

Embedding Models for Episodic Knowledge Graphs

Yunpu Ma^{a,b}, Volker Tresp^{a,b}, Erik A. Daxberger ^{1c}

^a*Siemens AG, Corporate Technology, Munich, Germany*

^b*Ludwig Maximilian University of Munich, Munich, Germany*

^c*ETH Zurich*

Abstract

In recent years a number of large-scale triple-oriented knowledge graphs have been generated and various models have been proposed to perform learning in those graphs. Most knowledge graphs are static and reflect the world in its current state. In reality, of course, the state of the world is changing: a healthy person becomes diagnosed with a disease and a new president is inaugurated. In this paper, we extend models for static knowledge graphs to temporal knowledge graphs. This enables us to store episodic data and to generalize to new facts (inductive learning). We generalize leading learning models for static knowledge graphs (i.e., Tucker, RESCAL, HolE, ComplEx, DistMult) to temporal knowledge graphs. In particular, we introduce a new tensor model, ConT, with superior generalization performance. The performances of all proposed models are analyzed on two different datasets: the Global Database of Events, Language, and Tone (GDELT) and the database for Integrated Conflict Early Warning System (ICEWS). We argue that temporal knowledge graph embeddings might be models also for cognitive episodic memory (facts we remember and can recollect) and that a semantic memory (*current* facts we know) can be generated from episodic memory by a marginalization operation. We validate this episodic-to-semantic projection hypothesis with the ICEWS dataset.

Keywords: knowledge graph, temporal knowledge graph, semantic memory, episodic memory, tensor models

¹Work done while at Siemens AG.

1. Introduction

In recent years a number of sizable Knowledge Graphs (KGs) have been developed, the largest ones containing more than 100 billion facts. Well known examples are DBpedia [1], YAGO [2], Freebase [3], Wikidata [4] and the Google KG [5]. Practical issues with completeness, quality and maintenance have been solved to a degree that some of these knowledge graphs support search, text understanding and question answering in large-scale commercial systems [5]. In addition, statistical embedding models have been developed that can be used to compress a knowledge graph, to derive implicit facts, to detect errors, and to support the above mentioned applications. A recent survey on KG models can be found in [6].

Most knowledge graphs are static and reflect the world at its current state. In reality, of course, the state of the world is changing: a healthy person becomes diagnosed with a disease and a new president is inaugurated. In this paper, we extend semantic knowledge graph embedding models to episodic/temporal knowledge graphs as an efficient way to store episodic data and to be able to generalize to new facts (inductive learning). In particular, we generalize leading approaches for static knowledge graphs (i.e., constrained Tucker, DistMult, RESCAL, HolE, ComplEx) to temporal knowledge graphs. We test these models using two temporal KGs. The first one is derived from the Integrated Conflict Early Warning System (ICEWS) data set which describes interactions between nations over several years. The second one is derived from the Global Database of Events, Language and Tone (GDELT) that, for more than 30 years, monitors news media from all over the world. In the experiments, we analyze the generalization abilities to new facts that might be missing in the temporal KGs and also analyze to what degree a factorized KG can serve as an explicit memory.

We propose that our technical models might be related to the brain’s explicit memory systems, i.e., its episodic and its semantic memory. Both are considered long-term memories and store information potentially over the life-time of an individual [7, 8, 9, 7]. The semantic memory stores general factual knowledge,

i.e., information we *know*, independent of the context where this knowledge was acquired and would be related to a static KG. Episodic memory concerns information we *remember* and includes the spatiotemporal context of events [10] and would correspond to a temporal KG.

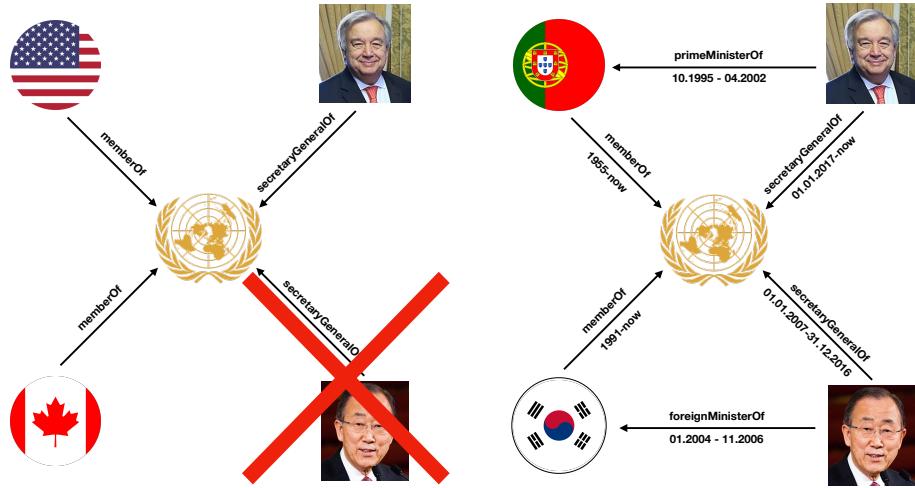


Figure 1: Illustrations of (left) a semantic knowledge graph and (right) an episodic knowledge graph. (Left) Every arrow represents a (subject, predicate, object) triple, with the annotation of the arrow denoting the respective predicate. The triple (Ban Ki-moon, SecretaryOf, UN) is deleted, since the knowledge graph has been updated with the triple (António Guterres, SecretaryOf, UN). (Right) Every arrow represents a (subject, predicate, object, timestamp) quadruple, where the arrow is both annotated with the respective predicate and timestamp. Here the quadruple involving is not deleted, since the attached timestamp reveals that the relationship is not valid at present.

An interesting question is how episodic and semantic memories are related. There is evidence that these main cognitive categories are partially dissociated from one another in the brain, as expressed in their differential sensitivity to brain damage. However, there is also evidence indicating that the different memory functions are not mutually independent and support one another [11]. We propose that semantic memory can be derived from episodic memory by marginalization. Hereby we also consider that many episodes describe starting and endpoints of state changes. For example, an individual might become sick

with a disease, which eventually is cured. Similarly, a president’s tenure eventually ends. We study our hypothesis on the Integrated Conflict Early Warning System (ICEWS) dataset, which contains many events with start and end dates. Figure 1 compares semantic and episodic knowledge graphs. Furthermore, Figure 2 illustrates the main ideas of building and modeling semantic and episodic knowledge graphs.

