Multilingual SubEvent Relation Extraction: A Novel Dataset and Structure Induction Method

[1]Viet Dac Lai*, [2]Hieu Man*, [3]Linh Ngo Van, [4]Franck Dernoncourt, [1]Thien Huu Nguyen
[1]Department of Computer Science, University of Oregon, Eugene, Oregon, USA
[2]VinAI Research, Vietnam
[3]Hanoi University of Science and Technology, Hanoi, Vietnam
[4]Adobe Research, Seattle, WA, USA
{vietl,thien}@cs.uoregon.edu, v.hieumdt@vinai.io
linhnv@soict.hust.edu.vn, franck.dernoncourt@adobe.com

Abstract

Subevent Relation Extraction (SRE) is a task in Information Extraction that aims to recognize spatial and temporal containment relations between event mentions in text. Recent methods have utilized pre-trained language models to represent input texts for SRE. However, a key issue in existing SRE methods is the employment of sequential order of words in texts to feed into representation learning methods, thus unable to explicitly focus on important context words and their interactions to enhance representations. In this work, we introduce a new method for SRE that learns to induce effective graph structures for input texts to boost representation learning. Our method features a word alignment framework with dependency paths and optimal transport to identify important context words to form effective graph structures for SRE. In addition, to enable SRE research on non-English languages, we present a new multilingual SRE dataset for five typologically different languages. Extensive experiments reveal the state-of-the-art performance for our method on different datasets and languages.

1 Introduction

In Information Extraction (IE), events are defined as things that happen/occur (Pustejovsky et al., 2003) or changes of state of real world entities (Walker et al., 2006). Due to their complexity, a general event (i.e., superevent) can involve multiple other events with finer granularity (i.e., subevents) that can be altogether mentioned in text to present necessary details (e.g., a *war* can contain multiple *attacks*, which, in turn, can contain different bombing events). To this end, our paper studies the problem of subevent relation extraction (SRE): given two event mentions in a document, a model needs to predict if one even is a part/subevent of the other one. Following previous work (Glavaš et al., 2014), our SRE problem requires that a subevent relation is only established if the subevent is both spatially and temporally contained in the superevent. Accordingly, SRE systems will need to effectively model document context to infer spatiotemporal evidences for subevent reasoning. Among others, SRE finds its important applications in summarization (Filatova and Hatzivassiloglou, 2004) and information retrieval (Glavaš and Šnajder, 2013).

To encode document context, existing models (Wang et al., 2020; Trong et al., 2022) have leveraged pre-trained language models, i.e., RoBERTa (Liu et al., 2019), to obtain representations for input documents for subevent prediction. However, an issue of existing SRE methods is that they only rely on the sequential format of documents (i.e., sequence of sentences/words) for representation learning. On the one hand, the sequential format does not provide mechanisms to highlight the most important context words or avoid irrelevant ones in input documents, potentially introducing noisy information in the representations for SRE. Further, due to the sequential nature of input texts, current SRE models cannot exploit effective structures/graphs that directly connect important context words to improve representation learning for SRE.

Motivated by recent works on relation extraction between entities (Zhang et al., 2018; Gupta et al., 2019; Sahu et al., 2019), one approach to improve sequential representation of input texts for SRE can be based on dependency trees of sentences (i.e., graph-based structures) where dependency paths (DP) between two input entity mentions have been shown to capture important context words. In particular, to adapt this idea to document level with multiple sentences, (Gupta et al., 2019) obtains dependency trees for each sentence whose roots are linked together to obtain connected dependency graphs for input documents. Afterward, the dependency graphs for documents are prune to preserve only the words along the dependency paths between two input mentions (called in-DP words)

*The first two authors contribute equally to this paper.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5559–5570
December 7-11, 2022 ©2022 Association for Computational Linguistics

for representation learning. However, for our SRE problem, important context words for subevent prediction can also be distributed outside the dependency paths, thus necessitating further techniques to identify other important words and connect them with the in-DP words to form better graph structures to represent input texts for SRE. For example, in the input text "*They implemented the proposal early last year. Following the plan, the performers collected data and developed frameworks to monitor human trafficking for the first step of the proposal.*", "*developed*" is a subevent of the "*implemented*" event for which the DP is "*implemented* $\rightarrow$ *collected* $\rightarrow$ *developed*". However, the word "*proposal*", which is important to connect "*implemented*" and "*developed*" to the same target for subevent recognition, is not included in the DP in this case. For convenience, we use non-DP words to refer to the words that do not belong to the DPs between two input event mentions for SRE.

