similarity dataset.

domain-representative sampling and alignment constraints towards learning representations that help effective prediction of intents and emotions. Moreover, we assess the domain sequences that cause a performance drop. For example, whenever SBSCK is trained at the end, the model shows reduced accuracy. The reason can be ascribed to the effect of interference of noisy knowledge over the previously trained cleaner domains. Training this noisy source ahead leads to positive knowledge transfer and hence produces sharpened performance outcomes. Despite these odds, our *DM-EMR* model records the least standard deviation implying reduced sensitivity to training order. Figure 3 (Right) shows the dynamics of our continual learning mode. It contains accuracy scores of a single run and displays the lower-triangular values. We see that our approach allows positive backward transfer, i.e., model performance in previous domains gradually increases with new domain knowledge. Our model achieves the best performance (see the bold-faced accuracy scores in Figure 3(Right)) in most domains after observing all the data ($\mathcal{D}_{1:6}$).

7.1.2 Ablation Study (RQ2)

We analyze different model configurations related to: (a) encoding: EVENTGRU, EVENTBERT, (b) pooling: attribute-augmented input (CLS), Mean Pooling (MP) and Attentive Pooling (AP) and (c) sampling: K-Means, CURE. Results of the study are given in Figure 5(Left). We evaluated different combinations of these strategies but only report the average accuracy scores of configurations having the best strategy at each category combined with the variants in the following category, i.e., for pooling strategies, we only report scores with EVENTBERT encoding strategy and so on. From the results, we ascertain that the best configuration comprises an EVENTBERT encoder supported by attentive pooling and CURE-based sample selection.

Methods	Intents	Emotions
Encoding Strategy		
EVENTGRU	60.92	70.61
EVENTBERT	**68.56**	**78.48**
Pooling Strategy		
MP	65.50	76.13
AP	67.28	76.69
CLS	**68.56**	**78.48**
Sampling Strategy		
K-Means	66.45	75.88
CURE	**68.56**	**78.48**

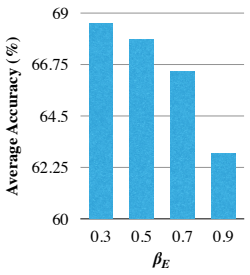

Figure 5: Results of our ablation study on a held-out validation set. **Left:** ***AvgAcc*** scores (%) on intent-emotion prediction task using different encoding, pooling & sampling strategies. **Right:** ***AvgAcc*** scores (%) to measure the effect of β_E in predicting intents.

Additionally, we measure the effect of β_E in the prediction of intents. As shown in Figure 5 (Right), the model performs significantly better for lower values β_E as more weight is assigned for the disentanglement of pragmatic aspects. However, we observe that a balanced loss function with $\beta_E = 0.5$ allows for consistently good performance in both intent-emotion prediction (Figure 5 (Right)) and hard similarity tasks (see Section 7.2). Despite other hyperparameters, changes to β_E determine the importance of incorporating semantic or pragmatic information in the ensuing event embedding.

7.2 Hard Similarity Task (RQ3)

By following the work of Ding et al. (Ding et al., 2019), we evaluate our social event representation on an extended dataset of event pairs containing: (a) similar event pair having minimum lexical overlap (e.g., people admired president / citizens loved leader) (b) dissimilar event pair with high lexical overlap (e.g., people admired president / people admired nature). A good performance in this task will ensure that similar events are pulled closer to each other than dissimilar events. Combining hard similarity datasets from (Ding et al., 2019) and (Weber et al., 2018), the total size of this balanced

dataset is 2,230 event pairs. Using our joint embedding h_C for an event text and triplet loss setup, we compute a similarity score between similar and dissimilar pairs. The baselines include: Knowledge-graph based embedding model (KGEB) (Ding et al., 2016), Neural Tensor Network (NTN) and its variants augmented with ATOMIC dataset based embeddings (Int, Senti) (Ding et al., 2019). We report the model's accuracy in assigning a higher similarity score for similar pairs than dissimilar pairs. Table 1b shows that our model outperforms the state-of-the-art method for this task.

7.3 Paraphrase Detection (RQ4)

Given a sentence pair, the objective is to detect whether they are paraphrases or not. For each sentence pair (s_1, s_2), we pass them through our model and obtain their respective h_C, given by vectors (u, v). We concatenate these vectors (u, v) with the element-wise difference $|u - v|$ and feed to a feed-forward layer. We optimize binary cross-entropy loss. For evaluation purposes, we compare our model against baselines like BERT and ESIM (Chen et al., 2016). Trained on a subset of dataset explained in Section 4.2, we choose an out-of-domain test dataset where samples stem from a dissimilar input distribution. To this end, Twitter URL paraphrasing corpus (Lan et al., 2017), referred to as TwitterPPDB, is selected. Table 2b contains results of our evaluation. The results testify the efficacy of our embeddings.

7.4 Social IQA Reasoning (RQ4)

We determine the quality of our latent social event representations by evaluating on a social commonsense reasoning benchmark – SocialIQA dataset (Sap et al., 2019b). Given a context, a question and three candidate answers, the goal is to select the right answer among the candidates. Following Sap et al.(Sap et al., 2019b), the context, question, and candidate answer are concatenated using separator tokens and passed to the BERT model. Additionally, we feed the context to our EVENTBERT model to obtain three embeddings h_I, h_R, h_C. While the original work computed a score l using the hidden state of $[CLS]$ token, we introduce a minor modification to this step as: $l = W_5 tanh(W_1 h_{CLS} + W_2 h_I + W_3 h_R + W_4 h_C)$, where $W_{1:4} \in \mathcal{R}^{d_H \times d_H}$ and $W_5 \in \mathcal{R}^{1 \times d_H}$ are learnable parameters. Similar to (Sap et al., 2019b), triple with the highest normalized score is used as the model's prediction. We fine-tune BERT models using our new scoring function with social event embedding (denoted as "w/") and compare against baselines (like GPT (Radford et al., 2018)) without our event embeddings (denoted as "w/o"). Results in Table 2a indicate that a simple enhancement procedure at the penultimate step can offer significant performance gains. Our findings also suggest that our enhanced model performed well for question types like 'wants' and 'effects' that weren't explicitly modeled in our embedding model. This confirms that our pragmatics-enriched embeddings lead to improved reasoning capabilities.