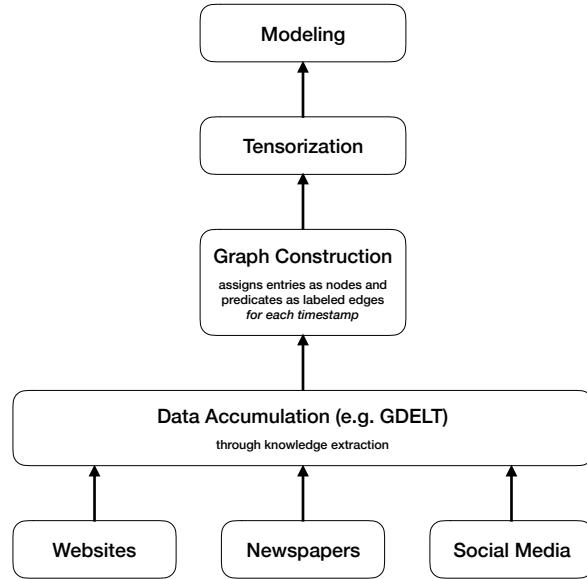


Figure 2: Illustration of the main idea behind the models presented in this paper. **Step 1:** Knowledge is extracted from unstructured data, such as websites, newspapers or social media. **Step 2:** The knowledge graph is constructed, where entities are assigned as nodes, and predicates as labeled edges; note that there is a labeled edge for each timestamp. **Step 3:** The knowledge graph is represented as a tensor; for semantic KGs, we obtain a 3-way tensor, storing (subject, predicate, object) triples, and for episodic KGs, we obtain a 4-way tensor, storing (subject, predicate, object, timestamp) quadruples. **Step 4:** The semantic and episodic tensors are decomposed and modeled via compositional or tensor models (see Section 2).

The paper is organized as follows. Section 2 introduces knowledge graphs,

the mapping of a knowledge graph to an adjacency tensor, and the statistical embedding models for knowledge graphs. We also describe how popular embedding models for KGs can be extended to episodic KGs. Section 3 shows experimental results on modelling episodic KGs. Finally, we present experiments on the possible relationships between episodic and semantic memory in Section 4.

2. Model Descriptions

A static or semantic knowledge graph (KG) is a triple-oriented knowledge representation. Here we consider a slight extension to the subject-predicate-object triple form by adding the value in the form $(e_s, e_p, e_o; \text{Value})$, where *Value* is a function of e_s, e_p, e_o and, e.g., can be a Boolean variable (*True* for 1, *False* for 0) or a real number. Thus $(\text{Jack}, \text{likes}, \text{Mary}; \text{True})$ states that Jack (the subject or head entity) likes Mary (the object or tail entity). Note that e_s and e_o represent the entities for subject index s and object index o . To simplify notation we also consider e_p to be a generalized entity associated with predicate type with index p . For the episodic KGs we introduce e_t , which is a generalized entity for time t .

To model a static KG, we introduce the three-way semantic adjacency tensor χ where the tensor element $x_{s,p,o}$ is the associated *Value* of the triple (e_s, e_p, e_o) . One can also define a companion tensor Θ_χ with the same dimensions as χ and with entries $\theta_{s,p,o}$. Thus, the probabilistic model for the semantic tensor χ is defined as $P(x_{s,p,o}|\theta_{s,p,o}) = \sigma(\theta_{s,p,o})$, where $\sigma(x) = 1/(1 + \exp(-x))$. Similarly, the four-way temporal or episodic tensor \mathcal{E} has elements $x_{t,s,p,o}$ which are the associated values of the quadruples (e_t, e_s, e_p, e_o) , with $t = 1, \dots, T$. Therefore, the probabilistic model for episodic tensor is defined with the corresponding companion tensor $\Theta_{\mathcal{E}}$ as

$$P(x_{t,s,p,o}|\theta_{t,s,p,o}) = \sigma(\theta_{t,s,p,o}) . \quad (1)$$

We assume that each entity e has a unique latent representation \mathbf{a} . In particular, the embedding approach used for modeling semantic and episodic knowledge

graphs assumes that $\theta_{s,p,o}^{sem} = f^{sem}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o})$, and $\theta_{t,s,p,o}^{epi} = f^{epi}(\mathbf{a}_{e_t}, \mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o})$, respectively. Here, the indicator function $f^{sem/epi}(\cdot)$ is a function to be learned.

Given a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, latent representations and other parameters (denoted as \mathcal{P}) are learned by minimizing the regularized logistic loss

$$\min_{\mathcal{P}} \sum_{i=1}^m \log(1 + \exp(-y_i \theta_i^{sem/epi})) + \lambda \|\mathcal{P}\|_2^2. \quad (2)$$

In general, most KGs only contain positive triples; non-existing triples are normally used as negative examples sampled with local closed-world assumption. Alternatively, we can minimize a margin-based ranking loss over the dataset such as

$$\min_{\mathcal{P}} \sum_{i \in \mathcal{D}_+} \sum_{j \in \mathcal{D}_-} \max(0, \gamma + \sigma(\theta_j^{sem/epi}) - \sigma(\theta_i^{sem/epi})), \quad (3)$$

where γ is the margin parameter, and \mathcal{D}_+ and \mathcal{D}_- denote the set of positive and negative samples, respectively.

There are different ways for modeling the indicator function $f^{epi}(\cdot)$ or $f^{sem}(\cdot)$. In this paper, we will only investigate multilinear models derived from tensor decompositions and compositional operations. We now describe the models in detail. Graphical illustrations of the described models are shown in Figure 3.

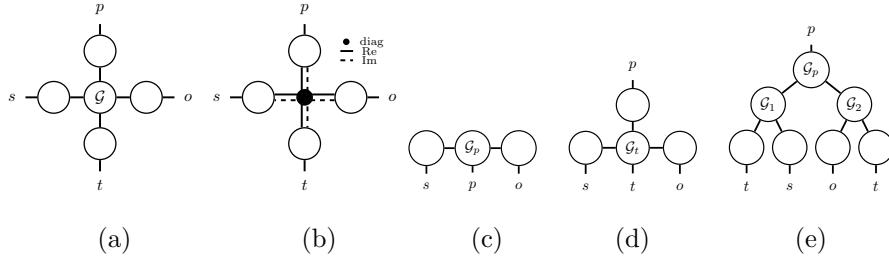


Figure 3: Illustrations of (a) episodic Tucker, (b) episodic ComplEx (where \bullet denotes contraction), (c) RESCAL, (d) ConT and (e) Tree. Each entity in the figure is represented as a circle with two edges, since the representation for an entity e is $\mathbf{a}_{e,i}$. In addition, \mathcal{G} represents the core tensor in Tucker, \mathcal{G}_p represents the matrix latent representation of predicate p in the RESCAL and Tree models, \mathcal{G}_t represents the three-dimensional tensor latent representation of timestamp t in the ConT model.

Table 1 and Table 2 summarize notations used throughout this paper for

easy reference, while Table 3 summarizes the number of parameters required for each model.²

Table 1: Summary of the general notations.

