Detecting Subevents using Discourse and Narrative Features

Mohammed Aldawsari & Mark A. Finlayson
School of Computing and Information Sciences
Florida International University
Miami FL, 33199
{malda021,markaf}@fiu.edu

Abstract

Recognizing the internal structure of events is a challenging language processing task of great importance for text understanding. We present a supervised model for automatically identifying when one event is a subevent of another. Building on prior work, we introduce several novel features, in particular discourse and narrative features, that significantly improve upon prior state-of-the-art performance. Error analysis further demonstrates the utility of these features. We evaluate our model on the only two annotated corpora with event hierarchies: HiEve and the Intelligence Community corpus. No prior system has been evaluated on both corpora. Our model outperforms previous systems on both corpora, achieving 0.74 BLANC F_1 on the Intelligence Community corpus and 0.70 F_1 on the HiEve corpus, respectively a 15 and 5 percentage point improvement over previous models.

1 Introduction

An event is something that occurs in a certain place at a certain time (Pustejovsky et al., 2003). Understanding events plays a major role in various natural language processing tasks such as information extraction (Humphreys et al., 1997), question answering (Narayanan and Harabagiu, 2004), textual entailment (Haghighi et al., 2005), event coreference (Choubey and Huang, 2018) and contradiction detection (De Marneffe et al., 2008). There has been a significant amount of work on automatic processing of events in text including systems for events extraction, event coreference resolution, and temporal relation detection (Araki, 2018; Ning et al., 2017). However, events are not atomic entities: they often have complex internal structure that can be expressed in a variety of ways (Huttunen et al., 2002; Bejan and Harabagiu, 2008; Hovy et al., 2013).

One of the unsolved problems related to event understanding is the detection of subevents, also referred to as *event hierarchy construction*. As described by Glavaš and Šnajder (2014a), there have been efforts that have focused on detecting temporal and spatial subevent containment individually. However, it is clear that subevent detection requires both simultaneously. The subevent relationship is defined in terms of (e_1,e_2), where e_1 and e_2 are events: event e_2 is a subevent of event e_1 if e_2 is spatiotemporally contained by e_1. More precisely, we say that an event e_1 is a parent event of event e_2, and e_2 is a child event of e_1 if (1) e_1 is collector event that contains a complex sequence of activities; (2) e_2 is one of these activities; and (3) e_2 is spatially and temporally contained within e_1 (i.e., e_2 occur at the same time and same place as e_1) (Hovy et al., 2013; Glavaš and Šnajder, 2014b). This subevent relationship is independent of other types of relationships, e.g., causal relationship between the events. Example 1 illustrates a text expression of a complex event hierarchy. Figure 1 shows a corresponding graphical representation of the hierarchy.

> *Egyptian police have said that five protesters were* ***killed***$_1$ *when they were* ***attacked***$_2$ *by an armed group near the Defense Ministry building in Cairo. The statement said that early this morning, the armed group* ***attacked***$_3$ *the demonstrators who have for days been staging their* ***protest***$_4$ *against the military government. . . . Police said that the* ***attack***$_5$ *on Wednesday* ***wounded***$_6$ *at least 50 protesters.*

Example 1: Excerpt from the HiEve corpus (Glavaš et al., 2014a). Events are in bold and given a numerical subscript for reference. In all the examples the identified events are gold annotations, but for clarity not all annotations are included.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4780–4790
Florence, Italy, July 28 - August 2, 2019.

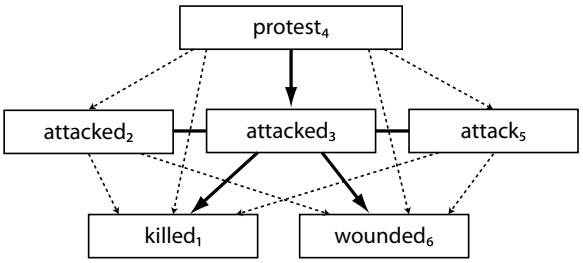

Figure 1: The corresponding event hierarchy of example 1. Bolded arrows indicate subevent relationships and bolded lines indicate event coreference relationships, when they are explicitly indicated in the HiEve annotations. Dashed lines indicate implicit subevent relationship.

In Figure 1, we see that **killed**$_1$ and **wounded**$_6$ are explicitly annotated as subevents of **attacked**$_3$, while that event in turn is a subevent of **protest**$_4$. Events **attacked**$_2$ and **attack**$_5$ are explicitly indicated as coreferent with **attacked**$_3$. These relationships induce the implicit subevent relations shown by dashed lines.

