Weakly Supervised Subevent Knowledge Acquisition

Wenlin Yao, Zeyu Dai
Maitreyi Ramaswamy, Ruihong Huang
Texas A&M University
College Station, TX
huangrh@cse.tamu.edu

Bonan Min
Raytheon BBN Technologies
10 Moulton Street, Cambridge, MA
bonan.min@raytheon.com

Abstract

Subevents widely exist in event descriptions. Subevent knowledge is useful for discourse analysis and event focused applications. Acknowledging the scarcity of subevent knowledge, we propose a weakly supervised approach to extract subevent tuples from text and build the first large scale subevent knowledge base. We first obtain the initial set of event pairs that are likely to have the subevent relation, following two observations that 1) subevents are temporally contained by the parent event, and 2) the definitions of the parent event can be used to further guide the identification of subevents. Then, we collect rich weak supervision using the initial seed subevent pairs to train a contextual BERT classifier and applies it to identify new subevent pairs. The evaluation showed that the acquire subevent pairs (142k) are of high quality (88% of accuracy) and cover a wide range of event types. The acquired subevent knowledge base has been shown useful for discourse analysis.

1 Introduction

Subevents, which elaborate and expand an event, widely exist in event descriptions. Knowing typical subevents of an event can help with analyzing several discourse relations (such as expansion and temporal relations) between text units. Furthermore, knowing typical subevents of an event is important for understanding the internal structure of the event (what is the event about?) and its properties (e.g., is this a violent or peaceful event?), and therefore has great potential to benefit event detection, event tracking, event visualization, event summarization and extreme event management among many other applications.

While being in high demand, little subevent knowledge can be found in existing knowledge bases. For instance, the widely used commonsense knowledge base ConceptNet (Speer et al., 2017) only contains around 30K individual subevent relation tuples. In this paper, we aim to extract subevent knowledge from text and build the first subevent knowledge base covering a large number of commonly seen events and their rich subevents.

Little research has focused on identifying the subevent relation between two events in a text. A few datasets annotated with subevent relations (Glavaš et al., 2014; Araki et al., 2014; O'Gorman et al., 2016) exist, but they are extremely small and usually contain dozens to one or two hundred documents. Subevent relation classifiers trained on these small datasets are not suitable to use to extract subevent knowledge from text, considering that subevent relations can be described in dramatically different contexts depending on topics and particular pairs of events. But, manually annotating a large corpus for subevent relationships exhaustively is too laborious.

We propose to conduct weakly supervised learning and train a wide-coverage contextual classifier to acquire new diverse event pairs of the subevent relation from text. Specifically, we create weak supervision in the form of event pairs that are likely to be in the subevent relation. Then, we populate these event pairs in a large text corpus to quickly label many (but noisy) training instances and train a contextual classifier for identifying the subevent relation in text.

We start by creating rich **weak supervision**. Considering that no existing knowledge base contains enough event pairs of the subevent relation, we aim to identify the initial set of subevent pairs from a text corpus. With no contextual classifier, it is difficult to extract subevent relation pairs because subevent relations are rarely stated explic-

itly. Instead, we propose a novel two-step approach to obtain the initial set of event pairs of the subevent relation following two key observations that (1) the subevents are temporally contained by the parent event, therefore, we can use common temporal relation indicators to identify candidate subevent relations[1], and (2) the definition of the parent event is useful to further guide the identification of subevents.

Specifically, we first use several preposition patterns indicating the temporal relation *contained_by* (e.g., event_A *during* event_B) to identify candidate subevent pairs, and then conduct an event definition-guided semantic consistency check to remove spurious subevent pairs that often include two temporally overlapped but semantically incompatible events. For example, a news article may report a *bombing* event that happened in parallel during a *festival*, but the *bombing* event is not semantically compatible with and therefore is not a subevent of the event *festival*, as informed by the common definition of *festival*:

> *A festival is an organized series of celebration events, or an organized series of concerts, plays, or movies, typically one held annually in the same place.*

The semantic consistency check clusters events based on similarities between events and similarities between the definition of a parent event and its subevents, and then identifies spurious subevent pairs as event pairs that have two events in two different event clusters.

