WARWICK
THE UNIVERSITY OF WARWICK

Manuscript version: Author's Accepted Manuscript
The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:
http://wrap.warwick.ac.uk/158111

How to cite:
Please refer to published version for the most recent bibliographic citation information.

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

warwick.ac.uk/lib-publications

Extracting Event Temporal Relations via Hyperbolic Geometry

Xingwei Tan[1], **Gabriele Pergola**[1], **Yulan He**[1 2]
[1]Department of Computer Science, University of Warwick, UK
[2]Alan Turing Institute, UK
{Xingwei.Tan, Gabriele.Pergola, Yulan.He}@warwick.ac.uk

Abstract

Detecting events and their evolution through time is a crucial task in natural language understanding. Recent neural approaches to event temporal relation extraction typically map events to embeddings in the Euclidean space and train a classifier to detect temporal relations between event pairs. However, embeddings in the Euclidean space cannot capture richer asymmetric relations such as event temporal relations. We thus propose to embed events into hyperbolic spaces, which are intrinsically oriented at modeling hierarchical structures. We introduce two approaches to encode events and their temporal relations in hyperbolic spaces. One approach leverages hyperbolic embeddings to directly infer event relations through simple geometrical operations. In the second one, we devise an end-to-end architecture composed of hyperbolic neural units tailored for the temporal relation extraction task. Thorough experimental assessments on widely used datasets have shown the benefits of revisiting the tasks on a different geometrical space, resulting in state-of-the-art performance on several standard metrics. Finally, the ablation study and several qualitative analyses highlighted the rich event semantics implicitly encoded into hyperbolic spaces.[1]

1 Introduction

Successful understanding of natural language depends, among other factors, on the capability to accurately detect events and their evolution through time. This has recently led to increasing interest in research for temporal relation extraction (Chambers et al., 2014; Wang et al., 2020) with the aim of understanding events and their temporal orders. Temporal reasoning has been proven beneficial, for example, in understanding narratives (Cheng et al., 2013), answering questions (Ning et al., 2020), or summarizing events (Wang et al., 2018).

[1]Source code is available at https://github.com/Xingwei-Warwick/hyper-event-TempRel.

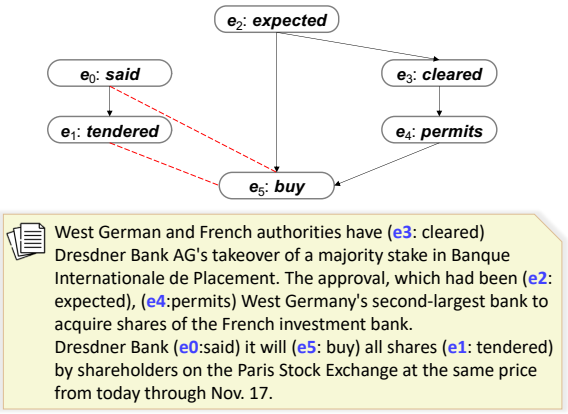

Figure 1: Events annotated with temporal relations from a document excerpt. Arrow lines represent the *Before* relations, while red dashed lines the *Vague* ones.

However, events that occurred in text are not just simple and standalone predicates, they rather form complex and hierarchical structures with different granularity levels (Fig. 1), a characteristic that still challenges existing models and restricts their performance on real-world datasets for temporal relation extraction (Ning et al., 2018a,c). Addressing such challenges requires models to not only recognize accurately the events and their hierarchical and chronological properties but also encode them in appropriate representations enabling effective temporal reasoning. Although this has prompted the recent development of neural architectures for automatic feature extraction (Ning et al., 2019; Wang et al., 2020; Han et al., 2019b), which achieves better generalization and avoids costly design of statistical methods leveraging hand-crafted features (Mani et al., 2006; Chambers et al., 2007; Verhagen and Pustejovsky, 2008), the inherent complexity of temporal relations still hinders approaches that just rely on the scarce availability of annotated data.

Some of the intrinsic limitations of the mentioned approaches are due to the adopted embedding space. Existing approaches to temporal relation extraction typically operate in the Euclidean

space, in which an event embedding is represented as a point. Although such embeddings exhibit a linear algebraic structure which captures co-occurrence patterns among events, they are not able to reveal richer asymmetric relations, such as event temporal order (e.g., '*event A happens before event B*' but not vice versa). Inspired by recent works on learning non-Euclidean embeddings, such as Poincarè embeddings (Nickel and Kiela, 2017; Tifrea et al., 2019) showing superior performance in capturing asymmetrical relations of objects, we propose to learn event embeddings in hyperbolic spaces (Ganea et al., 2018b).

Hyperbolic spaces can be viewed as continuous versions of trees, thus naturally oriented to encode hierarchical and asymmetrical structures. For instance, Sala et al. (2018) showed that hyperbolic spaces with just two dimensions (i.e., Poincaré disk) could easily embed tree structures with arbitrarily low distortion (Sarkar, 2011), while Euclidean spaces cannot achieve any comparable distortion even with an unbounded number of dimensions (Linial et al., 1994). Despite the hierarchical properties arising in modeling event relations, there are still very few studies on how to leverage those models for temporal event extraction. We propose a framework for temporal relation (TempRel) extraction based on the Poincaré ball model, a hyperbolic space well-suited for the efficient computation of embeddings based on the Riemannian optimization.

Our contributions can be summarized as follows:

- We propose an embedding learning approach with a novel angular loss to encode events onto hyperbolic spaces, which pairs with a simple rule-based classifier to detect event temporal relations. With only 1.5k parameters and trained in about 4 minutes, it achieves results on-par with far more complex models in the recent literature.
- Apart from directly learning hyperbolic event embeddings for TempRel extraction, we propose an alternative end-to-end hyperbolic neural architecture to model events and their temporal relations in hyperbolic spaces.
- We conduct a thorough experimental assessment on widely used datasets, with ablation studies and qualitative analyses demonstrating the benefits of tackling the TempRel extraction task in hyperbolic spaces.

2 Related Work

Our work is related to at least two lines of research: one about event temporal relation extraction and another on hyperbolic neural models.