In this work we propose a pairwise model that leverages new discourse and narrative features to significantly improve subevent relation detection. evaluate our model on two corpora, namely, the HiEve corpus (Glavaš et al., 2014a) and the Intelligence Community (IC) corpus[1] (Hovy et al., 2013). We build on feature sets proposed in previous work, but propose several important discourse and narrative level features. We show that our model outperforms current systems on the subevent detection task by a significant margin. An error analysis reveals why these features are important and further details on why the subevent detection task is difficult.

We begin the paper by discussing prior work on subevent detection task (§2). Then we introduce our model and the feature set (§3). Following that, we describe the corpora (§4.1) we used and the experimental setup (§4.2). We then present the evaluation metrics and the performance of our model (§4.3) as well as compare our model performance to previous works (§5). To the end, we show an extensive error analysis (§6) and conclude with a list of contributions (§7).

2 Related Work

There are two pieces of prior work that are most related to our work. Araki et al. (2014) proposed a logistic regression model to classify pairs of events into four classes: *coreference*, *subevent*, *sister*, and *no relation*. They then used sister relations and their parents to improve the system performance. Their model was trained and tested on 65 articles from the IC corpus developed by (Hovy et al., 2013). Similarly, Glavaš and Šnajder (2014b) used a logistic regression model to classify pairs of event into three classes: subevent relations (*SuperSub* and *SubSuper*) and *no relation*. They enforced structural coherence which improved the quality of the extracted event hierarchies by 7.6% F_1 score. They trained and tested their approach on the HiEve corpus developed by (Glavaš et al., 2014a). Both approaches were evaluated using different evaluations metrics. Araki et al. evaluated their model using BLANC evaluation metric (Recasens and Hovy, 2011) whereas Glavaš and Šnajder evaluated their model using the standard F_1 evaluation metric. Both works introduced a variety of features. The main contribution of our work is to note that the subevent detection task requires a better understanding of the discourse. Thus here we introduce several new features, including discourse structure and narrative structure. The error analysis (§6) demonstrates why these features are effective and also reveals more details on why subevent detection is difficult.

3 Features

In this section, we explain the features used in our model. As discussed, both the HiEve and IC corpus (Hovy et al., 2013; Glavaš et al., 2014a) are annotated with both subevent and event coreference relationships. We compute features over all pairs of events (e_1, e_2) where e_1 precedes e_2 in the text. Each pair of events is either related by a forward pointing parent-child relationship (PC), a backward pointing parent-child relationship (CP), or no relation (NoRel). Our features can be divided into five sets as shown in Table 1. In the following sections we first illustrate the features we directly obtained from prior work (§3.1); next we explain the features that were inspired by prior work but that we modified significantly (§3.2); and finally we introduce our new discourse and narrative features (§3.3).

3.1 Prior Features

We obtained most of the lexical and syntactic features, and several of the semantic features, directly

[1]The IC corpus is unfortunately not publically available; we obtained a copy from Hovy et al. (2013).

