The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction

Tommaso Caselli and **Piek Vossen**
Vrije Universiteit Amsterdam
De Boelelaan 1105 1081 HV Amsterdam (NL)
{t.caselli;p.t.j.m.vossen}@vu.nl

Abstract

This paper reports on the Event StoryLine Corpus (ESC) v0.9, a new benchmark dataset for the temporal and causal relation detection. By developing this dataset, we also introduce a new task, the StoryLine Extraction from news data, which aims at extracting and classifying events relevant for stories, from across news documents spread in time and clustered around a single seminal event or topic. In addition to describing the dataset, we also report on three baselines systems whose results show the complexity of the task and suggest directions for the development of more robust systems.

1 Introduction

Humans have an appetite for information to explain the things they observe. Our minds constantly mine the present for cues, merge this with information from the past, and derive models for reasoning and taking decisions. It is by means of such explanatory patterns, and by extension of *explanatory relations* among entities and events, that we understand the changing world.

The current stream of information poses a big challenge both to humans and systems to extract, organize, and represent events and their relations. News aggregation systems can easily monitor the burst and the development of a topic, or news story, but they fail in providing a content-based analysis. Given a topic or trending story, people still have to read the documents and reconstruct a unitary and coherent report mentally. Current NLP systems can identify complex information but they lack a method to connect it in a unitary and coherent message. Steps in this direction have been conducted but are very limited and do not cover the full story that is told by these documents (e.g. the textual entailment task, or script extraction).

Monitoring a news story from its beginning to end is a challenging task, which requires systems to be able to: 1) reconcile information from different sources distributed in time; 2) resolve deduplication of information; and 3) extract informative semantic structures.

It is surprising to observe how humans can perform these tasks with relative little effort. It has been suggested that this capacity is partly based on narrative strategies (Boyd, 2009; Gottschall, 2012). Such a structuring is possible thanks to a key component of narratives, the *plot structure* (Bal, 1997), which provides a chronological and logical ordering of events. This means that events are not simply ordered in time but they are selected and connected in such a way that their relations are meaningful, i.e., they give rise to a network of explanatory relations. Accessing and reconstructing plot structures for different topics would be beneficial for lots of Natural Language Understanding applications (question answering, summarization, co-reference resolution, event processing, and script extraction, among others).

One of the necessary step for a StoryLine Extraction task is to decide on a corpus to evaluate performance of systems. This paper presents such as resource: the Event StoryLine Corpus v0.9, specifically designed for the evaluation of systems aiming at reconstructing event-centric plot structures. The resource is still being extended with new annotated texts, but in the remainder of the paper we will refer to this first version. The corpus has been developed by applying annotation guidelines designed to mark-up the network of explanatory relations which can be realized between pairs of events in a document belonging to a specific topic. Furthermore, the guidelines

Proceedings of the Events and Stories in the News Workshop, pages 77–86,
Vancouver, Canada, August 4, 2017. ©2017 Association for Computational Linguistics

are compliant with other initiatives for event annotation: temporal processing (TimeML (Pustejovsky et al., 2003a) and Richer Event Description (RED) (O'Gorman et al., 2016)), event coreference (Event Coreference Bank+ (ECB+) (Cybulska and Vossen, 2014b)), and causal relations (Causal-TimeBank (Mirza and Tonelli, 2016), BECauSE (Dunietz et al., 2015), ROCStories (Mostafazadeh et al., 2016b) among others).

The remainder of the paper is structured as follows: Section 2 will explain the annotation scheme, describe the annotation layers of the Event StoryLines Corpus (ESC) v0.9, and report on agreement measures. Section 3 will describe experiments related to the development of baselines for the StoryLine Extraction task. In Section 4 a review of previous annotation initiatives is given, showing differences and commonalities between them and the ESC data. Finally, conclusions and future work are reported in Section 5. The annotated data, the evaluation scripts, and the baselines models are publicly available. [1]

2 The Event StoryLine Corpus v0.9

The primary goal of the ESC v0.9 dataset is to provide an intrinsic evaluation benchmark for the event-centric StoryLine Extraction task. The task can be best described as a combination of three basic subtasks:

- **Event Detection and Classification** Identify and classify events in each document which compose a topic, or a *seminal event*;
- **Temporal Anchoring of Events** Anchor each event mention to the temporal expression expressing the time of its happening, as well as to the Document Creation Time (DCT);
- **Explanatory Relation Identification and Classification** Select event pairs which are temporally and logically connected, and then, classify the storyline relation type.