Event TempRel Extraction Approaches to TempRel extraction are largely built on neural models in recent years. These models have been proven capable of extracting automatically reliable event features for TempRel extraction when provided with high-quality data (Ning et al., 2019), alleviating significantly the required human-engineer effort and yielding results outperforming the above mentioned methodologies. In particular, Ning et al. (2019) employed an LSTM network (Hochreiter and Schmidhuber, 1997) to encode the textual events, taking into account their global context and feeding their representations into a multi-layer perceptron for TempRel classification. In addition, to enhance the generalization to unseen event-tuples, they simultaneously trained a Siamese network bridging common-sense knowledge across event relations. Similarly, Han et al. (2019a) combined a bidirectional LSTM (BiLSTM) with a structured support vector machine (SSVM), with the BiLSTM extracting the pair of events and the SSVM incorporating structural linguistic constraints across them[2]. Wang et al. (2020) proposed a constrained learning framework, where event pairs are encoded via a BiLSTM, enhanced with common-sense knowledge from ConceptNet (Speer et al., 2017) and TEMPROB (Ning et al., 2018b), while enforcing a set of logical constraints at training time. The aim is to train the model to detect and extract the event relations while regularizing towards consistency on logic converted into differentiable objective functions, similarly to what was proposed in Li et al. (2019).

Hyperbolic Neural Models The aforementioned models are all designed to process data representations in the Euclidean space. However, several studies (Nickel et al., 2014; Bouchard et al., 2015) have shown the inherent limitations of the Euclidean space in terms of representing asymmetric relations and tree-like graphs (Nickel et al., 2014; Bouchard et al., 2015). Hyperbolic spaces, instead, are promising alternatives that have a natural hierarchical structure and can be thought of as continuous versions of trees. This makes them

[2]They compute the metrics differently and use an old version of MATRES. Thus, their results are not directly comparable

highly suitable and efficient to encode tree-like networks (Nickel and Kiela, 2017; Tran et al., 2020).

Previous works have explored their use in embedding taxonomy for network link prediction or modeling lexical entailment. In particular, Nickel and Kiela (2017) proposed to learn word hierarchies through a negative-sampling training, based on the distance metric on the Poincaré ball. Ganea et al. (2018a) generalized the idea of order embeddings (Vendrov et al., 2016) to the Poincaré ball, levering the projected areas to infer data relations. Ganea et al. (2018b) introduced a framework of hyperbolic neural networks composed of neural units learning and optimizing parameters in hyperbolic spaces. Their experiments show that, without increasing the number of parameters of the models, hyperbolic neural networks outperform their Euclidean counterparts on natural language inference and detection of noisy prefixes tasks. There have been a few attempts in revisiting NLP tasks in the hyperbolic space framework by generalizing hyperbolic neural activation functions for machine translation (Gulcehre et al., 2018), to detect hierarchical entity types (López and Strube, 2020), and for document classification (Zhang and Gao, 2020). Compared to the above works, our model is the first attempt in devising an end-to-end hyperbolic architecture showing the benefit of addressing event TempRel extraction in hyperbolic spaces.

3 Preliminaries

In this section, we give a brief introduction of hyperbolic geometry and hyperbolic neural networks.

Hyperbolic geometry. A hyperbolic space is a non-Euclidean space that has the same negative sectional curvature at every point (i.e., a constant negative curvature). Intuitively, that a space has constant curvature implies that it keeps the same "curveness" at every point. An example of constant positive curvature space is a perfect globe. On the other hand, an example of an n-D hyperbolic space is a hyperboloid in a $\mathbb{R}^{n+1}$ space.

One of the widely-used models of hyperbolic space is the Poincaré model, which is an open n-dimensional unit ball $\mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ equipped with the Riemannian metric tensor:

$$g_x^{\mathbb{D}} = (\lambda_x^2)^2 g^E, \text{ where } \lambda_x := \frac{2}{1 - \|x\|^2}, \quad (1)$$

$x \in \mathbb{D}^n$, $\|\cdot\|$ denotes the Euclidean norm, and g^E denotes the Euclidean metric tensor. The geodesics (i.e., the shortest path between two points) on the Poincaré ball are all intersections of circles with the unit ball $\mathbb{D}^n$ perpendicular to the boundary sphere. Based on the metric tensor $g_x^{\mathbb{D}}$, the distance of two points $x, y \in \mathbb{D}^n$ is defined as (Nickel and Kiela, 2017):

$$d_{\mathbb{D}}(x, y) = \operatorname{arcosh}(1 + 2\frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}). \quad (2)$$

The Poincaré norm is the distance between the origin and the given point:

$$\|x\|_{\mathbb{D}} = d_{\mathbb{D}}(0, x) = 2 \arctan(\|x\|). \quad (3)$$

Also, an angle $\angle ABC$ in the Poincaré model can be derived as (Ganea et al., 2018a):

$$\cos(\angle(z_1, z_2)) = \frac{g_x^{\mathbb{D}}(z_1, z_2)}{\sqrt{g_x^{\mathbb{D}}(z_1, z_2)}\sqrt{g_x^{\mathbb{D}}(z_1, z_2)}}, \quad (4)$$

where $z_1, z_2 \in T_x\mathbb{D}^n \setminus \{0\}$ are the initial tangent vectors of the geodesics connecting B with A, and B with C. An important property of Poincaré model is its conformality with a Euclidean space, which means their metrics define the same angles (Eq. 4$= \frac{\langle z_1, z_2 \rangle}{\|z_1\|\|z_2\|}$, where $\langle \cdot, \cdot \rangle$ denotes Euclidean inner product).