Next, we populate the identified event pairs in a text corpus to obtain hundreds of thousands of sentences containing an event pair, which are used to train a contextual classifier that can recognize the subevent relation in text. We train the contextual subevent relation classifier by fine-tuning the pre-trained BERT model (Devlin et al., 2019), inspired by recent success of BERT fine-tuned models applied to various NLP tasks. We then apply the contextual BERT classifier to identify new event pairs that have the subevent relation.

We have acquired 30K seed subevent pairs and 112K new subevent pairs, much bigger than subevent pairs (30K) that exist in ConceptNet. In the learned event pairs, 8,035 unique events have shown as a parent event, covering over half of the events annotated in two commonly used datasets for event extraction. Each event is associated with 18 subevents on average. Intrinsic evaluation demonstrates that the learned subevent tuples are of high quality (88% of accuracy) and valuable for event ontology building and exploitation.

The learned subevent knowledge has been shown useful for identifying subevent relations in text, including both intra-instance and cross-sentence cases. Further, when incorporated into a recent neural discourse parser, the learned subevent knowledge is beneficial to discourse parsing and has improved predictions on several types of implicit discourse relations including expansion, temporal and even comparison relations. The learned subevent knowledge will be publicly released for future applications.

2 Related Work

Subevent identification: Only a few studies have focused on identifying subevent relations in text. (Araki et al., 2014) built a logistic regression model to classify the relation between two events into full coreference (FC), subevent parent-child (SP), subevent sister (SS), and no coreference (NC). They improved the prediction of SP relations by performing SS prediction first and using SS prediction results in a voting algorithm. (Glavaš and Šnajder, 2014) trained a logistic regression classifier using a range of lexical and syntactic features and then used Integer Linear Programming (ILP) to enforce document-level coherence for constructing coherent event hierarchies from news. Recently, (Aldawsari and Finlayson, 2019) outperformed previous models on two datasets using a linear SVM classifier, by introducing several new features, in particular discourse features (i.e., rhetorical structure, reported speech, etc.) and narrative features (i.e., non-major mentions).

In addition, we are aware of previous research that study subevents specifically for social media (e.g., Twitter) applications (Shen et al., 2013; Meladianos et al., 2015; Pohl et al., 2012), in terms of both its definition of subevents and methodologies. For example, in previous research by (Shen et al., 2013), a subevent is defined as a topic that is discussed intensively in the Twitter stream for a short period of time before fading away. Accordingly, the subevent detection method relies on modeling the "burstiness" and "cohesiveness" properties of

[1]While subevents are also spatially contained by the parent event, we did not use this observation to identify candidate subevent relations because spatial relations are not frequently stated in text.

tweets in the stream.

Subevent Knowledge Acquisition: Due to the generalizability issue of supervised contextual classifiers trained on small annotated data, pilot research on subevent knowledge acquisition relies on heuristics (Badgett and Huang, 2016). The recent work (Bosselut et al., 2019; Sap et al., 2019) use generative language models to acquire subevent knowledge among many other types of commonsense knowledge.

Identification and Acquisition of other Event Relations: Compared to relatively little research devoted to subevent identification and acquisition, significantly more research has been done for identifying and extracting several other types of event relations, especially temporal relations (Pustejovsky et al., 2003; Chklovski and Pantel, 2004; Chambers and Jurafsky, 2008; Bethard, 2013; Llorens et al., 2010; D'Souza and Ng, 2013; Chambers et al., 2014; Pichotta and Mooney, 2016; Granroth-Wilding and Clark, 2016; Wang et al., 2017) and causal relations (Girju, 2003; Bethard and Martin, 2008; Riaz and Girju, 2010; Do et al., 2011; Riaz and Girju, 2013; Mirza and Tonelli, 2014, 2016).

3 Event Representations

We aim to identify event pairs that unambiguously show the subevent relation out of context. However, an event word can refer to a general type of events or more than one type of events, and therefore has varied meanings depending on contexts. To make individual events expressive and self-contained, we find and attach arguments to each event word and form event phrases. Specifically, we consider both verb event phrases and noun event phrases. We further require that at least one argument is included in an event pair which may be attached to the first or the second event. In other words, we do not consider event pairs in which neither event has an argument.