A storyline relation can be best described as a loose causal and temporal relation between a pair of event mentions, where one event mention explains/justifies the occurrence of the other event mention in the pair (more details are reported in Section 2.3). Relations can be classified either as `rising_action`, or `falling_action`.

An additional task is **Event Co-reference Resolution**, which aims at identifying co-referential chains of events mentions both at within- and cross-document levels. The availability of this information allows us to deduplicate information across event mentions by creating *event instances*, i.e. formal semantic representation in RDF compliant URIs that may integrate linguistic information with external resources, and thus, allow reasoning (Fokkens et al., 2013).[2] In the following sections, we will illustrate the components of the ESC Annotation Scheme and its annotation framework.

2.1 Basic Components: Events and Temporal Expressions

Events and temporal expressions are the basic components of the annotation scheme for the ESC v0.9 dataset.

The term "event" is used as a cover term to refer to any situations that can happen, occur, or hold. The use of the term event is a synonym to "eventuality" introduced by Bach (1986), covering both dynamic and static situations (i.e. events and states). The annotation of events in NLP is a topic that got a lot of interest and on which yet no consensus has been reached. In this work, we adopted a definition of events that is provided in the ECB+ Annotation Guidelines (Cybulska and Vossen, 2014a), which is compatible with definitions in ACE (Linguistic Data Consortium, 2005) and TimeML. In particular, an event is any punctual, durational, or stative situation which happens or holds, and which results from a combination of four components such as: 1) an **action** component referring to what happens or holds; 2) a **time** slot which is responsible for anchoring the action in time ; 3) a **location** component which links the action component to a place/location; and 4) a **participant** component, which illustrates the "who" or "what" is involved in the action component.

The annotation of the extent of events in ECB+ follows the solution adopted in TimeML. This means that for each event mention, regardless of its part-of-speech, only the lexical item which is the bearer of the action meaning is annotated. This normally corresponds to the head of the phrase

[1] `https://github.com/cltl/EventStoryLine.git`

[2] `http://groundedannotationframework.org/files/2013/05/GAF_Poster.pdf`

realizing the action component, i.e. *the minimal chunk*, as illustrated in the following example[3]. Annotated events are in bold.

1. This terrible **war** could have **ended** in a month

However, exceptions to this rule apply. Adopting an event-centric annotation framework, adherence to the text surface is not always maintained. For instance, cases of historically significant events which may be referred to with proper nouns, such as *World War II*, *the American Civil War*, are annotated with a unique action component tag. Similarly, as the annotation is also primarily focused towards event co-reference, pre-modifiers of events can be included in the action component tag any time they contribute to the identification of a unique event instance:

2. **6.1-magnitude quake** strikes Indonesia's Aceh.

Furthermore, ECB+ allows the annotation of present- and past-participles in modifier position as event mentions:

3. The **earthquake** [. . .] left hundred trapped in **collapsed** buildings.

Each action component is classified as belonging to one of seven possible classes. Five of them, *ACTION_OCCURRENCE*, *ACTION_ASPECTUAL*, *ACTION_REPORTING*, ACTION_STATE, and *ACTION_PERCEPTION*, mirror TimeML classes. The two additional classes , *ACTION_CAUSATIVE* and *ACTION_GENERIC*, have been introduced to annotate events expressing casual relations, and events which are not anchored to a specific time and location expressing generic actions (i.e. event mentions whose truthfulness is independent of the specific moment of utterance).