It is known (Ganea et al., 2018a) that an exponential map can be defined for each point $x \in \mathbb{D}^n$ to map any point $z \in \mathbb{R}^n (= T_x\mathbb{D}^n)$ onto $\mathbb{D}^n$:

$$\exp_x(z) = \frac{\frac{1}{\|z\|}\sinh(\lambda_x\|z\|)}{1 + (\lambda_x - 1)\cosh(\lambda_x\|z\|) + \lambda_x\langle x, \frac{z}{\|z\|}\rangle\sinh(\lambda_x\|z\|)} z + \frac{\lambda_x\big(\cosh(\lambda_x\|z\|) + \langle x, \frac{z}{\|z\|}\rangle\sinh(\lambda_x\|z\|)\big)}{1 + (\lambda_x - 1)\cosh(\lambda_x\|z\|) + \lambda_x\langle x, \frac{z}{\|z\|}\rangle\sinh(\lambda_x\|z\|)} x. \quad (5)$$

Hyperbolic neural networks. In order to provide an algebraic setting for a hyperbolic space, which is not a vector space, Ganea et al. (2018b) combine the formalism of Möbius gyrovector spaces with the Riemannian geometry of the Poincaré model. In particular, they replace the operations used in Euclidean multinominal logistic regression (MLR), Feed-Forward (FFNN) and Recurrent Neural Networks (RNN) with Möbius operations, which leads to the hyperbolic version of neural networks. The followings are the definition of the Hyperbolic Gated Recurrent Unit (HGRU):

$$
\begin{aligned}
z_t =& \sigma \log_0(((\mathbf{W}_z \otimes h_{t-1}) \oplus (\mathbf{U}_z \otimes x_t)) \oplus \mathbf{b}_z), \\
r_t =& \sigma \log_0(((\mathbf{W}_r \otimes h_{t-1}) \oplus (\mathbf{U}_r \otimes x_t)) \oplus \mathbf{b}_r), \\
\tilde{h}_t =& \varphi^{\otimes}((([\mathbf{W}_h \operatorname{diag}(r_t)] \otimes h_{t-1}) \oplus (\mathbf{U}_h \otimes x_t)) \oplus \mathbf{b}_h), \\
h_t =& h_{t-1} \oplus (\operatorname{diag}(z_t) \otimes ((-h_{t-1}) \oplus \tilde{h}_t)),
\end{aligned}
$$

where $\otimes$ is the Möbius product, $\oplus$ is the Möbius addition, $\varphi^{\otimes}$ is a hyperbolic non-linearity, $\operatorname{diag}(x)$ is the square diagonal matrix of x, and $\{\mathbf{W}_z, \mathbf{U}_z, \mathbf{b}_z, \mathbf{W}_r, \mathbf{U}_r, \mathbf{b}_r, \mathbf{W}_h, \mathbf{U}_h, \mathbf{b}_h\}$ is the parameter set.

In Hyperbolic MLR, the prediction probability of a given class $k \in 1, ..., K$ is computed as:

$$
\begin{aligned}
p(y = k|x) \propto \exp\big(\operatorname{sign}(\langle -\mathbf{p}_k \oplus x, \mathbf{a}_k \rangle) \\
\sqrt{g^{\mathbb{D}}_{\mathbf{p}_k}(\mathbf{a}_k, \mathbf{a}_k)} d_{\mathbb{D}}(x, \tilde{H}_{\mathbf{a}_k, \mathbf{p}_k})\big), \quad (6)
\end{aligned}
$$

where $x \in \mathbb{D}^n$ is the output vector of the previous layer, $p_k \in \mathbb{D}^n$ and $a_k \in T_{p_k}\mathbb{D}^n \setminus \{0\}$ are parameters.

4 Event Temporal Relation Extraction in the Hyperbolic Space

In this section, we propose two approaches to leverage a hyperbolic space for TempRel extraction. The first approach learns event embeddings that encode temporal order via a hyperbolic space; while the second one is an end-to-end hyperbolic neural network tailored for the TempRel extraction task.

4.1 Hyperbolic Event Embedding Learning

We first explore how to learn embeddings of events in a hyperbolic space while preserving their temporal orders. Temporal relations are asymmetric and transitive, exhibiting similar properties to hierarchical relations. Inspired by previous successes of hyperbolic embeddings for word hierarchies (Nickel and Kiela, 2017; Ganea et al., 2018a), we propose to learn event embeddings based on the Poincaré model to capture their temporal relations.

For a given text sequence containing an event pair (u, v), we first extract the contextualized embeddings of the event tokens[3], e_u and e_v, from their, for example, ELMo (Peters et al., 2018) or RoBERTa (Liu et al., 2019) sequence encodings, and use the exponential mapping function to map them onto a Poincaré ball. The embeddings are then further projected to a lower-dimensional space through a hyperbolic feed-forward layer:

$$
s_u = \text{HFFL}\big(\exp_0(e_u)\big) \quad s_v = \text{HFFL}\big(\exp_0(e_v)\big) \quad (7)
$$

where $\exp_0(\cdot)$ is the exponential map at the origin of a Poincaré ball as defined by Eq. (5), HFFL is a hyperbolic feed-forward layer, s_u and s_v are the final Poincaré embeddings of event u and v.

To encode temporal connections in the embeddings, we want to pull events that have temporal connections close to each other, while pushing events that have no temporal relations far apart. Thus, inspired by Poincaré embeddings (Nickel and Kiela, 2017), we define the first loss term:

$$
\mathcal{L}_1 = \sum_{(u,v) \in D} \log \frac{e^{-d_{\mathbb{D}}(s_u, s_v)}}{\sum_{v' \in N(u)} e^{-d_{\mathbb{D}}(s_u, s_{v'})}}, \quad (8)
$$

where D is the set of event pairs that have temporal connections, $N(u)$ is the set of events that have no temporal relations with the event u; s_u and s_v denote the Poincaré embedding of event u and v, respectively. For example, in the MATRES dataset, event pairs are annotated with one of the following four relations: BEFORE, AFTER, EQUAL and VAGUE. We can consider the first three relations as temporal connections, and regard VAGUE as *no relation* (Ning et al., 2019). Therefore, the set D contains event pairs in the training set that have BEFORE, AFTER, or EQUAL labels. The set $N(u)$ contains the events that are in the same documents with u, but cannot reach u using only BEFORE, AFTER, or EQUAL edge. It is worth noting that because we do not encode the relation type explicitly, the order of input event pair (u, v) matters. We explicitly model the BEFORE relation in the event pair (u, v) (i.e., u happens earlier than v). For event pairs with the AFTER relation, we simply swap the events since AFTER and BEFORE are reciprocal.