Feature Set or Feature	Representation	Description
Lexical		
Event Expression	Bag-of-Events	The surface form of e_1 and e_2.
Same Lemma	Binary	Whether e_1 and e_2 have the same lemma.
Temporal Signals*	Bag-of-Signals	If both events are in the same sentence, the temporal signals appearing in the sentence between the events, based on the temporal signals list from (Derczynski and Gaizauskas, 2010).
Event String Similarity	Numeric	The string similarity between surface forms of the events using a Levenshtein distance measure.
Syntactic		
Major POS	One-hot	The Major POS of e_1 and e_2 (e.g., Noun, Verb, or Adjective) [2 features].
Same Major POS	Binary	Whether the Major POS of e_1 and e_2 are the same.
POS Tag	One-hot	The POS Tag of e_1 and e_2. [2 features]
Same POS Tag	Binary	Whether the POS Tag of the e_1 and e_2 are the same.
Syntactic Dependency*	One-hot	The ancestor event of the other event in the dependency tree.
Determiner	Binary	Whether each event has a determiner. [2 features]
Semantic		
Semantic Frame	Binary	Whether e_1 and e_2 have the same semantic frame using SEMAFOR (Das et al., 2010).
Event Type*	One-hot	The event type of e_1 and e_2 extracted from the mapping from frames to event types (Liu et al., 2016). [2 features]
Same Event Type	Binary	Whether event types of e_1 and e_2 are the same.
VerbOccan Score	Numeric	The VerbOcean score (Chklovski and Pantel, 2004) between e_1 and e_2 for each of VerbOcean's five relations. [5 features]
Semantic Similarity*	Numeric	The cosine similarity between e_1 and e_2 embeddings using FastText (Mikolov et al., 2018) pre-trained model (wiki-news-300d-1M).
Most Likely Parent Event*	One-hot	Which event is most likely to be a parent of the other event if both exist in the training data (see §3.2).
WordNet Similarity	Numeric	The WordNet Similarity scores between e_1 and e_2 using (Lin, 1998; Wu and Palmer, 1994) similarity measures.[2 features]
Arguments		
Co-refering Event Arguments*	One-hot	Whether specific arguments of e_1 and e_2 corefer (Lee et al., 2017). Verb arguments are computed with Allennlp's SRL (Gardner et al., 2018; He et al., 2017), Nouns and Adjectives with SEMAFOR.
# of Coreferring Args	Numeric	The number of coreferring arguments between e_1 and e_2.
Event in the Other's Args	One-hot	Whether one event is mentioned in one of the other event's arguments, if both events are in the same sentence.
Discourse & Narrative		
Sentence Distance	Numeric	The number of sentences between e_1 and e_2.
Event Distance	Numeric	The number of events between e_1 and e_2.
Same Sentence	Binary	Whether e_1 and e_2 are in the same sentence.
Reported Speech	Binary	Whether an event mention is mentioned in a direct speech (see §3.3.1).
Non Major Mention	Binary	Whether the sentences, in which the events are mentioned, share co-referential non major mentions (see §3.3.2).
RST-DTs Relation	One-hot	The discourse relation between elementary discourse units (EDUs), where e_1 or e_2 are mentioned in, in Rhetorical Structure Tree Discourse Trees (RST-DTs; see §3.3.1).

Table 1: Features used in the model. Novel features are underlined. Features modified from prior work are marked with an asterisk.

from prior work on subevent detection (Araki et al., 2014; Glavaš and Šnajder, 2014b). We used spaCy (Honnibal and Montani) to compute lexical and syntactic features.

3.2 Modified Features

Five of our features were inspired by those in prior work, but we modified them for our system.

Temporal Signals We observed that if a sentence mentions two events from different event hierarchies, then a temporal signal often exists between them (e.g., *after* and *since*). This is illustrated by the first sentence in Example 6. To capture this we used a temporal signals list (Derczynski and Gaizauskas, 2010) to find intervening temporal signal words between the events, and encoded this as a bag of temporal signals.

Syntactic Dependency Both prior systems encoded a feature which captured whether one event in a pair was an immediate child (i.e., *governed*) of

the other. We expand that to checking for ancestry more generally. This is encoded as one-hot vector.

Event Type We use the mapping from frames to 33 ACE 2005 event types introduced in (Liu et al., 2016) to determine the event type of each event. Prior work relied on the IBM SIRE system to compute event types (Florian et al., 2010). This is encode as a one-hot vector.

Semantic Similarity We used the FastText (Mikolov et al., 2018) pre-trained model (`wiki-news-300d-1M`) to measure the semantic similarity between pairs of events. Prior work used the SENNA system for this feature (Collobert et al., 2011). This is encoded as a numeric feature.

Most Likely Parent Event Similar to (Araki et al., 2014), we count the number of times in the training data that a particular event lemma and POS pair is observed as a parent of another event lemma/POS pair. For a pair (e_1, e_2), if the lemma and POS of e_1 is more often found as a parent of e_2, this is encoded as the vector (1,0,0); if the opposite is true, this is encoded as (0,1,0). If there were no observations, this is encoded as (0,0,1). Prior work did not take into account the part of speech, or the direction of the subevent relationship.

Co-referring Event Arguments When matching arguments, we allowed ARG0 to match ARG0 or ARG1 and vice versa, and we also examined LOC and TMP modifying arguments. This is encoded as six-place binary vector for ARG0/ARG1, LOC, and TMP.

3.3 New Features

The new features are divided into three types: two discourse features (§3.3.1), one narrative feature (§3.3.2) and two semantic features (§3.3.3).