In previous work, in-DP words can be extended to find additional important context words for relation prediction by including non-DP words close to the DPs in the dependency graphs (Zhang et al., 2018) (i.e., based on syntactic distances). As such, this method does not consider contextual semantics of the words that can provide richer information for important word selection for SRE. To address this issue, we propose to leverage both syntactic and semantic evidences to determine the importance of a non-DP word for inclusion into the graph structure to represent input text for SRE. For syntactic information, we expect a word to be more important for subevent prediction if it is closer to the input event mentions in the dependency graphs. In addition, for semantic information, our intuition is to promote non-DP words that are more similar/related to in-DP words contextually to enhance the induced representations for SRE. However, combining syntactic and semantic similarities to compute overall importance scores to compare non-DP words is a non-trivial problem due to the different nature of the information. To this end, motivated by in-DP words as the anchors to induce graph structure representations for input texts, we propose to cast the problem of combining syntactic and semantic similarities to select important non-DP words into finding an optimal alignment between non-DP and in-DP words. A non-DP word is considered to be important for SRE and retained in the induced graph structures for input texts if it is aligned with one of the in-DP words. In this way, our approach facilitates the application of Optimal Transport (OT) methods to effectively integrate syntactic and semantic information into a single joint optimization problem to obtain the optimal alignment for non-DP word selection for SRE. In particular, to adapt to the goal of aligning two groups of points based on their transportation costs and distributions in OT, we will leverage semantic similarity to obtain transportation costs while syntactic distances in dependency graphs will be used to compute the distributions for in-DP and non-DP words to perform word alignment for SRE. The resulting word alignment will then be used to select important non-DP words and construct graph structures to learn representations for subevent prediction.

We evaluate our method over HiEve (Glavaš et al., 2014) and Intellgience Community (IC) (Hovy et al., 2013), popular public datasets for SRE. However, an issue with prior datasets and methods for SRE is that they are only developed and evaluated over English data. As such, a critical question for the generalization of SRE methods to non-Enlgish languages has not been explored in the literature. To address this issue, we further present a new multilingual dataset for SRE (called mSubEvent) for five languages, i.e., English, Danish, Spanish, Turkish, and Urdu, to enable future research in multilingual learning for SRE. Our dataset follows the annotation guidelines in HiEve to make it consistent with prior SRE work, introducing a large SRE dataset with more than 46K event mentions and 3.9K subevent relations for model development. We conduct extensive experiments over HiEve and our new dataset mSubEvent to demonstrate the effectiveness of the proposed method with state-of-the-art performance for SRE. Our experiments cover both monolingual learning (i.e., training and test data are from the same language) and cross-lingual transfer learning evaluation (i.e., training and test data comes from different language), thus highlighting the generalization across languages of the proposed method for SRE. To our knowledge, this is the first work that explores multilingual data and cross-lingual learning for SRE. Finally, we will publicly release the new mSubEvent dataset to provide baselines and resources for future research in this area.

2 Model

Following prior work (Trong et al., 2022), we utilize pairwise classification to formulate SRE. Given a document $D = [w_1, w_2, \dots, w_n]$ (of n words) with w_{e_1} and w_{e_2} as two input event mentions/triggers, a SRE model needs to classify the relation between w_{e_1} and w_{e_2} according to one of the three types for subevents, i.e., `PARENT-CHILD`, `CHILD-PARENT`, and `NOREL`. Here, the `NOREL` type is to indicate no subevent relation.

Input Encoding: In the first step, our model feeds the input document D into a pre-trained language model (PLM), i.e., RoBERTa (Liu et al., 2019), to obtain a representation vector v_i for each word $w_i \in D$. Here, we utilize the hidden vectors in the last transformer layer where vectors for the word-pieces in w_i are averaged to compute v_i. For convenience, let $V = v_1, v_2, \dots, v_n$ be the sequence of representation vectors for the words in D. Note that if the length of the input document exceeds the length limit in PLMs (i.e., 512 sub-tokens), we split the document into smaller segments to fit into the limit and run PLM over each segment separately to obtain the representations in V.

Structure Induction: As presented in the introduction, our method aims to transform the sequential format of D into a graph representation that can better capture important context and structures for representation learning for SRE. Motivated by the dependency path between w_{e_1} and w_{e_2} to capture important context for relation prediction (Zhang et al., 2018; Gupta et al., 2019), we first build a dependency graph T for D to initialize our graph construction process. In particular, we obtain dependency trees for the sentences in the document and connect the roots of the trees for consecutive sentences to create T. We leverage the Trankit toolkit (Nguyen et al., 2021) to generate dependency trees and ignore directions in the edges of the trees in the computation. As such, a property of the non-DP words in T is that they can involve both important and irrelevant context words for our subevent prediction problem (as demonstrated in the introduction). Accordingly, to compute an effective graph structure for D for SRE, our goal is to prune the dependency graph T so that only important context words are retained (i.e., removing irrelevant works). Using in-DP words in T as the anchor (i.e., presumably with important context), we aim to further select non-DP words that involve important context to perform the pruning of T for SRE. To this end, we propose to cast the non-DP word selection problem into an alignment problem between non-DP and in-DP words in which a non-DP word is considered as important for subevent prediction if it is aligned with one in-DP word in the alignment (i.e., extending the anchor in-DP words). To compute the alignment between the words for SRE, we propose to model both syntactic and semantic similarities between non-DP and in-DP words where Optimal Transport (OT) (Peyre and Cuturi, 2019) is leveraged to facilitate the information combination for optimal alignment computation.