General	
Symbol	Meaning
e_s	Entity for subject index s
e_o	Entity for object index o
e_p	Generalized entity for predicate index p
e_t	Generalized entity for time index t
\mathbf{a}_{e_i}	Latent representation of entity e_i
$\mathbf{a}(e_{t_{start}})$	Latent representation of starting timestamp
a_{e_i, r_i}	r_i -th element of \mathbf{a}_{e_i}
\tilde{r}	Rank/Dimensionality of \mathbf{a}_{e_i} for $i \in \{s, p, o\}$
\tilde{r}_t	Rank/Dimensionality of \mathbf{a}_{e_t}
$N_{e/p/t}$	Number of entities / predicates / timestamps

Tucker. First, we consider the Tucker model for semantic tensor decomposition of the form $\theta_{s,p,o}^{sem} = \sum_{r_1, r_2, r_3=1}^{\tilde{r}} a_{e_s, r_1} a_{e_p, r_2} a_{e_o, r_3} g^{sem}(r_1, r_2, r_3)$. Here, $g^{sem}(r_1, r_2, r_3) \in \mathbb{R}$ are elements of the core tensor $\mathcal{G}^{sem} \in \mathbb{R}^{\tilde{r} \times \tilde{r} \times \tilde{r}}$. Similarly, the indicator function of a four-way Tucker model for episodic tensor decomposition is of the form

$$\begin{aligned} \theta_{t,s,p,o}^{epi} &= \sum_{r_1=1}^{\tilde{r}_t} \sum_{r_2, r_3, r_4=1}^{\tilde{r}} \\ &a_{e_t, r_1} a_{e_s, r_2} a_{e_p, r_3} a_{e_o, r_4} g^{epi}(r_1, r_2, r_3, r_4), \end{aligned} \quad (4)$$

with a four dimensional core tensor $\mathcal{G}^{epi} \in \mathbb{R}^{\tilde{r}_t \times \tilde{r} \times \tilde{r} \times \tilde{r}}$. Note that this is a con-

²For DistMult, ComplEx, and HolE it is required that $\tilde{r} = \tilde{r}_t$. In our experiments (see Sections 3 and 4), in order to enable a fair comparison between the different models, we assume that the latent representations of entities, predicates, and time indices all have the same rank/dimensionality.

Table 2: Summary of the notations for semantic and episodic knowledge graphs.

Semantic knowledge graphs		Episodic knowledge graphs	
Symbol	Meaning	Symbol	Meaning
χ	Sem. adjacency tensor	\mathcal{E}	Epi. adjacency tensor
Θ_χ	Companion tensor of χ	$\Theta_{\mathcal{E}}$	Companion tensor of \mathcal{E}
$x_{s,p,o}$	Value of (e_s, e_p, e_o)	$x_{t,s,p,o}$	Value of (e_t, e_s, e_p, e_o)
$\theta_{s,p,o}^{sem}$	Logit of (e_s, e_p, e_o)	$\theta_{t,s,p,o}^{epi}$	Logit of (e_t, e_s, e_p, e_o)
$f^{sem}(\cdot)$	Sem. indicator function	$f^{epi}(\cdot)$	Epi. indicator function
\mathcal{G}^{sem}	Sem. core tensor	\mathcal{G}^{epi}	Epi. core tensor
$g^{sem}(\cdot)$	Element of \mathcal{G}^{sem}	$g^{epi}(\cdot)$	Element of \mathcal{G}^{epi}

straint Tucker model, since, as in RESCAL, entities have unique representations, independent of the roles as subject or object.

RESCAL. Another model closely related to the semantic Tucker tensor decomposition is the RESCAL model, which has shown excellent performance in modelling KGs [12]. In RESCAL, subjects and objects have vector latent representations, while predicates have matrix latent representations. The indicator function of RESCAL for modeling semantic KGs takes the form $\theta_{s,p,o}^{sem} = \sum_{r_1, r_2=1}^{\tilde{r}} a_{e_s, r_1} g_p(r_1, r_2) a_{e_o, r_2}$, where $g_p(r_1, r_2)$ represents the matrix latent representation for the predicate e_p . Then next two models, Tree and ConT, are novel generalizations of RESCAL to episodic tensors.

Tree. From a practical perspective, training an episodic Tucker tensor model is very expensive since the computational complexity is approximately \tilde{r}^4 . Tensor networks provide a general and flexible framework to design nonstandard tensor decompositions [13, 14]. One of the simplest tensor networks is a tree tensor decomposition (\mathcal{T}) of the episodic indicator function, which is illustrated in compositional operations. We now describe the models in detail. Graphical illustrations of the described models are shown in Figure 3(e). Therefore, we propose a tree tensor decomposition (\mathcal{T}) of the episodic indicator function. The tree \mathcal{T} is partitioned into two subtrees \mathcal{T}_1 and \mathcal{T}_2 , wherein subject e_s and time

e_t reside in \mathcal{T}_1 , while object e_o and an auxiliary time e_t reside in \mathcal{T}_2 . \mathcal{T}_1 and \mathcal{T}_2 are connected with e_p through two core tensors \mathcal{G}_1 and \mathcal{G}_2 . Thus, the indicator function can be written as

$$\theta_{t,s,p,o}^{epi} = \sum_{r_1, r_6=1}^{\tilde{r}_t} \sum_{r_2, r_3, r_4, r_5=1}^{\tilde{r}} a_{e_t, r_1} a_{e_s, r_2} g_1(r_1, r_2, r_3) g_p(r_3, r_4) g_2(r_4, r_5, r_6) a_{e_o, r_5} a_{e_t, r_6}. \quad (5)$$

Within \mathcal{T} , we reduce the four-way core tensor in Tucker into two three-dimensional tensors \mathcal{G}_1 and \mathcal{G}_2 , so that the computational complexity of \mathcal{T} is approximately \tilde{r}^3 .

ConT. ConT is another generalization of the RESCAL model to episodic tensors with reduced computational complexity of approximately \tilde{r}^3 . The idea is that another way of reducing the complexity is by contracting indices of the core tensor. Therefore, we contract the \mathcal{G} from Tucker with the time index giving a three-way core tensor \mathcal{G}_t for each time instance. The indicator function takes the form

$$\theta_{t,s,p,o}^{epi} = \sum_{r_1, r_2, r_3=1}^{\tilde{r}} a_{e_s, r_1} a_{e_p, r_2} a_{e_o, r_3} g_t(r_1, r_2, r_3). \quad (6)$$

In this model, the tensor \mathcal{G}_t resembles the relation-specific matrix \mathcal{G}_p from RESCAL. Later, we will see that ConT is a superior model for modeling episodic knowledge graphs due to the representational flexibility of its high-dimensional tensor \mathcal{G}_t for the time index.

Even though the complexity of Tree and ConT is reduced as compared to episodic Tucker, the three-dimensional core tensor might cause rapid overfitting during training. Therefore, we next propose episodic generalization of compositional models, such as DistMult [15], Hole [16] and ComplEx [17]. For those models, the number of parameters only increases linearly with the rank.