Verb Event Phrases: To ensure good coverage of regular event pairs, we consider all verbs[2] as event words except possession verbs[3]. The thematic patient of a verb refers to the object being acted upon and is essentially part of an event, therefore, we first consider the patient of a verb in forming an event phrase[4]. The agent is also useful to specify an event especially for an intransitive verb event, which does not have a patient. Therefore, we include the agent of a verb event in an event phrase if its patient was not found. The patient or agent of a verb is identified using dependency relations[5]. If neither a patient nor an agent was found, we include a preposition phrase (a preposition and its object) that modifies a verb in the event representation to form an event phrase. Example verb event phrases are "agreement be *signed*" and "*occupy* territory".

Noun Event Phrases: We include a preposition phrase (a preposition and its object)[6] that modifies a noun event in the event representation to form a noun event phrase. Example noun event phrases are "*ceremony* at location" and "*attack* on troops".

Note that many noun words do not refer to an event, therefore, we apply two strategies to quickly compile a list of noun event words. First, we obtain a list of derivative event nouns[7] (5028 event nouns) by querying each noun in WordNet (Miller, 1995) and checking if its root word form has a verb sense. Second, we identify five intuitive textual patterns, e.g., *participate in* EVENT[8], and extract their prepositional direct objects as potential noun events. We rank extractions first by the number of times they occur with these patterns and then by the number of unique patterns they occur with. We next quickly went through the top 5,000 nouns and manually removed non-event words, which results in 3154 noun event words.

Event Phrase Generalization: Including arguments into event representations generates specific event phrases though. In order to obtain generalized event phrase forms, we replace specific name arguments with their named entity types (Manning et al., 2014). We also replace personal pronouns with their type PERSON.

[2]We used POS tags to detect verb events.

[3]We determined that possession verbs, such as "own", "have" and "contain", mainly express the ownership status so we discarded these event phrases.

[4]In particular, we require a light verb (e.g., do, make, take etc.) to have a direct object because light verbs have little semantic content of their own.

[5]We use Stanford dependency relations (Manning et al., 2014). We identify the patient as the direct object of an active verb or the subject of a passive verb; we identify the agent as the subject of an active verb or the object of preposition *by* modifying a passive verb.

[6]We first consider an object headed by the preposition *of*, then an object headed by the preposition *by*, lastly an object headed by any other preposition.

[7]Derivative nouns ending with suffixes -er, -or are discarded.

[8]The five patterns are: *participate in* EVENT, *involve in* EVENT, *engage in* EVENT, *play role in* EVENT and *series*

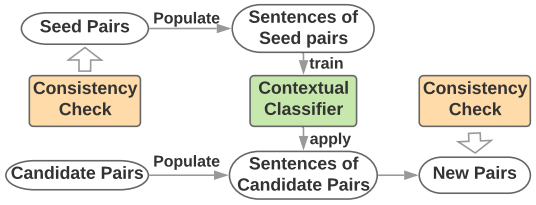

Figure 1: Overview of the Subevent Knowledge Acquisition System

4 Weak Supervision

Figure 1 shows the overview the weakly supervised learning system for subevent knowledge acquisition. We first identify seed event pairs that are likely to be in the subevent relation and then populate seed pairs in a large text corpus to quickly generate many subevent relation instances, which will be used to train the subevent relation classifier (Section 5). The trained contextual classifier will be further used to identify new event pairs of the subevent relation in text (Section 6). We use English Gigaword (Napoles et al., 2012) containing ten million news articles as the text corpus.

With no contextual classifier for subevent relation identification, we first use several temporal relation patterns (e.g., event_A *during* event_B) to identify candidate subevent pairs, and then conduct a definition-guided semantic consistency check to remove spurious subevent pairs that only show the temporal containment relation but not the subevent relation.

4.1 Seed Event Pair Identification

We use six preposition patterns (i.e., during, in, amid, throughout, including, and within) to extract candidate seed event pairs. Specifically, we use dependency relations to identify preposition patterns, e.g., *prep_during* and extract the governor and dependent word of each pattern. We then check whether each word is an event[9]. If yes, we form an event phrase for each event (Section 3) and obtain an event pair. To select seed subevent pairs, we consider event pairs that co-occur with at least two different patterns for at least three times. In total, we identified around 43K candidate seed pairs from the Gigaword corpus.

of EVENT.

[9]Note that we consider any verbs and only nouns that are in our noun event list.