In addition to the first loss term, we introduce a second novel loss term to enforce an angular property. As the example shown in Figure 2, we want to make $\angle\theta_1$ of a positive event pair (u, v) smaller. Based on preliminary tests, the angular loss can further enforce the first loss term which is driving the norm of u to be larger than the norm of v and thus increases the performance. Moreover, this angular property can help to distinguish VAGUE pairs by using a threshold on the $\angle\theta_2$ to determine whether to assign the event pair to the VAGUE label.

[3]The contextualized embeddings capture the context information around the event triggers through pre-trained language models. Thus, the resulting event embedding essentially also encodes the information about its subject and object.

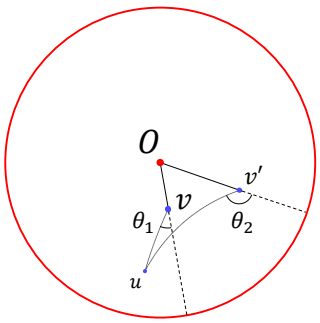

Figure 2: An illustration of the Poincaré embedding used to encode two events u and v with known temporal relation. θ_1 is the angle between the event pair (u, v), while θ_2 is the angle of an event pair (u, v') resulting from the negative sampling process.

We define the second loss term as the degree of $\angle\theta_1$, which is known to be equal to (Ganea et al., 2018a):

$$\mathcal{L}_2 = \sum_{(u,v)\in D} \arccos \left(\frac{\langle s_u, s_v\rangle(1+\|s_v\|^2) - \|s_v\|^2(1+\|s_u\|^2)}{\|s_v\| \cdot \|s_u - s_v\| \sqrt{1+\|s_u\|^2\|s_v\|^2 - 2\langle s_u, s_v\rangle}}\right). \quad (9)$$

Then, we train an HFFL based on the following objective function:

$$\mathcal{L} = \alpha\mathcal{L}_1 + (1-\alpha)\mathcal{L}_2, \quad (10)$$

where α is a hyperparameter to balance the importance of the two loss terms.

The training objective will push s_u towards the boundary, while pulling s_v close to the origin. Thus, the norm of s_u and s_v will provide key information to determine the temporal order of an event pair, while the angle $\angle\theta_1$ can help to distinguish VAGUE event pairs. We propose the following score function:

$$\text{score}(u, v) = \frac{\|s_u\|_{\mathbb{D}} - \|s_v\|_{\mathbb{D}}}{d_{\mathbb{D}}(s_u, s_v) + \epsilon}. \quad (11)$$

Based on the score function, different types of relations can be predicted using the following rules:

$$\text{TempRel}(u, v) = \begin{cases} \text{BEFORE}, & \text{if score} \in (t, 1) \\ \text{AFTER}, & \text{if score} \in (-1, -t) \\ \text{EQUAL}, & \text{if score} \in [-\epsilon, \epsilon] \\ \text{VAGUE}, & \text{if score} \in [-t, -\epsilon) \cup (\epsilon, t] \end{cases} \quad (12)$$

where the value of threshold $t \in (\epsilon, 1)$ is adjusted on the validation set.

4.2 Hyperbolic Neural Network for Temporal Relation Detection

Apart from the aforementioned approach, an alterative method is to train an end-to-end hyperbolic neural network using classification objective directly. We propose a hyperbolic neural network model for TempRel detection based on the operations defined on the usual Poincaré ball $\mathbb{D}^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$.

Given an input sequence containing an event pair (u, v), we first obtain its contextualized sentence representations from a pre-trained language model. The representations are denoted as a matrix $B \in \mathbb{R}^{l\times d}$, where l represents the sentence length and d is the dimension of word embeddings. The sentence representations are then projected onto a Poincaré ball and fed into a hyperbolic feed-forward layer, $C = \exp_0(B)$. The outputs are passed to a Hyperbolic Gate Recurrent Unit (HGRU) to derive the hidden state of each word. A position masking vector m_u or m_v, which has a value '1' in the position of an event and '0' otherwise, is applied to retrieve the hidden state representation of the corresponding event:

$$h_u = \text{HGRU}(C) \cdot m_u, \quad h_v = \text{HGRU}(C) \cdot m_v.$$

The HGRU on top of the RoBERTa output can further compose the information from the event triggers and corresponding subjects and objects.

Afterward, the two event hidden states are combined by performing weighted Möbius aggregation:

$$s_{uv} = \mathbf{W}_u \otimes h_u \oplus \mathbf{W}_v \otimes h_v \oplus \mathbf{b}_k. \quad (13)$$

The s_{uv} is further combined with the distance between the two event hidden states, $\text{d}_{\mathbb{D}}(h_u, h_v)$, before applying a hyperbolic non-linear function:

$$o = \varphi^{\otimes}\left(s_{uv} \oplus \text{d}_{\mathbb{D}}(h_u, h_v) \otimes \mathbf{w}_o\right). \quad (14)$$

The output o is then passed to a Hyperbolic Multinomial Logistic Regression (HMLR) layer (Eq. 6) to generate the event temporal relation classification result, $\hat{y} = \text{HMLR}(o)$. Figure 3 shows a schematic depiction of the network architecture.

Additionally, commonsense knowledge can be incorporated within the HGRU. We follow Ning et al. (2019) and use a Siamese network trained on TEMPROB[4], discretize its output, and turn the output into categorical embeddings. Then, we project the categorical embeddings onto a hyperbolic space