DistMult. DistMult [15] is a simple generalization of the CP model, by enforcing the constraint that entities should have unique representations. Episodic DistMult takes the form $\theta_{t,s,p,o}^{epi} = \sum_{i=1}^{\tilde{r}} \lambda_i a_{e_t, i} a_{e_s, i} a_{e_p, i} a_{e_o, i}$. Here, we require that vector latent representations of entities, predicates, and timestamps have

the same rank. DistMult is a special case of Tucker having a core tensor with only diagonal elements λ_i .

Hole. Holographic embedding (HolE) [16] is a state-of-art link prediction and knowledge graph completion method, which is inspired by holographic models of associative memory.

HolE uses circular correlation to generate a compositional representation from inputs e_s and e_o . The indicator of HolE reads $\theta_{s,p,o}^{sem} = \mathbf{a}_{e_p} \cdot (\mathbf{a}_{e_s} \star \mathbf{a}_{e_o})$, where $\star : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes the circular correlation $[\mathbf{a} \star \mathbf{b}]_k = \sum_{i=0}^{d-1} a_i b_{(k+i) \bmod d}$. We define the episodic extension of HolE as

$$\theta_{t,s,p,o}^{epi} = \mathbf{a}_{e_t} \cdot (\mathbf{a}_{e_p} \star (\mathbf{a}_{e_s} \star \mathbf{a}_{e_o})) . \quad (7)$$

As argued by [16], HolE employs a holographic reduced representation [18] to store and retrieve the predicates from e_s and e_o . Analogously, episodic HolE should be able to retrieve the stored timestamps from e_p , e_s and e_o . In the semantic case, e_p can be retrieved if existing triple relations are stored via circular convolution \star , and superposition in the representation $\mathbf{a}_{e_o} = \sum_{(s,p) \in \mathcal{S}_o} \mathbf{a}_{e_p} \star \mathbf{a}_{e_s}$, where \mathcal{S}_o is the set of all true triples given e_o . This is based on the fact that $\mathbf{a} \star \mathbf{a} \approx \delta$ [16]. Analogously, the stored timestamp e_t for an event can be retrieved if all existing episodic events are stored via \star , and superposition in the representation of e_o , $\mathbf{a}_{e_o} = \sum_{(t,s,p) \in \mathcal{S}_o} \mathbf{a}_{e_t} \star (\mathbf{a}_{e_p} \star \mathbf{a}_{e_s})$, where \mathcal{S}_o is the set of all true quadruples (t, s, p, o) given e_o . However, high order circular correlation/convolution will increase the inaccuracy of retrieval. Another motivation for our episodic extension (7) is that a compositional operator of the form $\mathbf{a}_{e_t} \cdot \tilde{f}$ allows a projection from episodic memory to semantic memory, to be detailed later.

ComplEx. Complex embedding (ComplEx) [17] is another state-of-art method closely related to HolE. It can accurately describe both symmetric and antisymmetric relations. HolE is a special case of ComplEx with imposed conjugate symmetry on embeddings [19]. Thus, ComplEx has more degrees of freedom, if compared to HolE. For the semantic complex embedding, the indicator function is $\theta_{s,p,o}^{sem} = \text{Re} \left(\sum_i^{\tilde{r}} a_{e_s,i} a_{e_p,i}, \bar{a}_{e_o,i} \right)$ with complex valued \mathbf{a} and where

the bar indicates the complex conjugate. To be consistent with the episodic HolE, the episodic complex embedding is defined as³

$$\theta_{t,s,p,o}^{epi} = \text{Re} \left(\sum_i^{\tilde{r}} a_{e_t,i} a_{e_s,i} a_{e_p,i} \bar{a}_{e_o,i} \right). \quad (8)$$

3. Experiments on Episodic Models

We investigate the proposed tensor and compositional models with experiments which are evaluated on two datasets:

ICEWS. The Integrated Conflict Early Warning System (ICEWS) dataset [20] is a natural episodic dataset recording dyadic events between different countries. An example entry could be (*Turkey, Syria, Fight, 12/25/2014*). These dyadic events are aggregated into a four-way tensor \mathcal{E} with 258 entities, 20 relation types, and 72 timestamps, which has in total 320,118 positive (e_t, e_s, e_p, e_o) quadruples⁴. This dataset was first created and used in [21]. From this ICEWS dataset, a semantic tensor is generated by extracting consecutive events that last until the last timestamp, constituting the *current*⁵ semantic facts of the world.

GDELT. The Global Database of Events, Language and Tone (GDELT) [20] monitors the world’s news media in broadcast, print and web formats from all over the world, daily since January 1, 1979⁶. We use GDELT as a large episodic dataset. For our experiments, GDELT data is collected from January 1, 2012 to December 31, 2012 (with a temporal granularity of 24 hrs). These events are aggregated into an episodic tensor \mathcal{E} with 1100 entities, 180 relation

³One can show that Eq. (7) is equivalent to Eq. (8) by converting it to the frequency domain [19]. Then, $\theta_{t,s,p,o}^{epi} \propto \omega_{e_t}^T (\bar{\omega}_{e_p} \odot \bar{\omega}_{e_s} \odot \omega_{e_o})$, where $\omega = \mathcal{F}(\mathbf{a}) \in \mathbb{C}^{\tilde{r}}$ are the discrete Fourier transforms of embeddings \mathbf{a} , and using the fact that ω is conjugate symmetric for real vector \mathbf{a} .

⁴Note that for an episodic event the dataset contains all the quadruples (e_{t_i}, e_s, e_p, e_0) for $t_i \in \{t_{start}, t_{start} + 1, \dots, t_{end} - 1, t_{end}\}$.

⁵*Current* always indicates the last timestamp/timestamps of the applied episodic KGs.

⁶<https://www.gdeltproject.org/about.html>

Table 3: Number of parameters for different models and the runtime of one training epoch on the GDELT dataset.

Model	Semantic	Episodic	Complexity	Runtime			
				rank 40	rank 60	rank 150	
DistMult	$(N_e + N_p + 1)\tilde{r}$	$(N_e + N_p + N_t + 1)\tilde{r}$	$\mathcal{O}(\tilde{r})$	35.2s	36.4s	53.7s	
Hole	$(N_e + N_p)\tilde{r}$	$(N_e + N_p)\tilde{r}$	$\mathcal{O}(\tilde{r} \log \tilde{r})$	42.8s	43.2s	59.0s	
ComplEx	$2(N_e + N_p)\tilde{r}$	$2(N_e + N_p + N_t)\tilde{r}$	$\mathcal{O}(\tilde{r})$	40.1s	42.4s	57.5s	
Tree	—	$N_e\tilde{r} + N_p\tilde{r}^2 + (N_t + 2\tilde{r}^2)\tilde{r}_t$	$\mathcal{O}(\tilde{r}^3)$	133.6s	160.2s	—	
ConT	—	$(N_e + N_p)\tilde{r} + N_t\tilde{r}^3$	$\mathcal{O}(\tilde{r}^3)$	95.4s	226.1s	—	
Tucker	$(N_e + N_p)\tilde{r} + \tilde{r}^3$	$(N_e + N_p)\tilde{r} + (N_t + \tilde{r}^3)\tilde{r}_t$	$\mathcal{O}(\tilde{r}^4)$	144.2s	387.9s	—	

types, and 366 timestamps, which has in total 2,563,561 positive (e_t, e_s, e_p, e_o) quadruples.