4.2 Definition-Guided Semantic Check

Many candidate seed pairs identified by the preposition patterns only show the temporal *contained_by* relation. In order to remove such spurious subevent pairs, we present a semantic consistency check approach guided by event definitions that clusters event phrases into groups so that any two event phrases within a group are semantically compatible. The intuition is that the definition of a parent event word describes important aspects of the event's meanings and signifies its potential subevents. For example, based on the definition of *festival*, events related to *"celebrations"*, such as *ceremony being held* and *set off fireworks*, are likely to be correct subevents of *festival*; however, *bomb explosion* and *people being killed* may be distinct events that only happen temporally in parallel with *festival*.

Clustering Event Phrases with Graph Propagation: The clustering method we use here is a graph propagation algorithm called Speaker-Listener Label Propagation Algorithm (SLPA) (Xie et al., 2011). SLPA has been shown one of the best algorithms for detecting overlapping clusters (Xie et al., 2013), which is preferred because an event can be a subevent of more than one event. For instance, *people being injured* can be a subevent of a conflict event (e.g., *protest*) or a disaster event (e.g., *earthquake*). In addition, SLPA is a self-adaptation model and can automatically converge to the optimal number of clusters, with no pre-defined number of clusters needed. We run the algorithm for 60 iterations.

Given a set of event pairs needing the semantic consistency check, we construct an undirected graph $G(V, E)$, where each node in V represents a unique event phrase. We connect event phrases with two types of weighted edges. First, for each given event pair in the subevent relation, we create an edge of weight 1.0 between the parent event and the child event. Second, we create an edge between any two event phrases if their similarity[10] is greater than a certain threshold[11], and the edge weight is their similarity score. If two event phrases are already connected because they are a subevent pair, we add their similarity score to the edge weight.

[10]To calculate the similarity between two event phrases, we pair each word from one event phrase (either the event word or an argument) with each word from the other event phrase and calculate the similarity between two word embeddings, then the similarity between two event phrases is the average of their word pair similarities. We used word2vec word embeddings.

[11]We used 0.3 as the similarity threshold.

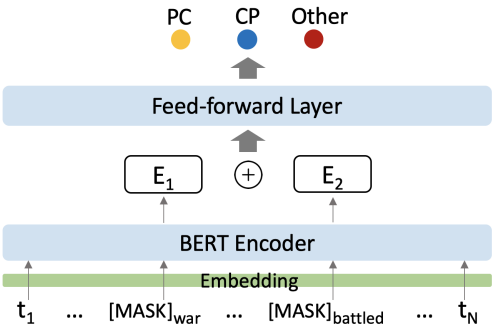

Figure 2: BERT-based Contextual Classifier

Next, we incorporate event definitions by adding new nodes and new edges to the graph. Specifically, for each event phrase that has children events in the candidate subevent pairs, we create a new node for its event word representing the event word definition. If the event word has multiple meanings and therefore multiple definitions, we consider at most 5 definitions retrieved from WordNet (Miller, 1995) and create one node for each definition, assuming each definition of the parent event will attract different types of children events. Then, we connect each definition node of a parent event with its corresponding children events, if their similarity[12] is over the same similarity threshold used previously.

5 The Contextual Classifier Using BERT

Recently, BERT (Devlin et al., 2019) pretrained on massive data has achieved state-of-the-art performance on various NLP tasks. We fine-tune a pretrained BERT model to build the contextual classifier for subevent relation identification.

BERT model is essentially a bi-directional Transformer-based encoder that consists of multiple layers where each layer has multiple attention heads. Formally, given a sentence with N tokens, each attention head transforms a token vector t_i into query, key, and value vectors q_i, k_i, v_i through three linear transformers. Next, for each token, the head calculates the self-attention scores for all other tokens of the input sentence against this token as the softmax-normalized dot productions between the query and key vectors. The output o_i of each attention head is a weighted sum of all value vectors:

$$o_i = \sum_{j=1}^{N} w_{ij} v_j, \; w_{ij} = \frac{\exp(q_i^T k_j)}{\sum_{l=1}^{N} \exp(q_i^T k_l)}$$

[12]The similarity between a definition node and a child event is calculated by exaustively pairing each non-stop word from the definition and each word from the child event phrase and taking the average of word pair similarities.