Temporal expression mark-up is inherited from TimeML following the *TIMEX3* annotation guidelines. We modified the original ECB+ annotation guidelines to be compatible with the *TIMEX3* TimeML ones by: 1) using the *TIMEX3* tag to annotate temporal expressions, 2) re-introducing the `type` attribute as part of the temporal expression tag; 3) re-introducing the attribute `value` for temporal expressions' normalization. We also allow the creation of empty *TIMEX3* tag, i.e. non-text consuming temporal expression markables corresponding to implicit, i.e. not realized in the text, beginning and/or end points of temporal expressions denoting a duration. In addition to this, temporal expressions which have been included in action tags as part of the action component description must be annotated also as independent temporal expressions. This means that we allow multiple annotations on overlapping tokens over different text expressions. We made this choice because these temporal expressions in most cases also function as temporal anchor of the event component.

2.2 Temporal Anchoring of Events (TLINKs)

Temporal information plays an essential role for StoryLine Extraction. At the same time, the annotation of temporal relations is by no means a trivial task.

Two types of temporal relations can be identified: 1) ordering relations, which involve elements of the same ontological type, e.g. pairs of events or temporal expressions; and 2) anchoring relations, which involve cross-type element relations, e.g. pairs of event and related temporal expression. Although both types of temporal relations are useful, they have different informational status. Following Pustejovsky and Stubbs (2011), we assume that the informational level of a temporal relation can be expressed as a function of the information contained in each temporal link and their closure. Under this assumption, anchoring relations expressing when an event mention occurred or its duration, are more informative than ordering relations. The former allow us to put event mentions on a specific point (or interval) on an imaginary timeline and, as a consequence, also gives us the ordering relations between event mentions.

The ESC Annotation Scheme expresses temporal relations using the TimeML *TLINK* tag and restricts them to anchoring relations. *TLINKs* between an event mention and a temporal expression are systematically annotated when an anchoring relation is instantiated. Anchoring relations may hold between an event mention and a temporal expression at intra- and inter-sentential levels In addition to this, each event mention is also connected to the Document Creation Time (DCT) of each document.

Limiting the annotation to anchoring relations

[3] All examples are taken from the ECB+ Annotation Guidelines or the ECB+ annotated data

is also a strategy to avoid the complexity of ordering relations between events. Most of the current solutions are not optimal, as they give the annotators too much freedom in the the selection of the event pairs (e.g. TimeML), or force the annotators to mark all possible relations (e.g. TimeBank-Dense (Cassidy et al., 2014)), or limit the annotations to the presence of explicit linguistic evidence (e.g. RED).

The temporal values in ESC are derived from the RED guidelines. We apply two sets of *TLINK* values according to the type of anchoring relation annotated: four values apply for relations between events and DCTs (namely `before`, `after`, `overlap`, and `contains`), while only one value (`contains`) applies to relations between events and temporal expressions. Annotators are also instructed on the directionality of the *TLINK*, which should always go from the temporal expression, or DCT, to the target event.

2.3 Explanatory Relation Annotation (PLOT_LINKs)

The annotation of explanatory relations between event pairs is encoded in the *PLOT_LINK* tag, following a previous proposal described in Caselli and Vossen (2016). *PLOT_LINK*s are specifically designed to capture the semantics of plot structures.

PLOT_LINK annotation is conducted in two steps: first, annotators have to identify all eligible relations between event pairs, and then they have to classify each relation as belonging to one of the two classes: `rising_action`, events which are circumstantial to, cause or enable another event, or `falling_action`, which explicitly mark speculations and consequences, i.e. events which are the (anticipated) outcome or the effect of another event.

*PLOT_LINK*s are related to causal and temporal relation annotation (Miltsakaki et al., 2004; Bethard et al., 2008; Mirza and Tonelli, 2014; Dunietz et al., 2015), but they differ in three ways: 1) they include the standard causal relations, i.e. *cause*, *enablement*, and *prevention*, but also additional event-event relations such as contingency, sub-event, entailment, and co-participation relations; 2) they are often not explicitly marked in the text through a relational structure; and 3) they are more specific than all events that stand in a temporal relation as they add explanatory information.

*PLOT_LINK*s can be positioned in between temporal and causal annotations by overcoming current shortcomings, such as creation of uninformative pairs of events, in the former case, and an extremely limited annotation in the latter, i.e. presence of an explicit causality trigger. Each pair of events in a *PLOT_LINK* relation is basically helping the reader (and the machine) to connect events in a meaningful way. In a nutshell, *PLOT_LINK*s aim at answering "why" something has happened. Given their event-centric nature, the answer to such a question must be another event mention explicitly stated in the document in analysis.