3.3.1 Discourse Features

We for the first time investigate the importance of discourse features for detecting subevents. We introduced two new features: rhetorical structure and reported speech.

Rhetorical Structure Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a hierarchical model aims to identify the discourse structure of a text. The text is first segmented into Elementary Discourse Units (EDUs) which in turn are linked in binary or multi-way discourse relations (see Carlson and Marcu, 2001). Rhetorical analysis has been shown to be beneficial in many NLP tasks including sentiment analysis (Somasundaran, 2010; Lazaridou et al., 2013; Bhatia et al., 2015), text generation (Prasad et al., 2005), information extraction (Maslennikov and Chua, 2007), question answering (Verberne et al., 2007) and coreference resolution (Cristea et al., 1998; Joty et al., 2015). Therefore we hypothesized that discourse structure could be useful to the subevent detection task. We employ the CODRA discourse parser (COmplete probabilistic Discriminative framework for performing Rhetorical Analysis; Joty et al., 2015) to build a discourse tree of each text. We use (Neumann, 2015) for post-processing the CODRA output to build a graph representing the result. We then extract the rhetorical relation between event mentions using the rhetorical relation between the EDUs in which the event are found. The feature is encoded as a one-hot vector covering all 16 main relation classes.

Consider Example 2. When applied to this text, the discourse parser identifies the relation between ***raid***$_3$ and ***killed***$_4$ as an *Elaboration* relation. Furthermore, the parser also captures a *Topic-Change* relation between **offensive**$_6$ and each of **killed**$_1$, **wounded**$_2$, **raid**$_3$, **killed**$_4$, and **injured**$_5$.

Although the discourse parser is useful primarily for providing information about inter-sentential relationships between events, it can also give useful information about intra-sentential relationships. Consider Example 3. For this text the discourse parser finds the *Background* relation between **abduction**$_1$ and each of **killed**$_2$ and **rescued**$_3$.

Reported Speech We also observed that

> *One Palestinian was* ***killed***$_1$ *and at least four others were* ***wounded***$_2$ *in an Israeli air* ***raid***$_3$ *near the southern Gaza town of Rafah on Sunday, Palestinian security sources said. ... Palestinian security sources said that one Palestinian bystander was* ***killed***$_4$ *and at least four others were* ***injured***$_5$*. ... Israeli troops continued a massive ground and air* ***offensive***$_6$ *in the Gaza Strip on Sunday.*

Example 2: Excerpt from IC corpus (Hovy et al., 2013). Events relevant to explaining the discourse features are bolded. Mentions relevant to explaining the narrative feature are underlined. Note that, for clarity, not all events marked in the corpus are bolded here (e.g., Reporting events such as *said*).

Mahsud, a former prisoner at Guantanamo Bay, is being hunted for the ***abduction***$_1$ *of two Chinese engineers, which ended last Thursday when commandos* ***killed***$_2$ *five kidnappers and* ***rescued***$_3$ *one Chinese.*

Example 3: A sentence where intra-sentential discourse relations are useful for discovering subevent relations.

subevents are often reported in direct and indirect speech. Direct speech is speech set off with quotes, while indirect speech is speech reported without quotes. We only considered direct speech in this work, primarily because it is easy to detect; however, subevents are also likely to be reported in indirect speech as can be seen in example 2 where **killed**$_4$ and **injured**$_5$ (which are subevents of **raid**$_3$) are mentioned in indirect speech.

3.3.2 Narrative Feature: Non-Major Mentions

We also introduced what we are calling a *narrative* feature that we found informative in detecting subevent relations. This feature recognizes that other entities mentioned in a sentence besides those in the event arguments can be useful in subevent detection. This feature is narrative in the sense that it takes into account whether an entity is central to the story in the text.

In particular, we observed that many sentences which shared an event hierarchy also share some coreferring mentions beside events argument. Despite this, certain entities are so central to the text that they are mentioned nearly everywhere and are thus no especially informative. Therefore we filter out these *major mentions* and encode as a binary feature whether or not the sentences contain the events share a non-major mention.

The trick, of course, is defining what is a *major mention*. A simple and effective way of filtering out major mentions is to measure the distribution of coreference chain lengths (normalized to the number of the corresponding article's chains), and discard all chains with a length above a certain threshold. This threshold can be tuned to the data. In our experiment we estimated the mean and standard deviation of the distribution of coreference chains in each text and filtered out chains that were longer than a single standard deviation above the mean. In Example 2, the threshold of the corresponding article is 2, thus *Palestinian security sources*, which is mentioned only twice, is not considered a major mention.