In this way, we can obtain N contextualized embeddings for all words $\{w_i, o_i\}_{i=1}^{N}$ using the BERT model. Figure 2 shows the overall structure of the classfier. To enforce the BERT encoder to look at context information other than the two event trigger words of a subevent pair, e.g., *war, person battle*, we replace the two event trigger words in a sentence with a special token `[MASK]` as the original BERT model did in masking. The contextualized embeddings at two event triggers' positions (two `[MASK]`'s positions) are concatenated and then fed into a feed-forward neural network with a softmax prediction layer for three-way classification, i.e., forward parent-child relationship (PC), backward parent-child relationship (CP), and no subevent relation (Other).

In our experiments, we use the pretrained $\text{BERT}_{\text{base}}$ model provided by (Devlin et al., 2019) with 12 transformer block layers, 768 hiddent size and 12 self-attention heads[13]. We train the classifier using cross-entropy loss and Adam (Kingma and Ba, 2015) optimizer with initial learning rate 1e-5, 0.5 dropout, batch size 16 and 3 training epochs.

Negative Training Instances: High-quality negative training instances that can compete with positive instances are important to empower the classifier to distinguish subevent relations from non-subevent relations. We include two types of negative instances to fine-tune the contextual BERT classifier. First, we randomly sample sentences that contain an event pair different from any seed pair or candidate pair (Section 6.1) as negative instances. We sample such negative sentences equal to five times of positive sentences, considering that most sentences in a corpus do not contain a subevent relation. Second, we observe that the subevent relation is often confused with temporal and causal event relations because a subevent is strictly temporally contained by its parent event. Therefore, to improve the discriminative capability of the classifier, we also include sentences containing temporally or causally related events as negative instances. Specifically, we apply a similar strategy - using patterns[14] to extract temporal and causal event pairs and then populate these pairs to collect sentences that contain a temporal or causal event pair. Event pairs that co-occur with temporal or

[13]Our implementation was based on https://github.com/huggingface/transformers.

[14]Three temporal patterns - "following", "before", "after" and seven causal patterns - "lead to", "result in", "result from", "cause", "cause by", "due to", "because of" are used.

causal patterns for at least three times are selected for population[15]. In total, we obtained around 1.8 million negative training instances.

6 Identifying New Subevent Pairs

We next apply the contextual BERT classifier to identify new event pairs that express a subevent relation. It is unnecessary to test on all possible pairs of events since two events that co-occur in a sentence often have no subevent relation. In order to narrow down the search space, we first identify candidate event pairs that are likely to have the subevent relation, and then apply the contextual classifier to examine instances of each candidate event pair to determine valid subevent pairs.

6.1 Candidate Event Pairs

We consider two types of candidate event pairs. First, the preposition patterns used to identify seed subevent pairs are again used to identify candidate event pairs, but with less strict conditions. Specifically, we consider event pairs that co-occur with any pattern for at least two times as candidate event pairs. In this way, the approach yields 1.4 million candidate event pairs from the Gigaword corpus.

Second, if an event pair of the subevent relation appears in a sentence, it is common to observe other subevents of the same parent event in the surrounding context. Therefore, we collect sentences that contain a seed subevent pair (*seed_parent, seed_subevent*), and identify additional subevents of the same parent event in the two preceding and two following sentences. Furthermore, we observe that the additional subevents often share the agent or patient with the event *seed_subevent*. As a consequence, we only consider such event phrases found in the surrounding sentences and pair them with the event *seed_parent* to create new candidate event pairs. Using this method, we extracted around 89K candidate event pairs from the Gigaword corpus.

6.2 New Subevent Pair Selection Criteria

We identify a candidate event pair as a new subevent pair only if the majority of its sentential contexts, correctly more than 50% of contexts, were consistently labeled as one specific subevent relation (PC or CP) by the BERT classifier. In addition, we require that at least three sentential instances of an event pair have been labeled as containing the subevent relation.

[15]In total, we collected 63K temporally related event pairs and 61K causally related event pairs.

Pairs to populate	P/R/F1
All seed pairs	44.9/25.3/32.4
+ Consistency check	55.9/26.2/35.7

Table 1: Performance of the Contextual Classifier on the RED dataset: micro-average Precision/Recall/F1-score (%) over PC and CP categories.

6.3 Statistics of Acquired Knowledge

Among around 1.5M candidate event pairs, 298K are identified as new subevent pairs. We also apply the definition-guided consistency check to filter newly identified pairs, which results in 112K subevent pairs after removing spurious pairs. To sum up, the full weakly supervised learning process acquires 142K subevent pairs, including 30K seed pairs and 112K classifier identified pairs, with 8,035 unique events shown as parent events. Each parent event is associated with 17.7 children events on average.