We assess the quality of episodic information retrieval on both datasets for the proposed tensor and compositional models. Since both episodic datasets only consist of positive quadruples, we generated negative episodic instances following the protocol of corrupting semantic triples given by Bordes [22]: negative instances of an episodic quadruple (e_s, e_p, e_o, e_t) are drawn by corrupting the object e_o to $e_{o'}$ or the timestamp e_t to $e_{t'}$, meaning that $(e_s, e_p, e_{o'}, e_t)$ serves as a negative evidence of the episodic event at time instance e_t , and $(e_s, e_p, e_o, e_{t'})$ is a true fact which cannot be correctly recalled at time instance $e_{t'}$. During training, for each positive sample in a batch we assigned two negative samples with corrupted object or corrupted subject.

The model performance is evaluated using the following scores. To retrieve the occurrence time, for each true quadruple, we replace the time index e_t with every other possible time index $e_{t'}$, compute the value of the indicator function $\theta_{t',s,p,o}^{epi}$, and rank them in a decreasing order. We filter the ranking as in [22] by removing all quadruples where $x_{t',s,p,o} = 1$ and $t \neq t'$, in order to eliminate ambiguity during episodic information retrieval. Similarly, we evaluated the retrieval of the predicate between a given subject and object at a certain time instance by computing and ranking the indicator $\theta_{t,s,p',o}^{epi}$. We also evaluated the

retrieval of entities by ranking and averaging the filtered indicators $\theta_{t,s',p,o}$ and $\theta_{t,s,p,o'}$. To measure the generalization ability of the models, we report different measures of the ranking: mean reciprocal rank (MRR), and Hits@n on the test dataset.

The datasets were split into train, validation, and test sets that contain the most frequently appearing entities in the episodic knowledge graphs. Training was performed by minimizing the logistic loss (2), and was terminated using early stopping on the validation dataset by monitoring the filtered MRR recall scores every $\{50, 100\}$ epochs depending on the models, where the maximum training duration was 500 epochs. This ensures that the generalization ability of unique latent representations of entities doesn't suffer from overfitting. Before training, all model parameters are initialized using Xavier initialization [23]. We also apply an l_2 norm penalty on all parameters for regularization purposes (see Eq. (2)).

In Table 3 we summarize the runtime for one training epoch on the GDELT dataset for different models at ranks $\tilde{r} = \tilde{r}_t \in \{40, 60, 150\}$. All experiments were performed on a single Tesla K80 GPU. In the following experiments, for compositional models we search rank in $\{100, 150\}$, while for tensor models we search optimal rank in $\{40, 50, 60\}$ since larger ranks could lead to overfitting rapidly. Loss function is minimized with Adam method [24] with the learning rate selected from $\{0.001, 1e-4, 5e-5\}$.

We first assess the filtered MRR, Hits@1, Hits@3, and Hits@10 scores of inferring missing entities and predicates on the GDELT test dataset. Table 4 summarizes the results. Generalizations on the test dataset indicate the inductive reasoning capability of the proposed models. This generalization can be useful for the completion of evolving KGs with missing records, such as clinical datasets. It can be seen that tensor models are able to outperform compositional models consistently on both entity and predicate prediction tasks. ConT has the best inference results on the entity-related tasks, while Tucker performs better on the predicate-related tasks. The superior Hits@1 result of ConT on the entity prediction indicates that there are easily to be fitted entities in the GDELT

Table 4: Filtered results of inferring missing entities and predicates of episodic quadruples evaluated on the GDELT dataset.

Method	Entity				Predicate			
	MRR	@1	@3	@10	MRR	@1	@3	@10
DistMult	0.182	6.55	19.77	43.70	0.269	12.65	30.29	59.40
HolE	0.177	6.67	18.95	41.84	0.256	11.81	28.35	57.73
ComplEx	0.172	6.54	17.52	41.56	0.255	12.05	27.75	56.60
Tree	0.196	8.17	21.00	44.65	0.274	13.30	30.66	60.05
Tucker	0.204	8.93	21.85	46.35	0.275	12.69	31.35	60.70
ConT	0.233	13.85	24.65	42.96	0.263	12.83	29.27	57.30

Table 5: Filtered results for entities and predicates recollection/prediction evaluated on the ICEWS dataset.

Method	Entity				Predicate			
	MRR	@1	@3	@10	MRR	@1	@3	@10
DistMult	0.222	9.72	22.48	52.32	0.520	33.73	62.25	91.13
HolE	0.229	9.85	23.49	54.21	0.517	31.55	65.47	93.59
ComplEx	0.229	8.94	23.53	57.72	0.506	30.99	61.46	93.44
Tree	0.205	10.48	19.84	42.81	0.554	36.62	67.25	94.70
Tucker	0.257	12.88	27.10	54.43	0.563	36.96	69.55	95.43
ConT	0.264	15.71	29.60	46.67	0.557	38.12	67.76	87.71

dataset along the timestamps. In fact, the GDELT dataset is unbalanced, and episodic quadruples related to certain entities dominate in the episodic Knowledge graph, such as quadruples containing the entities *USA*, or *UN*. Experiment results on balanced and extremely sparse episodic dataset will be reported in the following.

Next, Table 5 shows the MRR, Hits@1, Hits@3, and Hits@10 scores of inferring missing entities and predicates on the ICEWS test dataset. Similarly, we can read that tensor models outperform compositional models on both missing entity and predicate inference tasks. The superior Hits@1 result of ConT for the

missing entity prediction indicates again that the ICEWS dataset is unbalanced, and episodic quadruples related to certain entities dominate.

Table 6: Filtered recall scores for entities and timestamps recollection on the ICEWS (rare) training dataset.

Method	Rank	Timestamp		Entity	
		MRR	@3	MRR	@3
DistMult	200	0.257	27.0	0.211	21.9
HolE	200	0.216	20.8	0.179	16.3
ComplEx	200	0.354	40.3	0.301	33.2
Tree	40	0.421	55.3	0.314	35.7
Tucker	40	0.923	98.9	0.893	97.1
ConT	40	0.982	99.7	0.950	97.9

The recollection of the exact occurrence time of a significant past event (e.g. unusual, novel, attached with emotion) is also an important capability of episodic cognitive memory function. In order to manifest this perspective of proposed models, Table 6 shows the filtered MRR, and Hits@3 scores for the timestamps and entities recollection on the episodic ICEWS (rare) training dataset, where rank column registers the optimal and minimum rank $\tilde{r} = \tilde{r}_t$ having the outstanding recall scores. Figure 4 further displays the filtered MRR score as a function of rank. Unlike the original ICEWS, which contains many consecutive events that last from the first to the last timestamp leading to unreasonably high filtered timestamp recall scores, this ICEWS (rare) dataset consists of rare temporal events that happen less than three times throughout the whole time and starting points of events.