The Al-Qaeda linked Army of Ansar al-Sunna claimed responsibility Tuesday for a car bomb ***attack***$_1$ *which* ***killed***$_2$ *four Iraqi guardsmen . . .*

Example 4: A sentence where one event appears inside the argument for another event. Event **killed**$_2$ is a subevent of **attack**$_1$.

3.3.3 Semantic Features

Event in the Other's Arguments We observed that if an event hierarchy is expressed within a sentence, one of the events is often mentioned as part of the other event's arguments as can be seen in Example 4, where the **attack**$_1$ event appears as ARG0 of **killed**$_2$. Although this feature is related to the *Syntactic Dependency* feature, an event's arguments are not always syntactically dependent on the event head, so it adds useful information.

Number of Coreferring Arguments We also include the number of coreferring event arguments as numeric feature.

4 Experiment

Here we describe the corpora on which the experiment were performed and the evaluation metrics used to measure the performance of our model. Then we compare the performance of our model with previous models, specifically those of Araki et al. (2014) and Glavaš and Šnajder (2014b).

4.1 Corpora

As already mentioned, we used two corpora: the Intelligence Community (IC) (Hovy et al., 2013) corpus and HiEve corpus (Glavaš et al., 2014a) to train and test our model. The IC corpus contains 100 news articles in the Violent Event domain (*attacks, killings, wars, etc*). The HiEve corpus is an open domain corpus that also contains 100 news articles. Both corpora are annotated with both coreference and subevent relations. The inter-annotator agreement for the IC corpus is 0.467 Fleiss's kappa for subevent relations. The approach proposed for temporal relations by (Uz-Zaman and Allen, 2011) was used to measure the inter-annotator agreement in HiEve corpus, resulting in 0.69 F_1. There is a small conceptual difference between the annotation of subevent relations in both corpora. The annotation of subevents in the IC corpus follows (Hovy et al., 2013) where they argued that there are three degrees of event iden-

tity: *fully identical*, *quasi-identical* (a.k.a., partial co-reference) and *fully independent* (not identical). Quasi-identity in turn appears in two ways: *membership* or *subevent*. Membership is defined as when an event is a set of multiple instances of the same type of event and the other event is one of the instances. In Example 5, **attack**$_1$ and **operation**$_2$ are members of **blows**$_3$, not subevents. In contrast, the HiEve corpus considers the membership relation as a subevent relation. When training on the IC corpus we considered only the subevent relations, and ignore the membership relations.

> *The Al-Qaeda linked group which said it carried out the deadly* ***attack***$_1$ *against US soldiers in the Iraqi city of Mosul accused the United States ... The* ***operation***$_2$ *is one of the heaviest* ***blows***$_3$ *in the city of Mosul ...*

Example 5: Illustration of the *membership* quasi-identity relationship of Hovy et al. (2013)

For both corpora we extend the annotations by computing the transitive closure of both coreference and subevent relations according to the following rules, where e_1, e_2 and e_3 are event mentions, $\equiv$ indicates event coreference, $e_1 > e_2$ indicates e_1 is a parent of e_2 , and $e_1 < e_2$ indicates e_1 is a child of e_2. All of these rules are taken from the work by Glavaš et al. (2014a). We confirmed that this closure produces a consistent graph, and thus is insensitive to the order of computation of the closure. Table 2 shows the statistics of both corpora.

1. $(e_1 \equiv e_2)$ & $(e_2 \equiv e_3) \Rightarrow (e_1 \equiv e_3)$
2. $(e_1 > e_2)$ & $(e_2 > e_3) \Rightarrow (e_1 > e_3)$
3. $(e_1 < e_2)$ & $(e_2 < e_3) \Rightarrow (e_1 < e_3)$
4. $(e_1 > e_2)$ & $(e_2 \equiv e_3) \Rightarrow (e_1 > e_3)$
5. $(e_1 > e_2)$ & $(e_1 \equiv e_3) \Rightarrow (e_3 > e_2)$
6. $(e_1 < e_2)$ & $(e_2 \equiv e_3) \Rightarrow (e_1 < e_3)$
7. $(e_1 < e_2)$ & $(e_1 \equiv e_3) \Rightarrow (e_3 < e_2)$

4.2 Experimental Setup

We use Linear SVM classifier from scikit-learn package for classification over the gold annotated event mentions. Linear SVM can handle multi-class classification using a one-vs-rest scheme (Pedregosa et al., 2011). Most of the parameters are default parameters [2], but to address the issue of the data imbalance as shown in Table 3, we use the parameter `class_weight=balanced` to assign a higher misclassification penalty on the minority class (PC and CP). We conducted 5-fold cross-validation for the experiment. Average fold statistics are shown in Table 3.