[4]https://github.com/qiangning/TemProb-NAACL18

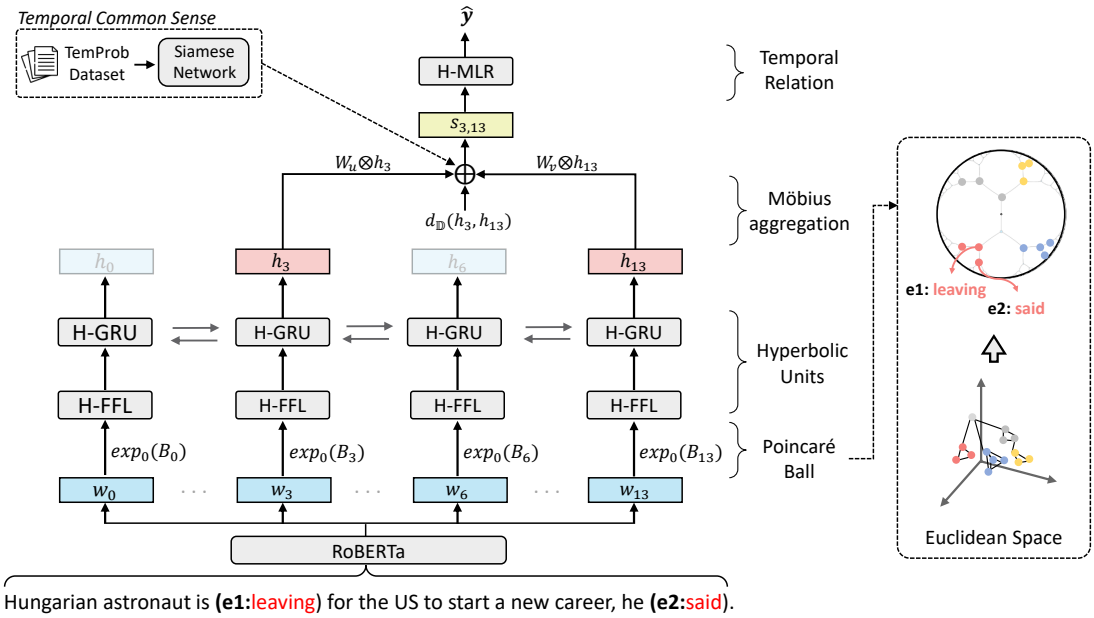

Figure 3: In the hyperbolic neural architecture for temporal relation extraction, sentence tokens are first associated to standard RoBERTa vectors (within the Euclidean space). They are subsequently mapped into a Poincaré ball and processed using Hyperbolic Feed-Forward Layers (H-FFL) and Hyperbolic-GRUs (H-GRU). Then, a masking process ensures that only the event-related vectors are aggregated via Möbius operations, along with their $d_{\mathbb{D}}$ distance and the relevant temporal common sense, extracted by a Siamese network pre-trained on TEMPROB knowledge base. Finally, the distribution over event temporal relations is derived using a Hyperbolic Multinomial Logistic Regression (H-MLR), analogous to a traditional Softmax layer in the Euclidean space.

and use a Riemannian optimizer to update them. The commonsense features can be directly combined with the components of Eq. 14.

4.3 The Use of Pre-trained Language Model

Both of our proposed methods incorporate pre-trained language models. We investigated several ways to utilize them, and include two of them in this paper. The first one follows Ning et al. (2019), which only uses the static output of the pre-trained models. This approach is fast and allows a fair comparison with the models in Ning et al. (2019).

The second approach fine-tunes the pre-trained language models during training on the TempRel Extraction objective. This approach can achieve better performance, but takes a longer time to train.

5 Experimental Setup

We describe the datasets, methodologies used in the recent literature for the TempRel extraction tasks. We also briefly present the parameter setup of our experiments. A description of the evaluation metrics can be found in the Appendix.

Dataset MATRES (Ning et al., 2018c) is a TempRel dataset that is composed of news documents.

Class	MATRES Train	MATRES Test	TCR
BEFORE	6,425	427	1,780
AFTER	4,481	271	862
EQUAL	418	30	4
VAGUE	1,416	109	0
Total	12,740	837	2,646

Table 1: The number of event pairs under each of the four relation classes in the MATRES and TCR datasets.

With the novel multi-axis annotation scheme, MATRES achieves much higher inter-annotator agreements (IAA) than previous temporal datasets, such as TB-Dense (Cassidy et al., 2014), RED (O'Gorman et al., 2016) and THYME-TimeML (Styler IV et al., 2014). MATRES consists of documents from three sources: TimeBank (183 documents), AQUAINT (72 documents), and Platinum (20 documents). We follow the official split in which TimeBank and AQUAINT are used for training, while Platinum is used for testing. We further split 20% of the training data as the validation set.

Temporal and Causal Reasoning (TCR) (Ning et al., 2018a) is another dataset that adopts the an-

Model	MATRES				TCR			
	P	R	Acc	F_1	P	R	Acc	F_1
CogCompTime (Ning et al., 2018d)	61.6	72.5	61.6	66.6	-	-	68.1	70.7
LSTM* (Ning et al., 2019)	70.2	76.5	67.3	73.2	79.6	75.7	75.7	77.6
LSTM+knowledge* (Ning et al., 2019)	70.2	80.1	70.1	74.8	79.3	76.9	76.9	78.1
LSTM+knowledge+ILP (Ning et al., 2019)	-	-	71.7	76.7	-	-	80.8	78.6
JCL-base (Wang et al., 2020)	67.7	80.3	-	73.5	-	-	-	-
JCL+multi-task+logic (Wang et al., 2020)	72.2	83.8	-	77.6	-	-	-	-
JCL-all (Wang et al., 2020)	73.4	**85.0**	-	78.8	-	-	-	-
Poincaré Event Embeddings (static RoBERTa)	72.6	82.1	71.9	77.1	81.4	80.1	80.1	80.7
Poincaré Event Embeddings (RoBERTa)	74.1	84.3	73.7	78.9	85.0	**86.0**	**85.0**	**85.5**
HGRU (static ELMo)	74.4	77.3	69.2	75.8	82.9	73.6	73.6	78.0
HGRU (static RoBERTa) + knowledge	73.4	84.3	73.4	78.5	83.2	83.2	83.2	83.2
HGRU (RoBERTa) + knowledge	**79.2**	81.7	**74.2**	**80.5**	**88.3**	79.0	79.0	83.5

Table 2: Experimental results on MATRES and TCR. Results presented in the top half are either directly taken from the cited papers or produced by employing the original source code supplied by the authors (models denoted by '*'). Results presented in the lower half are generated from our proposed models and their variants.

notation scheme defined in MATRES. It is a much smaller dataset, with just 25 documents and 2.6K TempRels. Due to the TCR limited size, we follow Ning et al. (2019) by using the temporal relations in TCR to test the model trained on MATRES. The statistics of the data used in our experiments is shown in Table 1.