PLOT_LINK relations are asymmetrical and non-transitive. Non-transitivity is justified by considering the nature of this type of relations. They apply at a local level of analysis between pairs of events, and cannot be transferred to a global level, i.e. inherited by the full chain of event mentions which contribute to the identification of a storyline. Although subjected to the chronological order of events, this type of relations aims at making explicit the coherence, or logical connections, of the events in a (news) story.

When annotating *PLOT_LINK*s, the (broad) "causal" dimension of the relation is more prominent than the temporal aspect. We are not filling-up a timeline, where the axiom of the Internal Directionality of Time[4] (Bonomi and Zucchi, 2001) holds, but we are looking for explanations of "why" events happened, according to the information that we are given in the document of analysis. Thus, in example 4, the relation between the events "earthquake" and "trapped" is obtained by answering the question "why were people trapped?" and not by means of transitive relation between the pairs *earthquake* `rising_action` *collapsed* and *collapsed* `rising_action` *trapped*.

4. The **earthquake killed** 14 and **left** hundred **trapped** in **collapsed** buildings.
   earthquake `rising_action` killed
   earthquake `rising_action` trapped
   earthquake `rising_action` collapsed
   collapsed `rising_action` trapped

Annotators are free to identify the pairs of events which may stand in a *PLOT_LINK* rela-

[4] Internal Directionality of Time: if it is true of my current position in time, t, that the event **e** occurred in the past of t, then it is true of any future position t' that **e** is in the past of t

tion. We did not create a predefined set of pairs of events which may stand in a plot link, as in the TimeBank-Dense corpus, as this will require to create a really large graph between all events occurring both in the same sentence and across all sentences. However, we limited the annotation of *PLOT_LINK*s to events which correspond to one of the following three classes: *ACTION_OCCURRENCE*, *ACTION_PERCEPTION*, *ACTION_STATE*. We label those events as "semantically full" or "semantically loaded" events. Event mentions in these classes do have a content component describing a situation, rather than expressing meta-level information on the events.[5] The class of *ACTION_REPORTING* is excluded as well. In this case, the meaningful information is represented by the "content" of a speech event rather than by the lexical expression that introduces it. This choice guarantees that only meaningful events are part of a storyline.

Finally, *PLOT_LINK*s also allow the annotation of explicit causal relations between pairs of events. Two binary attributes, `cause` and `caused_by`, must be selected in presence of explicit causal relations. Explicit causal relations are introduced either by *ACTION_CAUSATIVE* events, or causal signals such as conjunctions (e.g. *because*), prepositions (e.g. *by*, *from*, *for*, among others), and other connectives. An additional attribute, `signal`, has been created to annotate the "markers" of the causal relation. At this stage of development, the attribute is filled only when *ACTION_CAUSATIVE* events are used to signal the presence of a casual relation:

5. A massive **quake struck** off Aceh in 2004 , **sparking** a **tsunami**.
   quake `rising_action` tsunami
   `signal`= sparking
   `cause` = YES

2.4 Event Co-reference

Currently, the annotation of co-referential chains among event mentions has been inherited from ECB+ The ECB+ guidelines consider two event mentions, either in the same document or across documents, as co-referential when they refer to the same event instance, i.e. if they describe the same action component, and 1.) share the same participants; 2) share the same temporal anchor; and 3) share the same location.

2.5 Data

The ESC v0.9 dataset is currently composed by 22 topics from the ECB+ corpus concerning calamity events, i.e. natural disasters, shootings, killings, accidents, and trials, among others.

The corpus contains 258 documents, and a total of 7,275 event mentions (191 of which being negated mentions).[6] A total of 1,297 temporal expressions are present, 248 of them corresponds to DCTs, of which 22 are realized by empty *TIMEX3* tags. In the remainder of the cases, 10 articles, it was not possible to recover a DCT, neither from the articles, nor by searching the Web.

Following the extended anchoring relation approach for *TLINK*s, we annotated a total of 6,904 relations between events and DCT and events and temporal expressions. The breakdown of the distribution of the values is reported in Table 1.