[2] `penalty=l2,C=0.01, random_state=0, max_iter=1000, class_weight=balanced, multi_class=ovr.`

	IC	HiEve
# of sentences	1,973	1,377
# of tokens	48,737	34,917
# PC relations, original	472	609
# PC relations, transitive closure	1632	1802
# CP relations, original	257	351
# CP relations, transitive closure	1665	1846
# NoRel relations	48567	42094
Avg # of sents. per article	19.7	13.7
Avg # of sents. in an event boundary	6.2	8.3
Avg # of events per article	30.5	26.0
Avg # of events in each hierarchy	5.2	7.0
Avg # of hierarchies per article	3.29	2.19

Table 2: Statistics of IC and HiEve corpora.

4.3 Evaluation and Result

We use the same evaluation metrics used in previous models. (Araki et al., 2014) evaluated their model using BLANC evaluation metric (Recasens and Hovy, 2011) whereas (Glavaš and Šnajder, 2014b) evaluated their model using the standard F_1 evaluation metric. The results of the performance averaged across all five folds on the three classes (PC, CP and NoRel) are shown in Table 4 using both evaluation metrics on both corpora. Table 5 shows the comparison between our model and previous models. Although it is not clear to us how Araki et al. handled the direction of the subevent relation, we take the average of our model classes (PC and CP) and compare it with the subevent class in Araki et al.'s work. For Glavaš and Šnajder, we consider only their *coherent* model, which is the best model that does not use the gold coreference relations. Therefore, in Table 5, the reported result of all models are the average of both classes (PC and CP). From Table 5, we can see that our model outperforms both prior models, by 15 and 5 percentage points. We also see that the precision is lower than the recall which indicate that the subevent detection task is still a difficult and complex task that needs more work. In the next two sections we explain why the performance of our model is low on IC corpus compared to the HiEve corpus, as well as an extensive error analysis.

	IC corpus			HiEve corpus		
	Training	**Test**	**Total**	**Training**	**Test**	**Total**
# articles	80	20	100	80	20	100
# PC (avg.)	1299.2	332.8	1632	1484	318	1802
# CP (avg.)	1317.8	347.2	1665	1456.4	389.6	1846
# NoRel (avg.)	39469	9098	48567	35621.2	6472.8	42094

Table 3: Average statistics of the folds. PC stands for parent-child relation. CP stands for child-parent relation. NoRel stands for no relation.

		Evaluation Metrics							
		F_1 Score			BLANC				
					Pos Links		Neg Links		Avg
Corpus	Relation	P	R	F_1	P	R	P	R	F_1
HiEve	PC	0.576	0.807	0.67	0.661	0.832	0.989	0.973	0.857
	CP	0.661	0.832	0.733	0.576	0.807	0.990	0.971	0.825
	NoRel	0.98	0.945	0.962	0.980	0.945	0.625	0.830	0.836
IC	PC	0.469	0.564	0.506	0.455	0.549	0.982	0.973	0.735
	CP	0.454	0.550	0.492	0.468	0.564	0.983	0.975	0.743
	NoRel	0.966	0.905	0.958	0.966	0.949	0.461	0.557	0.729

Table 4: Our model result on IC corpus (Hovy et al., 2013) and HiEve corpus (Glavaš et al., 2014a) using BLANC and F_1 standard evaluation metrics. PC stands for parent-child relation. CP stands for child-parent relation.

5 Discussion

As shown in Table 4, our model performs worse on the IC corpus than on HiEve. This is not surprising given the large difference in annotation agreement between IC and HiEve as well as the the removal of *membership* relations on IC corpus (see §4.1). Beside its lower annotation agreement, the IC corpus is also domain specific, with events only related to the intelligence community. This make general resources and tools (e.g., VerbOcean, WordNet) less effective.