Optimal Transport: OT is an established method to find the optimal plan to transform one distribution to another. Given two distributions $p(x)$ and $q(y)$ over discrete domains $\mathcal{X}$ and $\mathcal{Y}$ (respectively), and the cost function $C(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ to map $\mathcal{X}$ into $\mathcal{Y}$, OT finds the optimal joint alignment/distribution $\pi^*(x, y)$ (over $\mathcal{X} \times \mathcal{Y}$) with marginals $p(x)$ and $q(y)$, i.e., the cheapest transportation from $p(x)$ to $q(y)$, by solving the following problem:

$$\begin{gathered}\pi^*(x,y) = \min_{\pi\in\Pi(x,y)} \sum_{\mathcal{Y}}\sum_{\mathcal{X}} \pi(x,y)C(x,y)dxdy \\ \textbf{s.t. } x \sim p(x) \text{ and } y \sim q(y),\end{gathered} \tag{1}$$

where $\Pi(x, y)$ involves all joint distributions with marginals $p(x)$ and $q(y)$. Here, the distribution $\pi^*(x, y)$ is a matrix whose entry (x, y) captures the probability of transforming the data point $x \in \mathcal{X}$ to $y \in \mathcal{Y}$ for the conversion of $p(x)$ to $q(y)$. Note that to obtain a hard alignment between data points $\mathcal{X}$ and $\mathcal{Y}$, we can align each row of $\pi^*(x, y)$ with the column with the highest probability, i.e., $y^* = \text{argmax}_{y\in\mathcal{Y}}\pi^*(x, y)$ for all $x \in \mathcal{X}$.

To adopt OT to solve our non-DP word selection problem, we propose to treat the in-DP words in T as the data points for domain $\mathcal{Y}$ while the non-DP words will be used for domain $\mathcal{X}$. As such, OT facilitates the integration of syntactic and semantic similarities into the computation of optimal alignment between in-DP and non-DP words by leveraging these information to compute the transformation cost function $C(x, y)$ and the probability distributions $p(x)$ and $p(y)$. In particular, to compute $p(x)$ and $q(y)$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we use syntactic distances of the words to the input event mentions. Formally, for each word $w_i \in D$, we obtain the lengths of the paths that connect w_i with the input event mentions w_{e_1} and w_{e_2} in the dependency graph T, i.e., d_i^1 and d_i^2, respectively.

The syntactic importance of w_i for SRE is then determined by $syn(w_i) = \max(d_i^1, d_i^2)$. Afterward, the distributions $p(x)$ and $p(y)$ can be obtained by normalizing the syntactic importance scores (with softmax) for the words in the corresponding sets of $\mathcal{X}$ and $\mathcal{Y}$. Next, for the transportation cost $C(x, y)$, we leverage the contextual semantics for the words x and y, measured by the Euclidean distance between their representation vectors v_x and v_y (i.e., in V), i.e., $C(x, y) = ||v_x - v_y||$.

In addition, to aid the selection of non-DP important words, we introduce an extra data point, called `NIL`, to the in-DP set $\mathcal{Y}$ so non-DP words in $\mathcal{X}$ aligned with `NIL` will be considered irrelevant and excluded from T for graph structure induction for SRE. As such, the representation for `NIL` is computed using average of the representation vectors of the in-DP words in $\mathcal{Y}$ (i.e., to used for the transportation cost $C(x, y)$). Also, we utilize the average syntactic importance scores for the words in $\mathcal{X}$ to serve as the syntactic score $syn(\texttt{NIL})$ for `NIL` (the distribution $p(x)$ can be obtained accordingly). In this way, solving Equation 1 returns the optimal alignment $\pi^*(x, y)$ that can provide hard alignment for the data points in $\mathcal{X}$ and $\mathcal{Y}$[1]. Let I be the subset of non-DP words in $\mathcal{X}$ that are not aligned with `NIL` in $\mathcal{Y}$ according to $\pi^*(x, y)$ (i.e., irrelevant words). To this end, to prune the dependency graph T for SRE, we can eliminate the words in I from T to produce a new graph that only involves induced important context words for subevent prediction. However, as the resulting graph might be disconnected, we further retain the words in the paths between any word in I and the input event mentions (i.e., w_{e_1} and w_{e_2}), generating a new graph T' to serve as our induced graph structure to represent the input document for SRE.

In the next step, given the induced structure T', we feed it into a Graph Convolutional Network (GCN) (Kipf and Welling, 2017; Nguyen and Grishman, 2018) to learn richer representation vectors for the words in T'. The representation vectors from the PLM (i.e., in V) serve as the inputs for GCN. As such, the induced hidden vectors in the last layer of GCN are denoted by $V' = v'_{i_1}, \ldots, v'_{i_{|T'|}}$. Finally, we obtain an overall representation vector A for D for SRE via the concatenation: $A = [v'_{e_1}, v'_{e_2}, \text{max\_pool}(v'_{i_1}, \ldots, v'_{i_{|T'|}})]$ where v'_{e_1} and v'_{e_2} are the GCN-induced representation vectors in V' for the input event mentions w_{e_1} and w_{e_2}. The representation A will then be sent into a feed-forward network FF with softmax in the end to compute a distribution $P(\cdot|D, w_{e_1}, w_{e_2}) = FF(A)$ over the possible subevent relations. The negative log-likehood function over $P(\cdot|D, w_{e_1}, w_{e_2})$ will be used to train our SRE model in this work.