Compared Methods We compare our proposed Poincaré event embedding (§4.1) and hyperbolic architecture for TempRel extraction (§4.2) with the following baselines:

CogCompTime (Ning et al., 2018d) is a pipeline system based on semantic features and structured inference.

LSTM denotes a TempRel detection model based on LSTMs (Ning et al., 2019). It has two variants, one with the incorporation of the commonsense knowledge (*LSTM+knowledge*) and one with the additional global inference via Integer Linear Programming (*LSTM+knowledge+ILP*).

Joint Constrained Learning (Wang et al., 2020) conducts joint training on both temporal and hierarchical relation extraction based on RoBERTa and Bi-LSTMs. It incorporates logic constraints and commonsense knowledge establishing the current state-of-the-art results on the MATRES dataset.

Parameter Setup Based on the results of our preliminary experiments, the contextualized embeddings are produced from RoBERTa in both of our proposed approaches. We also conducted preliminary experiments to determine the best dimension for the Poincaré embeddings, and found variation in performance having no statistical significance. We then adopted the 2D embeddings for the sake of simplicity and ease of visualization. More details about the model architecture and hyperparameter setting can be found in the Appendix.

6 Experimental Results

Overall Comparison Existing methodologies adopted for TempRel extraction commonly leverage several auxiliary components, such as external commonsense knowledge and multi-task objectives. Therefore, to better understand the impact made by the adoption of the hyperbolic geometry, we conduct additional experiments over ablated versions of the baseline models. In particular, in Table 2, results for the `LSTM` and `LSTM+knowledge` are produced by employing the original code provided by the authors[5]. Other results of compared methods are directly taken from the cited papers. For consistency and fair comparison with Ning et al. (2019), we test our method with inputs from the ELMo (Peters et al., 2018), which has been reported achieving the best overall results for the models proposed in (Ning et al., 2019).

We observe that the proposed Poincaré event embedding learning method presented in Section 4.1 outperforms `LSTM` and its variants

[5] https://github.com/qiangning/NeuralTemporalRelation-EMNLP19

which rely on fairly complex auxiliary features and constraints on both the MATRES and the TCR datasets, and produced more accurate event TempRel detection results even when compared to the JCL-base model. It is worth noticing that the Poincaré event embeddings (static RoBERTa) are trained with a shallow network with just 1.5k parameters and only takes about 4 minutes to train on a single RTX 2080 Ti. If further fine-tuning RoBERTa on MATRES, we observe a further improvement on F_1 by 1.8%.

Using our proposed alternative end-to-end HGRU model, HGRU (static ELMo) outperforms both the standard LSTM model and the variant incorporating commonsense knowledge (LSTM+knowledge) on MATRES. This verifies that the hyperbolic-based method is more efficient than its Euclidean counterparts. In order to fairly compare with the state-of-the-art model (Wang et al., 2020) on this task, we utilize RoBERTa and the auxiliary temporal commonsense knowledge since Wang et al. (2020) also fine-tunes RoBERTa on MATRES and uses external commonsense knowledge. The results show that HGRU (RoBERTa) + knowledge outperforms JCL-base (Wang et al., 2020) significantly by 7% in F_1. It even outperforms JCL-all, which further incorporates logic constraints and multi-task learning.

On the TCR dataset, the proposed hyperbolic-based methods also see similar improvement over existing methods. In terms of the difference between the two proposed methods, Poincaré Events Embeddings achieve a higher F_1 score on TCR. The reason is that HGRU tends to predict more VAGUE labels, but there is no VAGUE in TCR. Interestingly, both proposed methods predict less VAGUE labels when using static RoBERTa. A detailed breakdown of results for each temporal relation and training cost is presented in the Appendix.

Ablation Study In order to study the impact of different components of HGRU on event TempRel extraction, we conduct the ablation study.

First, the HGRU layer is removed and the contextual embeddings of events are directly fed into hyperbolic FFNN (HFFNN) while all the other hyperparameters are frozen. The resulting performance of HFFNN is significantly lower than HGRU, which indicates that the temporal information is spread across different time steps of the pre-trained language model output, and it is better

Model	**Acc**	F_1
HFFNN (static ELMo)	67.1	72.7
HFFNN (static RoBERTa)	67.9	73.6
HGRU (static ELMo) w/o $d_{\mathbb{D}}$	67.7	74.3
HGRU (static ELMo)	69.2	75.8
HGRU (static RoBERTa) w/o $d_{\mathbb{D}}$	71.0	76.1
HGRU (static RoBERTa) with EMLR	71.8	76.8
HGRU (static RoBERTa)	72.2	77.6
HGRU (static RoBERTa) + knowledge	73.4	78.5
HGRU (RoBERTa) + knowledge	**74.2**	**80.5**

Table 3: Ablation experiments on the MATRES dataset.

encoded by a recurrent architecture. HGRU w/o $d_{\mathbb{D}}$ shows the impact of the hyperbolic distance feature $d_{\mathbb{D}}(h_u, h_v)$ and its parameter $\mathbf{w}_o$ in Eq. 14. The results show that the hyperbolic distance between the hidden states of two events encodes relevant information to predict the event temporal relations.

Although, Ganea et al. (2018b) reported that mixing hyperbolic neural networks with Euclidean Multinomial Logistic Regression (EMLR) can at times achieve better performance than pure hyperbolic networks, yet we observe no significant difference on the MATRES dataset. Finally, as discussed earlier, fine-tunning RoBERTa gives better performance compared to static RoBERTa and ELMo.

Additionally, our ablation study on the Poincaré Event Embedding shows that without the angular loss ($\alpha = 1$, Eq. 10) it can only achieve 62.4 in F_1, compared to 77.1 while using it (static RoBERTa).

Case Study Figure 4 shows a set of Poincaré event embeddings resulting from the method proposed in section 4.1[6]. The events are numbered according to the temporal relations detected by the model (a smaller number denotes an earlier event). Among them, it is worth noting the two temporal paths between the *expected approval from the bank* (**e2**) and the *final acquisition of shares* (**e5**), i.e., $e2 \rightarrow e5$ and $e2 \rightarrow e3 \rightarrow e4 \rightarrow e5$. The first direct path is accompanied by a more fine-grained path, specifying the clearance granted from the authorities (**e3**) to permit (**e4**) the bank acquisition and the consequential buying of tendered shares (**e5**). The model has encoded more recent events closer to the origin, and events in the past closer

[6]Each point corresponds to the contextualized embedding of an event represented by its predicate, e.g., $e1$: *said*, $e2$: *expected*, etc. We denote each point in Figure 4 by a tuple *(subject, predicate, object)* for easy inspection.