Models	Dev	Test
w/o Social Event Embeddings		
GPT	63.3	63.0
BERT-base	63.3	63.1
BERT-large	**66.0**	**64.5**
w/ Social Event Embeddings		
BERT-base	65.1	64.0
BERT-large	**68.7**	**67.9**

(a) SocialIQA Dataset

Models	% Acc
ESIM	84.01
BERT	87.63
EVENTBERT$_{0.5}$	88.23
EVENTBERT$_{0.7}$	**90.16**

(b) TwitterPPDB dataset

Table 2: Accuracy scores (%) of different models on: (a) SocialIQA dev & test set, and (b) Twitter URL Paraphrasing corpus, TwitterPPDB. EVENTBERT models employ DR-EMR and subscript indicates value of β_E.

8 Conclusion

Humans rely upon commonsense knowledge about social contexts to ascribe meaning to everyday events. In this paper, we introduce a lifelong learning approach to effectively embed social events with the help of a growing set of social commonsense knowledge assertions acquired from different domains. First, we leverage social commonsense knowledge to sharpen social event embeddings with semantic and pragmatic attributes. Next, we employ domain-representative episodic memory replay (DR-EMR) to overcome catastrophic forgetting and enable positive knowledge transfer with the emergence of new domain knowledge. By evaluating on a corpus of social events aggregated from multiple sources, we establish that our model is able to outperform several baselines. Experimental results on downstream tasks like event similarity, reasoning, and paraphrase detection demonstrate the capabilities of our social event embeddings. We hope that our work will motivate further exploration into lifelong representation learning of social events and advance the research in inferring pragmatic dimensions from texts.

References

Catherine Adams, Janet Baxendale, Julian Lloyd, and Catherine Aldred. 2005. Pragmatic language impairment: case studies of social and pragmatic language therapy. *Child Language Teaching and Therapy*, 21(3):227–250.

Nabiha Asghar, Lili Mou, Kira A Selby, Kevin D Pantasdo, Pascal Poupart, and Xin Jiang. 2018. Progressive memory banks for incremental domain adaptation. *arXiv preprint arXiv:1811.00239*.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90.

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. 2013. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731.

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

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Pengxiang Cheng and Katrin Erk. 2018. Implicit argument prediction with event knowledge. *arXiv preprint arXiv:1802.07226*.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. *arXiv preprint arXiv:1302.4813*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *arXiv preprint arXiv:1906.01076*.

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

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense knowledge. *arXiv preprint arXiv:1909.05190*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2133–2142.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. 2001. Cure: an efficient clustering algorithm for large databases. *Information systems*, 26(1):35–58.

Caglar Gulcehre, Sarath Chandar, Kyunghyun Cho, and Yoshua Bengio. 2018. Dynamic neural turing machine with continuous and discrete addressing schemes. *Neural computation*, 30(4):857–884.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person reidentification. *arXiv preprint arXiv:1703.07737*.

Robert Hopper and Rita C Naremore. 1978. *Children's speech: A practical introduction to communication development*. HarperCollins Publishers.

Linmei Hu, Juanzi Li, Liqiang Nie, Xiao-Li Li, and Chao Shao. 2017. What happens next? future subevent prediction using contextual hierarchical lstm. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the*

European Chapter of the Association for Computational Linguistics, pages 336–344. Association for Computational Linguistics.

Ronald Kemker and Christopher Kanan. 2017. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. *arXiv preprint arXiv:1708.00391*.

Sang-Woo Lee, Chung-Yeon Lee, Dong-Hyun Kwak, Jiwon Kim, Jeonghee Kim, and Byoung-Tak Zhang. 2016. Dual-memory deep learning architectures for lifelong learning of everyday human behaviors. In *IJCAI*, pages 1669–1675.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Ashutosh Modi and Ivan Titov. 2013. Learning semantic script knowledge with event embeddings. *arXiv preprint arXiv:1312.5198*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Lauren Parsons, Reinie Cordier, Natalie Munro, Annette Joosten, and Renee Speyer. 2017. A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PloS one*, 12(4):e0172242.

Karl Pichotta and Raymond J Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*.

Shagun Sodhani, Sarath Chandar, and Yoshua Bengio. 2020. Toward training recurrent neural networks for lifelong learning. *Neural computation*, 32(1):1–35.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182.

Prashanth Vijayaraghavan and Deb Roy. 2021. Modeling human motives and emotions from personal narratives using external knowledge and entity tracking. In *Proceedings of The Web Conference 2021*.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. *arXiv preprint arXiv:1903.02588*.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2018. Event representations with tensor-based compositions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Barbara S Wood. 1976. Children and communication: Verbal and nonverbal language development.

Markus Wulfmeier, Alex Bewley, and Ingmar Posner. 2018. Incremental adversarial domain adaptation for continually changing environments. In *2018*

IEEE International conference on robotics and automation (ICRA), pages 1–9. IEEE.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3987–3995. JMLR. org.