3 Data Annotation

There exist several datasets with subevent relation annotation, including HiEve (Glavaš et al., 2014), IC (Hovy et al., 2013; Araki et al., 2014), and RED (O'Gorman et al., 2016). However, these datasets are only annotated for English data, thus unable to evaluate the generalization of models across multiple languages. To better evaluate the proposed model and enable future research on multilingual SRE, we introduce the first multilingual dataset (called mSubEvent) for SRE that provides human annotation for five typological different languages, i.e., English, Danish, Spanish, Turkish, and Urdu. The rest of this sections describes our annotation schema, data collection, and annotation efforts.

Annotation Scheme: A dataset for SRE needs to provide annotations for two tasks, i.e., event mention and subevent relation extraction. As such, we inherit the well-designed annotation guidelines from existing benchmark datasets for both tasks to be consistent with prior work. In particular, we employ the annotation guideline and definition for event mentions from the popular ACE-2005 dataset (Walker et al., 2006). As our dataset focuses on subevent relations, we only annotate event mention spans and do not provide event types to reduce annotation cost. We allow event mentions to span multiple consecutive words in a sentence to flexibly handle different languages. In addition, for subevent relation annotation, we follow the guidelines from HiEve (Glavaš et al., 2014), a popular dataset for SRE. Following recent work (Wang et al., 2020), our dataset assigns a relation label for each pair of annotated event mentions in a document using three labels, i.e., `PARENT-CHILD`, `CHILD-PARENT`, and `NOREL`.

Data Collection & Preparation: To enable public release of our dataset, we collect documents for annotation from Wikipeda of the five intended languages. In particular, we obtains document from five event-intensive topics/categories in Wikipedia, including aviation accidents, railway accidents, nat-

[1] We employ the entropy-based approximation of OT and solve it with the Sinkhorn algorithm (Peyre and Cuturi, 2019).

Language	Event	Relation
English	0.92	0.96
Danish	0.68	0.83
Spanish	0.84	0.78
Turkish	0.69	0.66
Urdu	0.65	0.88
Average	0.75	0.82

Table 1: Kappa agreement scores.

Language	#Docs	#Events	#Rels	#Cross
English	438	8,732	841	8.7%
Danish	519	6,909	904	36.1%
Spanish	746	11,839	545	22.0%
Turkish	1,357	14,179	1,068	64.4%
Urdu	531	4,975	586	27.3%
Total	3,591	46,634	3,944	34.7%

Table 2: Statistics of our mSubEvent dataset. #Rels represents the number of subevent relations while #Cross indicate the percentage of subevent relations that involve event mentions in different sentences.

ural disasters, conflicts, and economic crisis. To do that, we exploit the category hierarchy in Wikipedia where a category involves a group of finer topic subcategories. Given the initial list of five categories, we crawl articles associated with the categories and their descendants (i.e., subcategories, subsubcategories) up to a hierarchy depth of 6. Here, by exploiting the interlinks across languages, we are able to retrieve Wikipedia articles in non-English languages for the chosen categories. In the next step, the crawled articles are then cleaned by removing markup elements (e.g., lists, tables, images). Finally, the articles are split into sentences and tokenized into words by Trankit (Nguyen et al., 2021), a multilingual NLP toolkit.

Annotating Wikipedia articles can be challenging and overwhelming as the articles tend to be long and the number of possible mention pairs grows quadratically with respect to the number of event mentions in a document. As such, to facilitate the annotators, we follow prior practices for event annotation (Mostafazadeh et al., 2016; Ebner et al., 2020) to split the cleaned articles into shorter chunks that contain five consecutive sentences (called documents in this work). In this way, the annotators only need to process a shorter document at a time to improve their attention and quality of annotated data.

Human Annotation: We hire annotators from `upwork.com`, a global crowdsourcing platform. We only consider candidates who are native speakers in our target languages and fluent in English. These information are provided in the annotators' profile in the platform. The candidates are provided with annotation guidelines and instructions for annotation interface, i.e., based on the BRAT annotation tool in our case (Stenetorp et al., 2012). Afterward, the candidates are invited to perform a designed test for both event mention and subevent relation annotation. For each language, the top two candidates are chosen for the annotation job.

We divide our annotation task into two steps for event mention and subevent relation annotation. For each language, we annotate subevent relations over the outputs from event mention annotation (i.e., after event mention annotation has been completed and finalized for all documents). Given a sample of selected documents for a language, for each step, the two annotators for that language independently annotate event mentions/subevent relations for the documents. Each annotator will completely annotation one document at a time. Afterward, the annotation conflicts are presented to the annotators for further discussion and revision to produce the final version of annotated documents for the current task. This helps to achieve high agreement and consistency for our dataset.