The outstanding performance of ConT compared with other compositional models indicates the importance of large dimensionality of time latent representation for the episodic tensor reconstruction / episodic memory recollection. Recall that for ConT the real *dimension* of the latent representation of time is actually \tilde{r}^3 after flattening \mathcal{G}_t . This flexible latent representation for time could

compress almost all the semantic triples that occur at a certain instance ⁷.

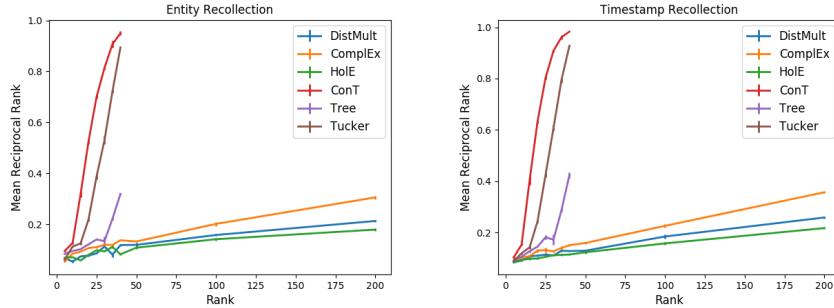


Figure 4: Filtered MRR scores vs. rank for the entities (left) and timestamps (right) recollection on the ICEWS (rare) training dataset.

4. Semantic Memory from Episodic Memory with Marginalization

We already discussed that a semantic KG might be related to a human semantic memory and that an episodic KG might be related to a human episodic memory. It has been speculated that episodic and semantic memory must be closely related, and that semantic memory is generated from episodic memory by some training process [28, 29]. As a very simple implementation of that idea, we propose that a semantic memory could be generated from episodic memory by marginalizing time. Thus, both types of memories would rely on identical representations and the marginalization step can be easily performed: Since probabilistic tensor models belong to the classes of sum-product nets, a marginalization simply means an integration over all time representations.

Thus, in the second set of experiments, we test the hypothesis that semantic

⁷This observation has its biological counterpart. In fact, the entorhinal cortex, which plays an important role in the formation of episodic memory, is the main part of the adult hippocampus that shows neurogenesis [25]. In an adult human, approximately 700 new neurons are added per day through hippocampal neurogenesis, which are believed to perform sensory and spatial information encoding, as well as temporal separation of events [26, 27].

memory can be derived from episodic memory by projection. In other words, a semantic knowledge graph containing *current* semantic facts can be approximately constructed after modeling a corresponding episodic knowledge graph via marginalization. A marginalization can be performed by activating all time index neurons, i.e., summing over all \mathbf{a}_{e_t} , since, e.g., Tucker decompositions are an instance of a so-called sum-product network [30]. However, events having start as well as end timestamps cannot simply be integrated into our *current* semantic knowledge describing what we *know* now. For example, (Ban Ki-moon, SecretaryOf, UN) is not consistent with what we *know* currently. To resolve this problem, we introduce two types of time indices, $e_{t_{start}}$ and $e_{t_{end}}$, having the latent representations $\mathbf{a}(e_{t_{start}})$ and $\mathbf{a}(e_{t_{end}})$, respectively. Those time indices can be used to construct the episodic tensor \mathcal{E}_{start} aggregating the start timestamps of consecutive events, as well as the episodic tensor \mathcal{E}_{end} aggregating the end timestamps⁸.

For the projection, instead of only summing over $\mathbf{a}(e_{t_{start}})$, we also subtract the sum over $\mathbf{a}(e_{t_{end}})$. In this way, we can achieve the effect that events that have terminated already (i.e., have an end time index smaller than the current time index) are not integrated into the current semantic facts. Now, to test our hypothesis that this extended projection allows us to derive semantic memory from episodic memory, we trained Hole, DistMult, ComplEx, ConT, and Tucker on the episodic tensors \mathcal{E}_{start} and \mathcal{E}_{end} as well as on the semantic tensor χ derived from ICEWS. Note that only these models allow projection, since their indicator functions can be written in the form $\theta_{t,s,p,o}^{epi} = \mathbf{a}_{e_t} \cdot \tilde{f}$, where \tilde{f} can be arbitrary function of \mathbf{a}_{e_s} , \mathbf{a}_{e_p} , and \mathbf{a}_{e_o} depending on the model choice⁹. The

⁸E.g., if the duration of a triple event (e_s, e_p, e_o) lasts from t_{start} to t_{end} , the quadruple $(e_s, e_p, e_o, e_{t_{start}})$ is stored in \mathcal{E}_{start} , while $(e_s, e_p, e_o, e_{t_{end}})$ is stored \mathcal{E}_{end} only if $t_{end} < T$ (where T is the last timestamp). In other words, events that last until the last timestamp do not possess e_{end} .

⁹For ConT, $\theta_{t,s,p,o}^{epi} = flatten(g_t) \cdot (\mathbf{a}_{e_s} \otimes \mathbf{a}_{e_p} \otimes \mathbf{a}_{e_o})$, where \otimes denotes the outer product. For ComplEx, $\theta_{t,s,p,o}^{epi} = \text{Re}(\mathbf{a}_{e_t}) \cdot \text{Re}(\mathbf{a}_{e_s} \odot \mathbf{a}_{e_p} \odot \bar{\mathbf{a}}_{e_o}) - \text{Im}(\mathbf{a}_{e_t}) \cdot \text{Im}(\mathbf{a}_{e_s} \odot \mathbf{a}_{e_p} \odot \bar{\mathbf{a}}_{e_o})$, where \odot denotes the Hadamard product. The Tree model cannot be written in this form since e_t

model parameters are optimized using the margin-based ranking loss (3)¹⁰.

Training was first performed on the episodic tensor \mathcal{E}_{start} , and then on \mathcal{E}_{end} with *fixed* \mathbf{a}_{e_s} , \mathbf{a}_{e_p} , and \mathbf{a}_{e_o} obtained from the training on \mathcal{E}_{start} , since we assume that latent representations for subject, object, and predicate of a consecutive event do not change during the event. Note that after training in this way, we could recall the starting and terminal point of a consecutive event (see the episodic tensor reconstruction experiments in Section 3), or infer a *current* semantic fact solely from the latent representations instead of rule-based reasoning.