7 Evaluation

7.1 Precision of the Contextual Classifier

The contextual classifier is the key component of our learning approach. We evaluate the performance of the BERT contextual classifier on identifying both forward (PC) and backward (CP) subevent relations against all the other event-event relations (e.g., temporal, causal or coreference relations, etc.) on the Richer Event Description (RED) corpus (O'Gorman et al., 2016) comprehensively annotated with rich event-event relations. Since the contextual classifier mainly performs at the sentence level, we only consider event pairs in the RED dataset that co-occur in the same sentence.

Table 1 shows the comparisons between two training settings - the BERT classifier either trained on sentences identified by candidate seed pairs (43K in total), or trained on sentences identified by seed pairs (30K in total) passing the definition-guided semantic consistency check. We can see that conducting the semantic check improves the precision of the trained classifier by 11% with no loss on recall. Overall, without using any annotated data, the classifier achieves the precision of 56%.

7.2 Quality of Acquired Subevent Pairs

We randomly sampled around 1% of acquired subevent pairs, including 300 from seed subevent

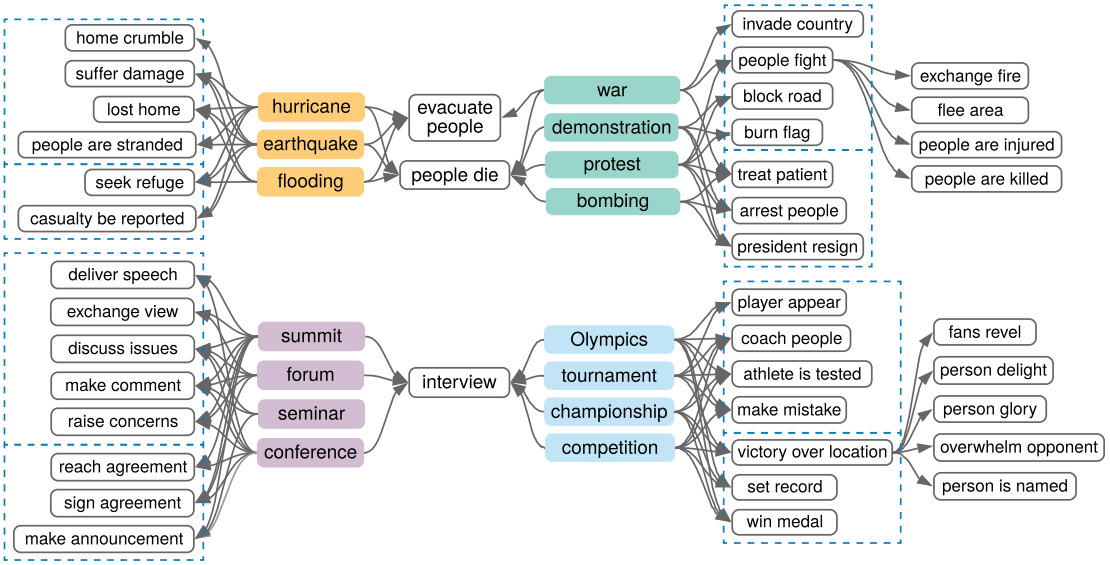

Figure 3: Example Subevent Knowledge Graph (→ denotes Parent→Child relation). Four colors indicate four groups of parent events where parent events in the same group commonly share children events. Children events circled by the same blue dash box describe a stage of development of parent events.

Datasets	Events	Coverage
ACE (Doddington et al., 2004)	987	0.63
KBP (Ellis et al., 2015)	1539	0.57

Table 2: Event Coverage on ACE and KBP datasets

	intra-sentence		cross-sentence	
	PC	CP	PC	CP
RED	290	240	303	112
HiEve	254	236	1536	1581

Table 3: Statistics of RED and HiEve Datasets

pairs and 1,216 from newly learned subevent pairs, and asked two human adjudicators to judge whether a PC (Parent-Child) or CP (Child-Parent) relation exists between two events. The two annotators achieved the kappa inter-agreement score of 0.65 on annotating 200 common event pairs (the remaining subevent pairs were evenly split between two annotators). According to human annotations, the accuracy of seed pairs is 91.7% and the accuracy of newly learned event pairs is 87.1%, with the overall accuracy of 88%.