Data Analysis: Table 1 shows our Kappa scores for annotation agreements of event mention and subevent relation annotation over five languages. Note that these scores are computed by comparing the independent annotations of the annotators over the documents (i.e., before the discussion to resolve conflicts). As can be seen, the scores are very close to either substantial or almost perfect agreement for all the tasks and languages, thus demonstrating high quality of our multilingual SRE dataset. We also find that non-English languages tend to have lower annotation agreement scores for both annotation tasks, thus highlighting the challenges of SRE for non-English languages that necessitate further research effort in this area. In addition, Table 2 show major statistics. The #Cross column in the table shows that all languages in our dataset involve event mentions in different sentences for the subevent relations (i.e., cross-sentence relation), thus necessitating document-level context modeling. Among the five languages, English has the smallest percentage for cross-sentence relation that further reveals the challenge of SRE for non-

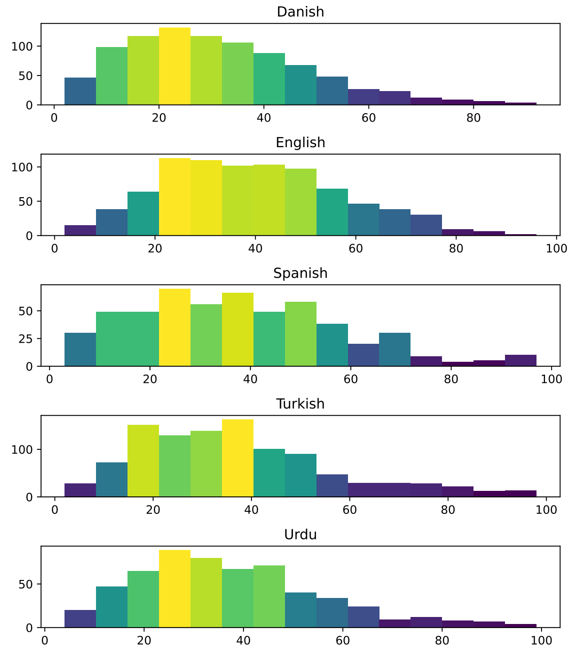

Figure 1: Distributions of distances between two event mentions with subevent relations. Distances are measured via the number of words.

English languages.

To provide more insight for our multilingual SRE dataset mSubEvent, Figure 1 shows the distributions of distances between two event mentions with subevent relations for five languages in mSubEvent. As can be seen, a majority of event mention pairs are 10 to 50 words away from each other in the documents, suggesting diverse levels of context information between event mentions that must be captured by SRE models for mSubEvent.

4 Experiments

Datasets & Hyper-Parameters: Similar to prior work (Wang et al., 2020, 2021; Trong et al., 2022), we evaluate our proposed model with optimal transport (called OT-SRE) on the popular datasets for SRE, i.e., **HiEve** (Glavaš et al., 2014) and Intelligence Community (IC) (Hovy et al., 2013). In particular, HiEve provides subevent and coreference relation annotation for events over 100 news articles using four relation labels, i.e., PARENT-CHILD, CHILD-PARENT (for subevents), COREF (for coreference), and NOREL (for no relation). To make it comparable, we utilize the same data split and setting as the current work with best-reported performance for HiEve (Wang et al., 2020; Trong et al., 2022), featuring 80 documents for training (2,423 subevent relations and 0.4 probability for down-sampling of negative examples) and 20 documents for testing (817 subevent relations). For IC, it also annotates 100 news articles for four subevent and coreference relations as in HiEve. Following the same setting in the current state-of-the-art method for IC (Wang et al., 2021), we discard relations with implicit event mentions and compute transitive closure for both subevent relations and coreference to obtain annotation for all event mention pairs as in HiEve (Glavaš et al., 2014). Also, IC is divided into three portions with 60/20/20 documents for training/development/test data respectively.

In addition, we evaluate the SRE models on the new multilingual dataset mSubEvent to provide baselines for future research. Here, we randomly split the documents for each language in mSubEvent into three separate parts with a ratio of 3/1/1 for training, development, and test data (respectively). We will use mSubEvent to evaluate SRE models in both monolingual and cross-lingual transfer learning experiments.

We fine-tune the hyper-parameters for our OT-SRE model over English development data of mSubEvent and apply the selected values for all experiments for consistency. In particular, the selected hyper-parameters for our model include: 2 layers for the GCN and feed-forward (i.e., FF) models with 512 dimensions for the hidden vectors, 5e-5 for the learning rate with Adam optimizer, and 16 for the batch size. Finally, we utilize the the RoBERTa$_{base}$ model (Liu et al., 2019) to encode input texts for HiEve as in prior work (Wang et al., 2020; Trong et al., 2022). For mSubEvent, we use the multilingual pre-trained language models (base versions), i.e., mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020), for multilingual text encoding.