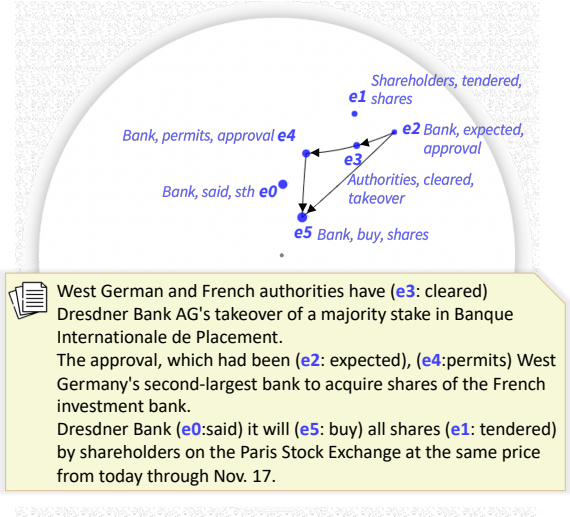

Figure 4: A document excerpt from the MATRES dataset and the related temporal event embedding generated by the Poincaré embedding method.

to the border, while simultaneously shaping a hierarchical structure to link their information with different granularity, for a resulting hierarchical structure with asymmetric connections.

7 Conclusion

In this paper, we proposed to model event temporal relations overcoming the limitations of Euclidean spaces, and designing two TempRel extraction methods using hyperbolic geometry. The first approach highlighted the convenience of learning event embedding in the Poincaré ball, achieving performance on-par with recent methodologies by using just a simple rule-based classifier. Then, we designed a hyperbolic neural network, incorporating temporal commonsense, outperforming state-of-the-art models on the standard datasets. Finally, a qualitative analysis pointed out the inherent advantage of employing hyperbolic spaces to encode asymmetric relations. In the future, we plan to extend our frameworks to a wider spectrum of event relations, including causal and sub-event relations.

Acknowledgements

This work was funded in part by the UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1). YH is supported by a Turing AI Fellowship funded by the UK Research and Innovation (grant no. EP/V020579/1).

References

Gary Becigneul and Octavian-Eugen Ganea. 2019. Riemannian adaptive optimization methods. In *Conference Track Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*.

Guillaume Bouchard, Sameer Singh, and Théo Trouillon. 2015. On approximate reasoning capabilities of low-rank vector spaces. In *Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches, AAAI Spring Symposium Series*.

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

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic. Association for Computational Linguistics.

Yao Cheng, Peter Anick, Pengyu Hong, and Nianwen Xue. 2013. Temporal relation discovery between events and temporal expressions identified in clinical narrative. *Journal of biomedical informatics*, 46.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. 2018a. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018b. Hyperbolic neural networks. In *Advances in neural information processing systems*, pages 5345–5355.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. 2018. Hyperbolic attention networks. In *International Conference on Learning Representations*.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with

shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Max Kochurov, Rasul Karimov, and Sergei Kozlukov. 2020. Geoopt: Riemannian Optimization in PyTorch. *ArXiv*, abs/2005.02819.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

N. Linial, E. London, and Y. Rabinovich. 1994. The geometry of graphs and some of its algorithmic applications. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, SFCS '94, page 577–591.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Federico López and Michael Strube. 2020. A fully hyperbolic neural model for hierarchical multi-class classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 460–475, Online. Association for Computational Linguistics.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chungmin Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760.

Maximilian Nickel, Xueyan Jiang, and Volker Tresp. 2014. Reducing the rank in relational factorization models by including observable patterns. In *Advances in Neural Information Processing Systems*, volume 27, pages 1179–1187.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6204–6210.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Qiang Ning, Hao Wu, and Dan Roth. 2018c. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018d. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. *Computing News Storylines*, page 47.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. 2018. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 4460–4469.

Rik Sarkar. 2011. Low distortion delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Conference on Graph Drawing*, GD'11, page 355–366.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

A. Tifrea, G. Becigneul, and O.-E. Ganea. 2019. Poincaré GloVe: Hyperbolic word embeddings. In *7th International Conference on Learning Representations (ICLR)*.

Lucas Vinh Tran, Yi Tay, Shuai Zhang, Gao Cong, and Xiaoli Li. 2020. Hyperml: A boosting metric learning approach in hyperbolic space for recommender systems. In *WSDM*, pages 609–617.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the tarsqi toolkit. In *COLING 2008: Companion Volume: Demonstrations*, pages 189–192.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2018. Event phase oriented news summarization. *World Wide Web*, 21(4):1069–1092.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chengkun Zhang and Junbin Gao. 2020. Hype-han: Hyperbolic hierarchical attention network for semantic embedding. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3990–3996.

Appendix

A Model Architecture and Hyperparameter Setting

The implementation of the proposed models is based on the *geoopt* package (Kochurov et al., 2020). RoBERTa used in the experiments is downloaded from Huggingface[1] (Wolf et al., 2020). ELMo is downloaded from AllenNLP[2].

The implementation architecture of the TempRel HGRU model is shown in Table A2. We use the RoBERTa base model, whose output dimension d_1 is 768. The dimension of the hidden states d_2 is set to 128. The commonsense embedding dimension d_3 is set to 32. The FFNN output dimension d_4 is 64. The activation function before the HMLR layer is ReLU. We use Riemannian Adam (Becigneul and Ganea, 2019) to optimize the hyperbolic parameters, and the standard Adam optimizer for parameters in the Euclidean space. The learning rate for RoBERTa fine-tuning is 1×10^{-5}, and for other parameters in our proposed models is set to 1×10^{-3}.