7.3 Coverage of Acquired Subevent Pairs

To see whether the acquired subevent pairs have good coverage of broad event types, we compare the unique parent events (8,035 events) of the acquired event pairs with events annotated in two datasets, ACE (Doddington et al., 2004) and KBP (Ellis et al., 2015), both with rich event types annotated and being commonly used for event extraction evaluation. Shown in Table 2, 63% of events in ACE and 57% of events in KBP match and have children events in the acquired subevent pairs.

In addition, we compare our learned 142K subevent pairs with the 30K ConceptNet subevent pairs. Interestingly, the two sets only have 240 event pairs in common, which shows that our learning approach extracts subevent pairs from real texts that are often hard to discover by crowd sourcing, the approach used by ConceptNet.

7.4 Example Subevent Knowledge Graph

Figure 3 shows an example knowledge graph constructed using acquired subevent pairs, which illustrates our two striking observations. First, we can draw a partition of the event space at multiple granularity levels by grouping events based on subevents they share, e.g., the upper and the lower sections of the graph illustrate distinct events sharing no subevent, and each section consists of two groups that share subevents extensively within a group and share fewer subevents between groups. Second, subevents encode event semantics and reveal different development stages of the parent events, e.g., subevents of natural disaster events show two disaster stages *response* and *recovery*.

7.5 Improving In-text Subevent Relation identification on Two Datasets

To find out whether the learned subevent knowledge can be used to improve contextual subevent relation identification, we conducted experiments on

Method	Macro	Acc	Comparison	Contingency	Expansion	Temporal
Base Model	50.8/47.8/49.0	56.42	43.8/39.0/41.3	44.7/51.3/47.8	66.6/65.7/66.2	48.2/35.0/40.6
+ Subevent (ours)	53.0/48.2/50.1	56.98	52.3/37.8/43.9	44.3/47.8/46.0	64.4/68.6/66.4	51.0/38.5/43.9

Table 4: Multi-class Classification Results on the PDTB dataset. We report accuracy (Acc), macro-average (Macro) Precision/Recall/F1-score (%) over four implicit discourse relation categories (Comparison, Contingency, Expansion and Temporal) as well as performance on each category.

Method	RED	HiEve
Train and test on intra-sentence event pairs		
Basic Classifier	64.6/47.6/54.8	43.9/45.3/44.6
+ subevent links	68.3/47.7/56.2	**46.2**/45.3/45.8
+ avg children emb	**69.1**/47.3/56.1	46.0/46.2/46.1
+ both	68.7/**48.3/56.7**	45.5/**46.9/46.2**
Train and test on cross-sentence event pairs		
Basic Classifier	54.9/41.9/47.5	29.0/29.1/29.1
+ subevent links	56.0/44.2/49.4	**30.9**/31.4/**31.2**
+ avg children emb	56.8/**44.7/50.1**	30.5/31.2/30.9
+ both	**57.4**/44.1/49.9	30.5/**31.8**/31.1

Table 5: 5-fold Cross-validation Results on RED and HiEve dataset: Precision/Recall/F1-score (%) micro-average performance over PC and CP categories.

the RED and HiEve[16] (Glavaš et al., 2014) datasets, both annotated with subevents. In our experiments, we consider intra-sentence event pairs and cross-sentence event pairs separately, with dataset statistics shown in Table 3. We conduct document level 5-fold cross-validation to evaluate our models.

We train the same BERT model using RED or HiEve annotations. Note that for cross-sentence event pairs, we simply append the second sentence after the first sentence by inserting in between the special token `[SEP]` used in the original BERT.

We propose two methods to incorporate learned subevent knowledge into the contextual subevent relation classifier. **1) Subevent links** For each event pair, we check if it is learned as a subevent pair and encode this information in a vector. **2) Averaged children event embedding.** Subevents elaborate and expand the parent event, therefore, children events of a given parent can be used to enhance the representation of the parent event. To do so, we retrieve children events of each event and calculate the weighted mean of word embeddings considering all the words in children events.

Finally, the subevent link vector or/and averaged subevent embeddings are concatenated with two event representations extracted from the BERT encoder (E1 and E2 in Figure 2) for a three-way classification. Results are shown in Table 5. We can see that compared to basic BERT classifier, incorporating learned subevent knowledge in either way achieves better performance on both precision and recall, for both intra-sentence and cross-sentence cases.