Baselines: For HiEve, we compare our proposed SRE model with the following baselines using the same data setting: **StructLR** (Glavaš et al., 2014) with feature engineering, **TacoLM** (Zhou et al., 2020) with temporal common sense knowledge, **Joint** (Wang et al., 2020) with joint subevent and temporal relation extraction, **EventSeg** (Wang et al., 2021) with event-based text segmentation, and **SCS** (Trong et al., 2022) with selection of best context sentences for SRE. Similarly, for IC, we consider **Joint**, **EventSeg**, and **SCS** for the baselines. Note that **SCS** and **EventSeg** have the state-of-the-art (SOTA) performance for HiEve and IC

Model	F1 score		
	PC	CP	Avg
HiEve			
StructLR (Glavaš et al., 2014)	52.2	63.4	57.7
TacoLM (Zhou et al., 2020)	48.5	49.4	48.9
Joint (Wang et al., 2020)	62.5	56.4	59.5
EventSeg (Wang et al., 2021)	58.6	57.9	58.3
SCS (Trong et al., 2022)	68.7	63.2	65.9
OT-SRE (ours)	**70.3**	**67.4**	**68.9**
IC			
(Araki et al., 2014)	-	-	26.2
Joint (Wang et al., 2020)	42.1	49.5	45.8
EventSeg (Wang et al., 2021)	44.6	51.6	48.1
SCS (Trong et al., 2022)	47.5	51.8	49.7
OT-SRE (ours)	**48.9**	**52.6**	**50.8**

Table 3: Model performance on test data of HiEve and IC. We focus on the performance for `PARENT-CHILD` (PC), `CHILD-PARENT` (CP), and their micro-average to be consistent with prior evaluation for SRE (Trong et al., 2022; Wang et al., 2021).

Model	English	Danish	Spanish	Turkish	Urdu
mBERT					
PLM	36.5	30.2	23.6	39.0	34.1
EventSeg	41.1	41.7	37.4	42.8	43.1
SCS	46.8	45.9	40.6	44.0	50.1
OT-SRE	**49.3**	**48.9**	**42.1**	**50.1**	**52.2**
XLMR					
PLM	40.1	33.1	34.9	41.9	45.2
EventSeg	42.3	40.0	41.3	42.9	51.1
SCS	48.1	41.8	**43.2**	45.1	51.6
OT-SRE	**49.5**	**50.0**	42.7	**52.2**	**52.4**

Table 4: Model performance (F1 scores) for monolingual settings in mSubEvent.

Model	Danish	Spanish	Turkish	Urdu
mBERT				
PLM	23.6	22.6	13.5	11.7
EventSeg	29.0	32.2	16.5	16.4
SCS	**34.6**	36.4	18.9	19.9
OT-SRE	33.1	**37.1**	**19.0**	**27.4**
XLMR				
PLM	25.1	25.4	17.4	18.4
EventSeg	28.5	31.3	20.9	21.4
SCS	41.2	33.7	19.3	22.5
OT-SRE	**42.8**	**34.4**	**22.6**	**26.0**

Table 5: Model performance (F1 scores) for cross-lingual learning settings in mSubEvent using English as the source languages. The languages in each column indicates the target languages.

(respectively) in the literature. We run the code for **SCS** (Trong et al., 2022) and **EventSeg** (Wang et al., 2021) from the original papers to obtain their performance for IC and HiEve (respectively) for completeness.

Performance Comparison: Table 3 presents the performance of the models on the test data of HiEve and IC. To be comparable with previous work (Glavaš et al., 2014; Trong et al., 2022), our model is trained for all the four relation labels in HiEve (i.e., including `COREF`); however, the performance for comparison is only measured according to the F1 scores of the subevent relations, i.e., `PARENT-CHILD`, `CHILD-PARENT`, and their micro-average. The most important observation from the table is that the proposed model OT-SRE significantly outperforms all the baseline models ($p < 0.01$) with substantial gaps for both HiEve and IC. In particular, for HiEve, OT-SRE is better than the prior SOTA method SCS by 3% over the average F1 score for subevent relations. OT-SRE is better than the prior SOTA methods for HiEve (i.e., SCS) and IC (i.e., EventSeg) by 3% and 2.7% (respectively) over the average F1 score for subevent relations. This results thus clearly demonstrate the effectiveness of our OT-based approach for graph structure induction to optimize representation learning for SRE.

Multilingual Evaluation: We further evaluate SRE models over multiple languages using the mSubEvent dataset. We focus on the best baselines, i.e., EventSeg and SCS, in Table 3 in this experiment. In addition, for reference, we report the performance of the **PLM** model that directly uses the representation vectors learned by the multilingual PLMs (i.e., in V) to form the overall representations for subevent prediction, i.e., $A = [v_{e_1}, v_{e_2}, \text{max\_pool}(v_1, \ldots, v_n)]$. As such, we first explore monolingual learning settings where models are trained and tested on data of the same language. In particular, Table 4 shows the monolingual performance of the SRE models for five languages in mSubEvent when either mBERT or XLMR is used for multilingual text encoding. As can be seen, OT-SRE is also significantly better than all baseline models over different languages in mSubEvent, thus highlighting the ability to generalize to different languages of the OT-induced graph structures for SRE. Importantly, we find that the performance of the models over mSubEvent is still far from being satisfactory (i.e., much worse than that for HiEve). Future research will have ample opportunities to improve the performance on mSubEvent.