TLINK Value	DCT	TIMEX3
CONTAINS	522	2816
BEFORE	52	n.a.
AFTER	3283	n.a.
OVERLAP	160	n.a.

Table 1: *TLINK* value per DCT and temporal expression in the document.

As for the *PLOT_LINK*s, a total of 2,265 explanatory relations have been annotated, with an average of 8.7 relations per document. 1,147 relations have been classified as `rising_action`, while 1,118 as `falling_action`. By extending the manually annotated relations with within-document event co-reference chains, we reach a total of 5,519 *PLOT_LINK*s, almost three times the average relation per document, i.e. 21.39. This results in 2,653 `rising_action` and 2,844 `falling_action` relations, respectively. Finally, only 117 explicit causal relations have been identified.[7]

[5] We consider event mentions contributing to the assessment of the factuality profiling of an event mention, including cognitive events, events belonging to the class *ACTION_ASPECTUAL*, which functions as lexical morphosyntactic markers of the the internal temporal structure of a situation, and *ACTION_CAUSATIVE* as meta-level event mentions, and thus they are excluded.

[6] Event annotation is directly inherited from ECB+, where only sentences containing relevant mentions of the topic were annotated.

[7] Note that this can be extended further using the cross-document event coreference chains of ECB+

The annotation of the ESC v0.9 corpus has been conducted by 2 experts following a multi-step process and using the web-based tool CAT (Bartalesi Lenzi et al., 2012). In the first phase, both annotators went through a training phase to familiarize with the task, and were allowed to discuss and compare their annotations, especially for the *PLOT_LINK* task. This phase led to a revision of the annotation guidelines, by introducing more specific rules to select event pairs. In the second phase, the inter-annotator agreement was calculated on a subset of the ESC v0.9 dataset. In particular, given that the basic components, i.e. event mentions, temporal expressions, and event co-referential chains, are directly inherited from the ECB+ corpus, the agreement was calculated only for anchoring (i.e. *TLINK* tags) and explanatory relations (i.e. *PLOT_LINK* tags). Inter-annotator agreement has been computed using the Dice coefficient, both for relation detection and relation classification. Two different subsets of the ESC v0.9 corpus have been used for the two relations: one seminal event[8] for *TLINK*s and 4 seminal events[9] for *PLOT_LINK*s. We made this choice because of the different nature of the two types of relations. Results are reported in Table 2. The scores for *PLOT_LINK*s have been computed as an average over the 4 seminal events.

Relation Type	**Identification**	**Classification**
TLINK	0.767	0.744
PLOT_LINK	0.638	0.638

Table 2: Inter-annotator agreement: Dice coefficient at token level.

One of the most interesting observations on the *PLOT_LINK* analysis is that the agreement may vary according to the type of seminal event. For instance, the highest agreement has been observed for T19: a shooting accident :Dice 0.723 for relation identification, and 0.728 for relation classification. The lowest agreement was found for an escape from prison (T3): Dice 0.48 for relation identification, and 0.471 for relation classification. The results, although preliminary, suggest that different types of seminal events may be narrated in different ways following different story patterns (e.g. more or less linear stories).

3 Experiments: Baselines

In this section, we describe the experimental results for a number of StoryLine Extraction baseline systems on the ESC v0.9 dataset. The outcomes of these experiments will be useful to compare the performance of future (and more complex) systems, as well as to have a preliminary assessment of the complexity of the task.

The ESC v0.9 dataset has been divided into a development set, consisting of 6 seminal events[10] and a test set of 16 seminal events[11]. The test subset contains a total of 4,027 *PLOT_LINK*s when extended with within-document event co-reference chains. All experiments have been conducted considering gold data for event mention extent, temporal expression extent and values, and event co-reference.

Three baselines have been developed: 1) OP: selection of event pairs in relations that mimic the textual order of presentation; 2) PPMI1: selection of event pairs using Positive Pointwise Mutual Information (PPMI) obtained from a set of selected seed pairs and the manually annotated pairs from the development set; 3) PPMI-CONTAINS: selection of the event pairs using PPMI as in the PPMI1 model but restricting the sets of events to those which share the same temporal anchors, i.e. have a *TLINK* of type `contains`.