We investigated the importance of each of the five feature sets (Table 1) to our model by retraining it while leaving out one set at time. In order of importance they are (1) Syntactic, (2) Semantic, (3) Discourse & Narrative, (4) Lexical, and (5) Arguments. The importance of the syntactic features derived from the fact that children events are most often mentioned in the same sentence as their parent events. The three most important features among the Semantic features are *Most Likely Parent Event*, *Event Type*, and *Semantic Frame*. For the Lexical feature set, the *Event Feature* and *Temporal Signals* are the most important.

6 Error Analysis

Inspection of the results revealed several types of errors, aside from the usual noise introduced by the various sub-components, such as the discourse parser or co-reference systems. We cluster the errors into three types: (1) an event pair that should be classified as *PC* but classified as *CP* and vice versa (about 28%); (2) an event pair is wrongly classified as NoRel (missed subevent relation; about 12%); (3) an event pair that is actually NoRel is wrongly classified as subevent (PC or CP; about 60% of the errors).

Type 1: PC as CP or vice versa About a third of the model errors were this type. Most of the errors are a result of an incorrect *Event Type* feature. This feature plays a major role in capturing the direction of the subevent relation. For example, if an event e_1 with event type *Die* occurs in the text before an event e_2 with event type *Attack*, then the direction of the relation is mostly *child-parent* relation. But if e_2 occurs before e_1, then the direction of the relation is mostly *parent-child*. If the event type is unknown for one of the event mentions, then our model commonly usually fails to capture the direction.

Type 2: Incorrect NoRel Most of the type 2 errors occur when an event is far away from its related event, in terms of number of intervening sentences. The larger the distance between events the more likely the model makes this error. For this type of error, we calculated the average number of sentences and the average number of events intervening between a missed pair of event, which the model should capture its subevent relation, and

Corpus	Model	F_1 Score			BLANC				
					Pos Links		Neg Links		Avg
		P	R	F_1	P	R	P	R	F_1
IC	Araki et al. (2014)	-	-	-	0.144	0.333	0.993	0.981	0.594
	Araki et al. Re-Impl.	0.242	0.285	0.262	-	-	-	-	-
	Our model	0.461	0.557	**0.499**	0.461	0.557	0.983	0.974	**0.739**
HiEve	Glavaš and Šnajder (2014b)	0.766	0.565	0.65	-	-	-	-	-
	Glavaš and Šnajder Re-Impl.	-	-	-	0.562	0.750	0.983	0.971	0.813
	Our model	0.618	0.82	**0.701**	0.618	0.82	0.99	0.972	**0.841**

Table 5: Our model performance compared to previous models (Araki et al., 2014; Glavaš and Šnajder, 2014b). Each row represent the average of both classes parent-child (PC) and child-parent (CP). Because the prior systems both did not report both metrics, we approximated the metrics for those systems by reimplementing them.

found that when the distance is greater than 9 sentences and the number of events is greater than 14 events, the more likely the model would conduct this error. Subevents tend to be close to their parents in the text as shown in Table 2. Moreover, we observed that the *Non-Major Mention* (§3.3.2) and *Discourse Relation* features (§3.3.1), were less useful the larger the distance between the events.

Type 3: False Positive PC or CP Most of the errors were of this type. There were a variety of causes, but the most common was when a sentence contained multiple event hierarchies. Consider Example 6 where the sentence contains two different event hierarchies, namely, one hierarchy containing **offensive**$_3$ and another containing **abduction**$_4$.

> *Over 90 Palestinians and one Israeli soldier have been* ***killed***$_1$ *since Israel* ***launched***$_2$ *a massive air and ground* ***offensive***$_3$ *into the Gaza Strip on June 28, three days after the* ***abduction***$_4$ *of one Israeli soldier by Palestinian militants in a cross-border* ***raid***$_5$.

Example 6: Excerpt from IC corpus (Hovy et al., 2013) showing a passage that results in an error of Type 3.

In this example, **killed**$_1$ and **launched**$_2$ are subevents of **offensive**$_3$, whereas **abduction**$_4$ is a subevent of **raid**$_5$. When processing this example the discourse parser failed to capture the discourse relation between **offensive**$_3$ and **abduction**$_4$ because both events are in the same EDU. Moreover, even though we introduced features such as temporal signals (*after*, *since*, etc.) to capture subevent relation between intra-sentential events, this error can still occur if the intra-sentential events are syntactically related (i.e., **killed**$_1$ syntactically dominates **abduction**$_4$, or there is a causal relation between events).