In addition, Table 5 investigates model performance in the cross-lingual transfer learning setting

ID	Model	CP	PC	Avg.
1	**OT-SRE (full)**	**70.3**	**67.4**	**68.9**
2	- OT	67.8	62.2	65.0
3	- Pruning	60.3	65.8	63.1
4	- GCN	64.3	67.6	66.0
5	- OT-GCN	63.7	57.1	60.4
6	- Syntax in OT	69.1	65.7	67.4
7	- Semantic in OT	65.3	66.8	66.1
8	- DP	69.1	67.2	68.2

Table 6: Ablation study on HiEve test data. We report the the performance for `PARENT-CHILD` (PC), `CHILD-PARENT` (CP), and their micro-average.

where models are trained over English training data (i.e., the source language) and directly evaluated on test data of other languages (i.e., the target languages). It is clear from the table that the cross-lingual performance in Table 5 is inferior to the English monolingual performance in Table 4, thus emphasizing the challenge of cross-lingual knowledge transfer for subevent recognition for future work. Finally, Table 5 further demonstrates better ability to learn transferable representations across languages of OT-SRE to yield the best cross-lingual performance for SRE. We attribute this to the advantages of the induced graph structures to represent input texts in OT-SRE that can be more general across languages than the sequential text order in the baseline methods.

Ablation Study: We study the ablated models of OT-SRE to understand the contribution of the designed components in the our model. Table 6 reports the performance over test data of HiEve for the ablation study. In particular, lines 2 and 3 in the table indicate the baselines where the OT component is not included to induce the graph structure T' for input document. Instead, the DP between the event mentions (i.e., in line 2 with **-OT**) or the full dependency graph T (i.e., in line 3 with **- Pruning**) is leveraged as the graph structure for representation learning. As can be seen, both lines 2 and 3 lead to significantly worse performance for ST-SRE, thus demonstrating the importance of the OT component to induce optimal graph structures to represent input texts for SRE.

In addition, in lines 4 and 5, we study variants of OT-SRE that eliminates the GCN component. In particular, in line 4 with **- GCN**, we still employ the OT component to compute the graph structure T'; however, instead of using GCN-induced representations, the overall representation for prediction is computed over PLM-induced representations in V, i.e., $A = [v_{e_1}, v_{e_2}, \text{max\_pool}(v_j|w_j \in T')]$ where the max-pooling is done for the words in the computed graph structure T'. For line 5 with **- OT-GCN**, both the OT and GCN components are removed from OT-SRE. The overall representation is thus also computed with the PLM-induced representations V, i.e., $A = [v_{e_1}, v_{e_2}, \text{max\_pool}(v_j|w_j \in D)]$, using a max-pooling operation over the entire input text D. It is clear from the table that GCN is helpful to learn better representations for SRE as removing it will significantly hurts the performance for OT-SRE in both lines 4 and 5.

Further, line 6 (**- Syntax in OT**) evaluates OT-SRE when syntactic information (i.e., the important scores $syn(w_i)$) is not used to obtain the domain distributions $p(x)$ and $p(y)$ in the OT component. Instead, uniform distributions are leveraged for $p(x)$ and $p(y)$ in this case. Also, for line 7 (**- Semantic in OT**), this variant avoids semantic information with contextual representations in V to compute the transformation cost $C(x, y)$ for OT. Instead, it employs a simple constant cost function $C(x, y) = 1$. As such, the superior performance of OT-SRE over these ablated models shows that both syntactic and semantic information are critical for the OT component to ensure the best performance for OT-SRE. Finally, in line 8 (i.e., **- DP**), our OT-SRE model only includes the two input event mentions/triggers in domain (Y). As such, domain $\mathcal{X}$ for alignment in OT will contain all other words in D, including the words on the dependency path. The worse performance in line 8 shows that only using event mentions as the anchor for OT alignment is not optimal, necessitating dependency paths to provide better starting points to extend to effective graph structures for SRE.

Case Study: We perform a case study to analyze the examples in HiEve that can be successfully predicted by OT-SRE, but fail the baseline without OT (i.e., in line 2 of Table 6 to directly use DP for representation). A major observation in our analysis is that OT-SRE can find important context words beyond the DP to aid subevent prediction. For example, consider the sentence “*Over 90 Palestinians and one Israeli soldier have been* ***killed since*** *Israel launched a massive* ***offensive*** *into the Gaza Strip on June 28.* with “*killed*” and “*offensive*” as the event mentions. While the DP “*killed → launched → offensive* does not provide clear context information to recognize the subevent relation, our OT-SRE is able to align the DP with the word "*since*" to facilitate SRE. A similar example can be found in "*No one has been arrested over Sunday's* ***attack*** *in Kabul and the Taliban have denied any involvement. Arsala Rahmani has been* ***killed*** *by enemies of Afghanistan. Both NATO and the US embassy in Kabul have also condemned the* ***assassination***." with the event mentions "*attack*" and "*killed*". The important context word "*assassination*" does not belong to the DP between the event mentions, but it is successfully included in the graph structure by OT-SRE for correct prediction.