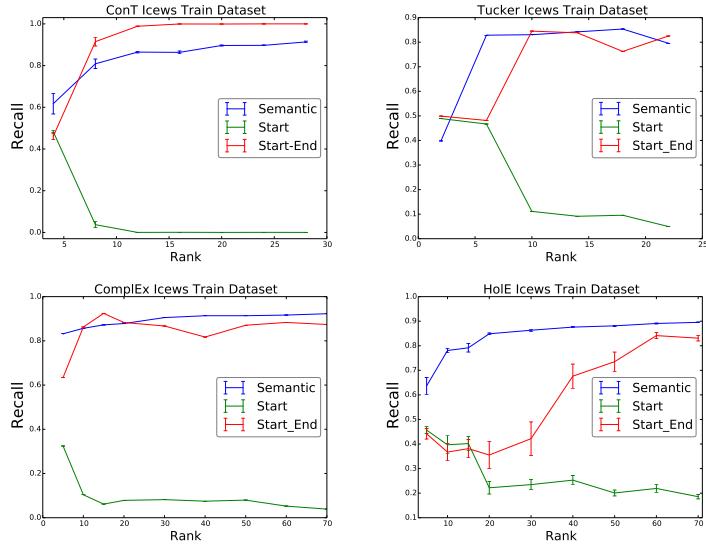


Figure 5: Recall scores vs. rank for the episodic-to-semantic projection on the ICEWS dataset with two different projection methods.

To evaluate the projection, we compute the recall and area under precision-recall-curve (AUPRC) scores for the projection at different ranks on the ICEWS

resides in both subtrees \mathcal{T}_1 and \mathcal{T}_2 .

¹⁰For the projection experiment, we omit the sigmoid function in Eq. (3), train and interpret the multilinear indicator $\theta_{t,s,p,o}^{epi} = \mathbf{a}_{e_t} \cdot \tilde{f}(\mathbf{a}_{e_s}, \mathbf{a}_{e_p}, \mathbf{a}_{e_o})$ directly as the probability of episodic quadruple. Only in this way of training, a projection is mathematically legitimate.

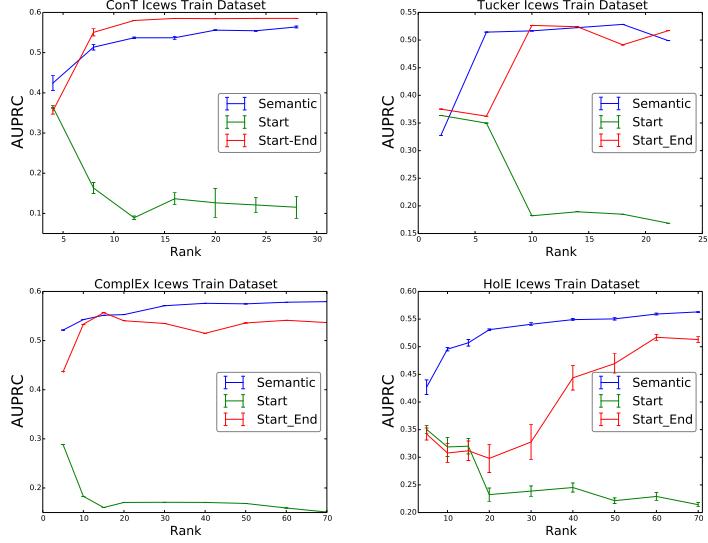


Figure 6: AUPRC scores vs. rank for the episodic-to-semantic projection on the ICEWS dataset with two different projection methods.

training dataset, and compare them with the scores obtained from training the semantic tensor separately. The semantic dataset contains positive triples, which are episodic events that continue until the last (current) timestamp, e.g. (António Guterres, SecretaryOf, UN, *True*), along with negative triples extracted from already terminated episodic events, e.g. (Ban Ki-moon, SecretaryOf, UN, *False*). During the test phase of projection, a triple from the semantic dataset is given with non-specified time index, e.g. ($e_s, e_p, e_o, \text{True/False}, t$). Then, for the first method considering only the starting point of an episodic event, the projection to semantic space is computed as

$$\theta_{s,p,o}^{proj} = \left[\sum_{t_{start}=1}^T \mathbf{a}(e_{t_{start}}) \right] \cdot \tilde{f}, \quad (9)$$

while for the second method considering both starting and terminal points, the projection is computed as

$$\theta_{s,p,o}^{proj} = \left[\sum_{t_{start}=1}^T \mathbf{a}(e_{t_{start}}) - \sum_{t_{end}=1}^T \mathbf{a}(e_{t_{end}}) \right] \cdot \tilde{f}. \quad (10)$$

Then, the scores are evaluated by taking the label of the given semantic triple as the target, and taking $\theta_{s,p,o}^{proj}$ as the prediction. The goal of this test is to check how well the algorithms can project a given consecutive event $(e_s, e_p, e_o, t_{start} \cdots t_{end})$ to semantic knowledge space using only the marginalized latent representation of time. All other experimental settings are similar to those in Section 3, and the experiments were repeated four times on different sampled training datasets.

Figure 5 shows the recall scores for the two different projection methods on the training dataset in comparison to the separately trained semantic dataset. Due to limited space, we only show four models: ConT, Tucker, ComplEx, and HolE. As we can see, only the marginalization considering both starting and terminal time indices allows a reasonable projection from episodic memory to the *current* semantic memory. Again, ConT¹¹ exhibits the best performance, with its recall score saturating after $\tilde{r} \approx 15$. In contrast, HolE shows insufficient projection quality with sizable errors, especially at small ranks, which is due to its higher-order encoding noise. To show that the two types of latent representations of time do not simply eliminate each other for a correct episodic projection, Figure 6 shows the AUPRC scores evaluated on the training dataset. Overall, this experiment supports the idea that semantic memory is a long-term storage for episodic memory, where the exact timing information is lost.

For a fair comparison, in the last experiment we report the recall scores of the semantic models obtained by projecting the episodic models with respect to the temporal dimension. We compare two projection methods, the Start projection which only considers the staring point of episodic events (see Eq. 9), and the Start-End projection which takes both the starting and terminal points of episodic events into consideration. In addition, we report the recall scores on two semantic datasets. The first one contains genuine semantic facts, while the second dataset contains false semantic triples which should already be ruled out

¹¹Note that since ConT doesn't have a direct semantic counterpart, we instead use the semantic results obtained using RESCAL. This is reasonable since ConT can be viewed as a high-dimensional (i.e., episodic) generalization of RESCAL.

Table 7: Filtered and raw Hits@10 scores for the episodic-to-semantic projection. Two projection methods, Start (Eq. 9), Start-End (Eq. 10), are compared. Furthermore, semantic ICEWS dataset with genuine semantic triples, and semantic ICEWS dataset with false triples are used for the projection experiments. Various projection scores are compared with the scores which are obtained by directly modeling the semantic ICEWS dataset with genuine semantic triples.