Based on this observation, we ran an experiment on the IC corpus to examine the impact on subevent detection of having two different events in the same sentence. We construct a subset of the IC corpus (58 articles) which excluded all articles that contain at least one sentence with two different event hierarchy, and re-ran our main experiment. Under these conditions, the model performance increased by 6 and 4.6 points F_1 on *PC* and *CP* classes, respectively (because of the smaller set, we used 3 folds instead of 5). Returning to the original corpus, we observed that two different event hierarchies are mostly found in compound and complex sentences, and one of the them is usually background event. This observation indicates that splitting compound or complex sentences into two simple sentences in advance might be useful in detecting subevents. Even though the discourse parser does this splitting automatically, this split is not currently propagated to the other features.

7 Contributions

We present a model to detect subevent relation in news articles which outperforms the two prior approaches by 15 and 5 percentage points, respectively. Our model involves several novel discourse and narrative features, as well as a small number of feature modifications. Our error analysis indicates that having two event hierarchies in the same sentence is a major problem, as well as having significant separation between a parent and child event.

Acknowledgments

Mr. Aldawsari was supported by a doctoral fellowship from Prince Sattam Bin Abdulaziz University, and thanks Dr. Sultan Aldossary for his advice and support. This work was also supported

by US National Science Foundation grant number IIS-1749917 to Dr. Finlayson. Both authors would like to thank Ed Hovy for providing the IC Corpus for our use.

References

Jun Araki. 2018. *Extraction of Event Structures from Text*. Ph.D. thesis, Carnegie Mellon University.

Jun Araki, Zhengzhong Liu, Eduard H Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4553–4558, Lisbon, Portugal.

Cosmin Adrian Bejan and Sanda M Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, pages 2881–2887, Marrakech, Morocco.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2212–2218, Lisbon, Portugal.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*.

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 (EMNLP)*, pages 33–40, Barcelona, Spain.

Prafulla Kumar Choubey and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume I*, pages 485–495, Melbourne, Australia.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Dan Cristea, Nancy Ide, and Laurent Romary. 1998. Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-ICCL)*, pages 281–285, Montreal, Canada.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A Smith. 2010. Probabilistic frame-semantic parsing. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 948–956, Los Angeles, CA.

Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08 HLT)*, pages 1039–1047, Columbus, OH.

Leon Derczynski and Robert Gaizauskas. 2010. USFD2: Annotating temporal expresions and tlinks for tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, pages 337–340, Los Angeles, CA.

Radu Florian, John F Pitrelli, Salim Roukos, and Imed Zitouni. 2010. Improving mention detection robustness to noisy input. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 335–345, Cambridge, MA.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

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

Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. 2014a. Hieve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th Language Resources and Evaluation Conference (LREC)*, pages 3678–3683.

Aria D Haghighi, Andrew Y Ng, and Christopher D Manning. 2005. Robust textual inference via graph matching. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 387–394, Vancouver, Canada.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Volume I*, pages 473–483, Vancouver, Canada.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://github.com/explosion/spaCy; Last accessed on May 31, 2019.

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

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 75–81, Madrid, Spain.

Silja Huttunen, Roman Yangarber, and Ralph Grishman. 2002. Complexity of event structure in ie scenarios. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7, Taipei, Taiwan.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Volume I*, pages 1630–1639, Sofia, Bulgaria.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197, Copenhagen, Denmark.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 296–304, San Francisco, CA.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging framenet to improve automatic event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Volume I*, pages 2134–2143, Berlin, Germany.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Mstislav Maslennikov and Tat-Seng Chua. 2007. A multi-resolution framework for information extraction from free text. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 592–599, Prague, Czech Republic.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, and Armand Puhrsch, Christian andJoulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 693–701, Geneva, Switzerland.

Arne Neumann. 2015. discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 309–312, Vilnius, Lithuania.

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 (EMNLP)*, pages 1027–1037, Copenhagen, Denmark.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32, Birmingham, UK.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*, pages 28–34. Stanford, CA.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Swapna Somasundaran. 2010. *Discourse-level relations for Opinion Analysis*. Ph.D. thesis, University of Pittsburgh.

Naushad UzZaman and James Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 351–356, Portland, OR.

Suzan Verberne, LWJ Boves, NHJ Oostdijk, and PAJM Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR*

Conference on Research and Development in Information Retrieval (SIGIR), pages 735–736, Amsterdam, The Netherlands.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 133–138, Las Cruces, NM.