5 Related Work

Early methods for SRE have exploited various contextual features for input texts (i.e., feature engineering) for machine learning models (Glavaš et al., 2014; Araki et al., 2014; Aldawsari and Finlayson, 2019). To alleviate feature engineering, recent works have explored deep learning models to induce representations for SRE from data, introducing joint inference with temporal relations (Wang et al., 2020; Zhou et al., 2020) and large PLMs (Yao et al., 2020; Wang et al., 2021; Trong et al., 2022). Existing datasets for SRE include HiEve (Glavaš et al., 2014), IC (Hovy et al., 2013; Araki et al., 2014), and RED (O'Gorman et al., 2016). However, none of such methods and datasets considers graph structure induction for input texts and multilingual learning for SRE as we do. Regarding related work on event-event relation extraction, we also note recent studies for other types of relations between events, including causal (Caselli and Vossen, 2017; Zuo et al., 2020; Man et al., 2022; Tran Phu and Nguyen, 2021), coreference (Nguyen et al., 2016; Choubey et al., 2020; Minh Tran et al., 2021; Phung et al., 2021), and temporal (Ning et al., 2017; Tran Phu et al., 2021) relations. Finally, optimal transport has also been recently used to solve NLP problems (Veyseh and Nguyen, 2022; Guzman-Nateras et al., 2022); however, none of previous work has employed OT for subevent relation extraction as we do.

6 Conclusion

We present a novel method for subevent relation extraction that leverages optimal transport to induce effective graph structures for input texts to improve representation learning. The graph structure representation is able to directly capture important context words and their connections to facilitate SRE. In addition, we introduce the first multilingual dataset for SRE that provides human annotation for five languages with high quality. Extensive experiments demonstrate the effectiveness of our method with state-of-the-art performance on different datasets and learning settings. Our new dataset also offers ample opportunities for future research. In the future, we plan to extend our method and dataset to other event-event relations.

Ethical Considerations

In this work we present a dataset annotated over the publicly accessible articles of `wikipedia.org`. Complying with the discussion presented by Benton et al. (2017), research with human subject information is exempted from the required full Institutional Review Board (IRB) review if the data is already available from public sources (as with Wikipedia) or if the identity of the subjects cannot be recovered.

Acknowledgement

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112 and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning. This research is also based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the Better Extraction from Text Towards Enhanced Retrieval (BETTER) Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ARO, ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This document does not contain technology or technical data controlled under either the U.S. International Traffic in Arms Regulations or the U.S. Export Administration Regulations.

References

Mohammed Aldawsari and Mark Finlayson. 2019. Detecting subevents using discourse and narrative features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages

4780–4790, Florence, Italy. Association for Computational Linguistics.

Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization. In *Text Summarization Branches Out*, pages 104–111, Barcelona, Spain. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2013. Event-centered information retrieval using kernels on event graphs. In *Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing*, pages 1–5, Seattle, Washington, USA. Association for Computational Linguistics.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3678–3683, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pankaj Gupta, Subburam Rajaram, Hinrich Schütze, and Thomas Runkler. 2019. Neural relation extraction within and across sentence boundaries. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 323–330, Seattle, Washington. Association for Computational Linguistics.

Hieu Minh Tran, Duy Phung, and Thien Huu Nguyen. 2021. Exploiting document structures and cluster consistencies for event coreference resolution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4840–4850, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Thien Huu Nguyen, Adam Meyers, and Ralph Grishman. 2016. New york university 2016 system for kbp event nugget: A deep learning approach. In *Proceedings of Text Analysis Conference (TAC)*.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. 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. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Gabriel Peyre and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. In *Foundations and Trends in Machine Learning*.

Duy Phung, Hieu Minh Tran, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Learning cross-lingual representations for event coreference resolution with multi-view alignment and optimal transport. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 62–73, Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316, Florence, Italy. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Minh Tran Phu, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Fine-grained temporal relation extraction with ordered-neuron LSTM and graph convolutional networks. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 35–45, Online. Association for Computational Linguistics.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Hieu Man Duc Trong, Nghia Trung Ngo, Linh Van Ngo, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Amir Pouran Ben Veyseh and Thien Nguyen. 2022. Word-label alignment for event detection: A new perspective via optimal transport. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 132–138, Seattle, Washington. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium*.

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, Online. Association for Computational Linguistics.

Haoyu Wang, Hongming Zhang, Muhao Chen, and Dan Roth. 2021. Learning constraints and descriptive segmentation for subevent detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5216–5226, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly Supervised Subevent Knowledge Acquisition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5345–5356, Online. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

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

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.