7.6 Improving Discourse Parsing

We expect subevent knowledge to be useful for analyzing discourse relations between two text units in general because subevent descriptions often elaborate and provide a continued discussion of a parent event introduced earlier in text. We chose a recent discourse parsing system proposed by Dai and Huang (2019) which can easily incorporate external event knowledge as a regularizer into a two-level hierarchical BiLSTMs (Base Model) for paragraph-level discourse parsing. The experimental setting is exactly the same as in (Dai and Huang, 2019)[17].

Table 4 reports the performance of discourse parsing on PDTB 2.0 (Prasad et al., 2008). Incorporating the acquired subevent pairs (142K) into the Base Model improves the overall macro-average F1-score and accuracy by 1.1 and 0.6 points respectively, which is non-trivial considering the challenges of implicit discourse relation identification. The performance improvements are on multiple categories, including temporal relations, expansion and even comparison categories.

8 Conclusion

We have presented a novel weakly supervised learning framework for acquiring subevent knowledge and built the first large scale subevent knowledge base containing 142K subevent tuples. Evaluation showed that the acquire subevent pairs are of high quality (88% of accuracy) and cover a wide range of event types. The acquired subevent knowledge

[16]HiEve annotated 3,200 event mentions and their subevents as well as coreference relations in 100 documents. We first extended the subevent annotations using transitive closure rules and coreference relations (Glavaš et al., 2014; Aldawsari and Finlayson, 2019), for instance, $e_1 > e_2$ and $e_2 > e_3$ implies $e_1 > e_3$, and $e_1 > e_2$ and $e_2 \equiv e_3$ implies $e_1 > e_3$, where $>$ and $\equiv$ denote Parent-Child subevent relation and coreference relation respectively.

[17]The source code was provided by the first author of (Dai and Huang, 2019).

base is useful for modeling event semantics and event space exploration and has been demonstrated to be useful for in-text subevent relation identification and discourse analysis in general. In the future, we would like to explore the uses of the subevent knowledge base for event-oriented applications such as event detection and event tracking.

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.

Jun Araki, Zhengzhong Liu, Eduard H Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *LREC*, pages 4553–4558.

Allison Badgett and Ruihong Huang. 2016. Extracting subevents via an effective two-phase approach. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 906–911.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 177–180. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for knowledge graph construction. In *Association for Computational Linguistics (ACL)*.

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 and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*, volume 94305, pages 789–797. Citeseer.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. 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 2974–2985.

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.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, page 1. Lisbon.

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In *HLT-NAACL*, pages 918–927.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and reults. In *Proceedings of the TAC KBP 2015 Workshop*, pages 16–17.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2014. Constructing coherent event hierarchies from news stories. In *Proceedings of TextGraphs-9: the workshop on Graph-based Methods for Natural Language Processing*, pages 34–38.

Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. 2014. Hieve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI*, pages 2727–2733.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–60.

P. Meladianos, G. Nikolentzos, F. Rousseau, Y. Stavrakas, and M. Vazirgiannis. 2015. Degeneracy-based real-time sub-event detection in twitter stream. In *Proceedings of the 9th AAAI international conference on web and social media (ICWSM)*, pages 248–257.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING*, pages 2097–2106.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th International Conference on Computational Linguistics*, pages 64–75.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. 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.

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.

Karl Pichotta and Raymond J Mooney. 2016. Learning statistical scripts with lstm recurrent neural networks. In *AAAI*, pages 2800–2806.

Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. 2012. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st International Conference on World Wide Web*, pages 683–686. ACM.

R. Prasad, N. Dinesh, Lee A., E. Miltsakaki, L. Robaldo, Joshi A., and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *lrec2008*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 361–368. IEEE.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*. Citeseer.

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

Chao Shen, Fei Liu, Fuliang Weng, and Tao Li. 2013. A participant-based approach for event summarization using twitter streams. In *HLT-NAACL*, pages 1152–1162.

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

Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 57–67.

Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43.

Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. 2011. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 344–349. IEEE.

Acknowledgements

This work was supported by DARPA/I2O and U.S. Army Research Office Contract No. W911NF-18-C-0003 under the World Modelers program. The views, opinions, and/or findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Department of Defense or the U.S. Government. 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.