The seed pairs for the PPMI based models have been extracted from the SemEval 2012 Task-2: Measuring Degrees of Relational Similarity (Jurgens et al., 2012). In particular, we extracted words pairs from the test set Phase-1 Answers corresponding to class-8 (CAUSE-PURPOSE), retaining only word pairs in the categories Cause:Effect, Cause:Compensatory Action, Action/Activity, and Prevention, where both words express events. This initial set of seed elements has been further extended by looking for “cause”, “enablement”, and “entails” relations in SUMO (Niles and Pease, 2001, 2003) and in WordNet (Miller, 1995). This resulted in a list of 1,609 unique seed pairs. PPMI has been computed using the DISSECT Toolkit (Dinu et al., 2013), and pair frequencies have been extracted from Google bigrams(Brants and Franz, 2006). Rather than identifying a unique threshold for eligible pairs, we looked for a range of PPMI values.

[8] T37

[9] T3, T19, T37, T41

[10] T5, T7, T8, T32, T33, T35

[11] T1, T12, T13, T14, T16, T18, T19, T20, T22, T23, T24, T3, T30, T37, T4, T41

	PLOT_LINK Detection			PLOT_LINK Classification		
Baseline Model	P	R	F1	P	R	F1
OP	0.156	**0.988**	**0.265**	0.07	**0.97**	**0.14**
PPMI1	0.137	0.174	0.137	0.065	0.098	0.068
PPMI-CONTAINS	**0.227**	0.091	0.121	**0.114**	0.05	0.064

Table 3: Results of three baselines models on *PLOT_LINK* identification and classification .

This has been identified by normalizing the PPMI scores between 0 and 1, computing average and standard deviation. This allowed us to identify a minimum and a maximum normalized score[12] for PPMI, representing the boundaries of the range inside which event pairs in a *PLOT_LINK* relation can be identified and selected.

As for the extraction of the events in a *PLOT_LINK* relation from the test data, co-occurrence frequencies were computed per pairs of eligible event types (i.e. *ACTION_OCCURRENCE*, *ACTION_PERCEPTION*, *ACTION_STATE*) both at sentence and at document level. PPMI values were obtained by applying the same procedure used for the seed pairs. In the PPMI1 model, all event pairs whose score is within the range obtained from the seed pairs were selected. On the other hand, in the PPMI-CONTAINS model, the event pairs were further filtered by applying the temporal anchor constraints, i.e. they must both have a *TLINK* of type `contains` with the same temporal expression.

As for relation classification, i.e. the assignment of the values `rising_action` or `falling_action` to an event pair, we decided to always assign the `rising_action` value, i.e. the most frequent value from the manually annotated data. In addition to this, we also aimed at evaluating the impact of the order of presentation of the information in a document on *PLOT_LINK*s.

In Table 3, we report on the aggregated results, i.e. average score over the test data, of the three baselines. The relation detection subtask limits the evaluation to the correctness/validity of the event pairs identified by each model against the extended gold data. On the other hand, in the classification subtask, both the event pair and the relation value must be correct. This means that if the *PLOT_LINK* value is wrong but the event pair is correct, then the entire *PLOT_LINK* is considered incorrect. Standard Precision (P), Recall (R), and F1-score (F1) apply for both subtasks.

The results, though preliminary, highlight the complexity of the task. Not surprisingly the best Recall value is obtained by the OP model. The creation of all possible pairs between eligible event types clearly gives rise to a lot of False Positive pairs (P=0.156), showing that even when only events in relevant sentences of specific topic are selected, there is still information which is not to be included in a storyline. For instance, there could be references to events which occurred in the past and which do not have any explanatory relations with the event mentions referring to the current topic, and presented to the reader for comparison or as additional background knowledge.

Different observations apply to the PPMI-based models. In PPMI1, we can observe a big drop in Recall (-0.841) and as well as in Precision, though lower (-0.019). On the other hand, temporal containment seems to facilitate the aggregation of the relevant pairs of a storyline, as shown by Precision (P=0.227). At this stage of the implementation, there is a lack of connection between events in different temporal anchors, thus limiting the connections between event pairs and having a negative impact on the Recall.