For the Poincaré event embeddings (§4.1), the contextualized embeddings are produced from RoBERTa-base. We conducted preliminary experiments to determine the best dimension for the Poincaré embeddings, and found variation in performance having no statistical significance. We then adopted the 2D embeddings for the sake of simplicity and ease of visualization. The weight α for balancing two loss terms is set to 0.5. The number of negative samples is set to 1.

B A Breakdown of Evaluation Results by Temporal Relation Types

Table A1 shows the model performance with respect to each relation type. The two proposed methods show increased performance on BEFORE and AFTER relations which take the majority in the MATRES and TCR datasets. The F_1 scores on EQUAL and VAGUE are considerably low. This is probably due to the limited number of EQUAL instances. For the VAGUE instances, even human annotators are unable to determine the relations. Interestingly, our TempRel HGRU with RoBERT fine-tuning has significantly higher F_1 score on VAGUE compared to LSTM or Poincaré event embeddings. Our hypothesis for this paper is that hyperbolic geometry can improve the extraction of asymmetric relations, but VAGUE is symmetric. The improvement on VAGUE is unexpected. Further investigation is needed to find out the reason behind this phenomenon.

[1] https://huggingface.co/
[2] https://github.com/allenai/allennlp

Relation	MATRES P	MATRES R	MATRES F_1	TCR P	TCR R	TCR F_1
LSTM (Ning et al., 2019)						
Before	73.0	85.0	78.6	87.8	80.1	83.8
After	66.5	80.4	72.8	66.4	75.5	70.6
Equal	0	0	0	0	0	0
Vague	33.3	3.7	6.6	-	-	-
Poincaré Event Embeddings (ours)						
Before	75.5	90.2	82.2	90.9	87.0	88.9
After	71.8	84.5	77.6	77.0	81.2	79.1
Equal	0	0	0	0	0	0
Vague	37.5	3.8	5.1	-	-	-
TempRel HGRU (ours)						
Before	82.4	85.7	84.0	95.3	77.5	85.5
After	74.6	84.5	79.2	77.4	82.6	80.0
Equal	0	0	0	0	0	0
Vague	30.2	23.9	26.7	-	-	-

Table A1: A breakdown of evaluation results by temporal relation types.

C Training Cost

The training time of the proposed TempRel HGRU (static RoBERTa) is about 2 minutes per epoch on the MATRES dataset (single Nvidia RTX 2080Ti GPU, batch_size=250). Evaluation on the validation set takes around 23 seconds. The best validation score tends to appear at around the 10th epoch. The proposed Poincaré event embedding method is much more efficient. The entire training of the Poincaré event embeddings (static RoBERTa) only takes about 4 minutes (single Nvidia RTX 2080 Ti, 20 epochs).

Our HGRU with RoBERTa fine-tuning takes about 32 minutes per training and validation epoch to train on a single Nvidia RTX 3090 GPU. The best validation score also tends to appear at around the 10th epoch.

For comparison, the state-of-the-art model (Wang et al., 2020) requires about 8.5 minutes per epoch in training if not fine-tuned the RoBERTa. If fine-tuning RoBERTa, it takes about 32.5 minutes per epoch for training (including validation).

D Evaluation Metrics

For the evaluation scores used in Table 2, we follow the widely adopted evaluation metrics proposed in Ning et al. (2019), which is also adopted by Wang et al. (2020). In particular, given the four tempo-

Input		Sentences $\{s_i\}_{i=1}^{N}$ each of which containing one of more event pairs (u, v) with a token sequence $s_i = \{x_{ij}\}_{j=1}^{L}$, the masks (m_u, m_v) indicate the position of the events in a sentence.
Contextual embedding	ELMo or RoBERTa	Encoded by a pre-trained language model $\{x_{ij}\}_{j=1}^{L} - \{\text{ELMo/RoBERTa}\} \rightarrow: \{c_{ij}\}_{j=1}^{L} \in \mathbb{R}^{d_1 \times L}$
Hyperbolic encoding	HGRU	Map the contextual embeddings onto hyperbolic space $\{\boldsymbol{c}_{ij}\}_{j=1}^{L} - \{\text{expmap}_0\} - \{\text{HGRU}\} \rightarrow: \boldsymbol{h}_u, \boldsymbol{h}_v \in \mathbb{R}^{d_2}$
H-distance		Compute the distance between two points on the Poincaré model $d_{\mathbb{D}}(\boldsymbol{h}_u, \boldsymbol{h}_v) \in \mathbb{R}$
Commonsense	VerbNet	Extract score from the pre-trained VerbNet $\boldsymbol{x}_u, \boldsymbol{x}_v - \{\text{VerbNet}\} \rightarrow: \boldsymbol{k} \in \mathbb{R}$, discretize the output score $\boldsymbol{k} - \{discretize\} - \{embedding\} \rightarrow: \boldsymbol{g} \in \mathbb{R}^{d_3}$
Combine	H-concat	Hyperbolic concatenation $\boldsymbol{h}_u \oplus \boldsymbol{h}_v \oplus d_{\mathbb{D}}(\boldsymbol{h}_u, \boldsymbol{h}_v) \oplus \boldsymbol{g} - \{relu\} \rightarrow: \boldsymbol{o} \in \mathbb{R}^{d_4}$
Classification	HMLR	Hyperbolic multinomial logistic regression $\boldsymbol{o} - \{HMLR\} - \{Softmax\} \rightarrow: \hat{y}$

Table A2: TempRel HGRU framework and hyperparameters

ral relation types: BEFORE, AFTER, EQUAL and VAGUE, a confusion matrix can be built with the row $C_{[k,:]}, k \in \{b, a, e, v\}$ representing the gold label of class k; the column $C_{[:,k]}$ representing model prediction of class k. The metrics are computed as follows:

Accuracy. $Acc = (C_{b,b} + C_{a,a} + C_{e,e} + C_{v,v})/S$, where S is the sum of all the values in the confusion matrix.

Precision, recall, and F_1. The *precision* $P = (C_{b,b} + C_{a,a} + C_{e,e})/S_1$, where S_1 indicates the sum of the first three columns in the confusion matrix. The *recall* $R = (C_{b,b} + C_{a,a} + C_{e,e})/S_2$, where S_1 denotes the first three rows in the confusion matrix. $F_1 = 2PR/(P + R)$.