Method	Start		Start-End		Start (false)		Start-End (false)		Semantic	
	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw	Filter	Raw
DistMult	3.8	3.6	5.6	5.0	4.0	3.8	3.8	3.6	59.3	32.4
HolE	5.8	5.4	5.5	5.1	4.7	4.5	5.6	5.2	56.1	31.3
ComplEx	4.1	3.7	4.9	4.4	3.9	3.7	3.8	3.6	60.1	29.4
Tucker	14.8	13.1	15.1	13.4	11.3	10.3	11.8	10.9	46.5	23.7
ConT	30.9	24.6	40.8	30.3	23.0	19.9	22.6	19.3	43.8	20.4

through the projection.

Two different projections are performed on two semantic datasets, the genuine one and the false one. Theoretically, the recall scores on the genuine semantic dataset should be higher than those on the false dataset. Thus, the model hyper-parameters are chosen by monitoring the difference between the recall scores Hits@10 on the genuine and false semantic datasets.

Table. 7 reports the filtered and raw Hits@10 metrics for different models, projection methods, and datasets. Moreover, we also compare the projection with the recall scores obtained by directly modeling the genuine semantic dataset using the corresponding semantic models ¹². The ConT model has the best projection performance, since its projected recall scores on the genuine dataset are much higher than those obtained on the false semantic dataset. Moreover, the Start-End projection method based on the ConT model is the only combination which achieves similar results compared to the corresponding semantic model. One can also notice that all the projected compositional models are only able to tell whether a semantic triple is already ruled out or not before the last

¹²Note that we use the RESCAL model as the corresponding semantic model for the ConT.

timestamp, however they can not provide good inference results on the genuine semantic dataset.

5. Conclusion

This paper described the first mathematical models for the declarative memories: the semantic and episodic memory functions. To model these cognitive functions, we generalized leading approaches for static knowledge graphs (i.e., Tucker, RESCAL, HolE, ComplEx, DistMult) to 4-dimensional temporal/episodic knowledge graphs. In addition, we developed two novel generalizations of RESCAL to episodic tensors, i.e., Tree and ConT. In particular, ConT has superior performance overall, which indicates the importance of introduced high-dimensional latent representation of time for both sparse episodic tensor reconstruction and generalization.

Our hypothesis is that perception includes an active semantic decoding process, which relies on latent representations of entities and predicates, and that episodic and semantic memories depend on the same decoding process. We argue that temporal knowledge graph embeddings might be models for human cognitive episodic memory and that semantic memory (facts we know) can be generated from episodic memory by a marginalization operation. We also test this hypothesis on the ICEWS dataset, the experiments show that the *current* semantic facts can only be derived from the episodic tensor by a proper projection considering both starting and terminal points of consecutive events.

Acknowledgements. This work is funded by the *Cognitive Deep Learning* research project in Siemens AG.

References

References

- [1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, *The semantic web* (2007) 722–735.
- [2] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: *Proceedings of the 16th international conference on World Wide Web*, ACM, 2007, pp. 697–706.
- [3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, AcM, 2008, pp. 1247–1250.
- [4] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* 57 (10) (2014) 78–85.
- [5] A. Singhal, Introducing the knowledge graph: things, not strings, *Official google blog*.
- [6] M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A review of relational machine learning for knowledge graphs, *Proceedings of the IEEE*.
- [7] H. Ebbinghaus, *Über das gedächtnis: untersuchungen zur experimentellen psychologie*, Duncker & Humblot, 1885.
- [8] R. C. Atkinson, R. M. Shiffrin, Human memory: A proposed system and its control processes, *Psychology of learning and motivation* 2 (1968) 89–195.
- [9] L. R. Squire, Memory and brain.
- [10] E. Tulving, Episodic and semantic memory: Where should we go from here?, *Behavioral and Brain Sciences* 9 (03) (1986) 573–577.

- [11] D. L. Greenberg, M. Verfaellie, Interdependence of episodic and semantic memory: evidence from neuropsychology, *Journal of the International Neuropsychological society* 16 (05) (2010) 748–753.
- [12] M. Nickel, V. Tresp, H.-P. Kriegel, A three-way model for collective learning on multi-relational data, in: *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 809–816.
- [13] A. Cichocki, Era of big data processing: A new approach via tensor networks and tensor decompositions, in: *International Workshop on Smart Info-Media Systems in Asia (SISA-2013)*, 2013.
- [14] A. Cichocki, Tensor networks for big data analytic and large-scale optimization problems, in: *Second Int. Conference on Engineering and Computational Schematics (ECM2013)*, 2013.
- [15] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, *International Conference on Learning Representations (ICLR)*.
- [16] M. Nickel, L. Rosasco, T. Poggio, Holographic embeddings of knowledge graphs, in: *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [17] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, in: *International Conference on Machine Learning*, 2016, pp. 2071–2080.
- [18] T. A. Plate, Holographic reduced representations, *IEEE Transactions on Neural Networks* 6 (3) (1995) 623–641.
- [19] K. Hayashi, M. Shimbo, On the equivalence of holographic and complex embeddings for link prediction, *CoRR* abs/1702.05563.
URL <http://arxiv.org/abs/1702.05563>
- [20] M. D. Ward, A. Beger, J. Cutler, M. Dickenson, C. Dorff, B. Radford, Comparing gdelt and icews event data, *Analysis* 21 (2013) 267–297.

- [21] A. Schein, J. Paisley, D. M. Blei, H. Wallach, Bayesian poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 1045–1054.
- [22] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Advances in neural information processing systems, 2013, pp. 2787–2795.
- [23] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks., in: Aistats, Vol. 9, 2010, pp. 249–256.
- [24] D. Kingma, J. Ba, Adam: A method for stochastic optimization, Proceedings of the 3rd International Conference on Learning Representations (ICLR).
- [25] W. Deng, J. B. Aimone, F. H. Gage, New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory?, *Nature reviews. Neuroscience* 11 (5) (2010) 339.
- [26] O. Lazarov, C. Hollands, Hippocampal neurogenesis: learning to remember, *Progress in neurobiology* 138 (2016) 1–18.
- [27] K. L. Spalding, O. Bergmann, K. Alkass, S. Bernard, M. Salehpour, H. B. Huttner, E. Boström, I. Westerlund, C. Vial, B. A. Buchholz, et al., Dynamics of hippocampal neurogenesis in adult humans, *Cell* 153 (6) (2013) 1219–1227.
- [28] J. L. McClelland, B. L. McNaughton, R. C. O’reilly, Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory., *Psychological review* 102 (3) (1995) 419.
- [29] L. Nadel, A. Samsonovich, L. Ryan, M. Moscovitch, Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results, *Hippocampus* 10 (4) (2000) 352–368.

- [30] H. Poon, P. Domingos, Sum-product networks: A new deep architecture, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, IEEE, 2011, pp. 689–690.