By observing the results on the classification task, it immediately appears that the textual order of presentation of the information badly correlates with *PLOT_LINK* values. The low results were in part expected given the distribution of the `rising_action` and `falling_action` relations in the test data. To better understand the results, we run an additional evaluation on the baselines by taking into account only same sentence pairs. In this case, we observed that all baselines increase the Precision (P=0.123 for OP, P=0.095 for PPM1, and P=0.151 for PPMI-CONTAINS) and downgrade the Recall scores. Given the evaluation framework for classification, this suggests that, at least when in the same sentence, there is a tendency to narrate the events following a logi-

[12] Average PPMI value=0.582; standard deviation=0.181; minimum PPMI value=0.4; maximum PPMI value=0.763

cal order, not only a temporal one. However, this does not hold anymore when cross-sentence relations are taken into account.

4 Related Work

Frameworks and models for understanding narratives have mainly focused on fictional texts (Lehnert, 1981; Goyal et al., 2010; Mani, 2012) Modern day news reports still reflect narrative structures but they have proven difficult for automatic tools (Rospocher et al., 2016). To the best of our knowledge, previous work on StoryLine Extraction is limited, if we exclude the contribution by Caselli and Vossen (2016). However, there are several related works in NLP dealing with related tasks. The extraction of *causal relations* is the nearest task. One of the most prominent work is represented by the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004), where explicit and implicit causal relations are annotated between discourse units.

The Causal-TimeBank (Mirza and Tonelli, 2016) has introduced a TimeML-based annotation of causal relations between events on top of the TempEval-3 TimeBank data. Casual relations are annotated by means of a *CLINK* tag and only explicit causal relations are marked-up, i.e. the relation must be signaled by a linguistic markers (e.g. a preposition or a causal verb). This results in 318 *CLINK*s, 296 of which are in same-sentence. The RED guidelines (O'Gorman et al., 2016) combines event co-reference, temporal and causal relations. In particular, causal relations are expressed by means of `precondition` and `cause` values, allowing both same sentence and adjacent sentence relations, thus aiming at achieving a richer semantic representations of event relations. The BECauSe Corpus 2.0 (Dunietz et al., 2015) focuses on causal language, by representing what causal relationships are expressed in a text/document, rather than taking into account real world causality. Causal relations are annotated only in presence of a causal connective (i.e. a lexical item signaling the causal relation). The annotation scheme is very rich as it allows the mark-up of overlapping relations (e.g. temporal, correlation, hypothetical, among others) as well.

Another relevant work is the CaTeRs annotation scheme (Mostafazadeh et al., 2016b). In CaTeRs, causal relations between events are annotated from a "commonsense reasoning" perspective rather than starting from linguistic markers, inspired by the mental model theory of causality. The scheme identifies 9 classes of causal relations as well as 4 classes of temporal relations. The scheme has been applied over 320 stories from the ROCStories Corpus (Mostafazadeh et al., 2016a), which collects everyday stories (e.g. "got a phone call") composed by 5 sentences. The main goal of the annotation is to focus on those causal and temporal relations which may facilitate the learning of stereotypical narrative structures.

In this work, we have extended the set of event-event relations to be annotated using the notion of explanatory relation. In our work both implicit and explicit relations are annotated, allowing the annotation at both intra- and inter-sentential levels. In addition to this, the availability of within- and cross-document event co-reference chains allows the extension of the annotated data across documents, providing access to a larger, "global" level of analysis.

5 Conclusion and Future Works

This paper presents the Event StoryLine Corpus v0.9, the first benchmark corpus for a StoryLine Extraction task, i.e. temporally and logically connected sequences of events related to a specific topic from documents spread in time. We also presented three baseline systems with their performance on the data base. This task aims at moving away from current approaches on timeline and causal relation extraction. With respect to the former task, storylines aim at the chronologically ordering only of events that are relevant to a story, thus cleaning timeline structures. At the same time, storylines extend causal relation extraction by covering both explicit and implicit causal relations between events, both at a intra- and inter-sentential levels. This facilitates the learning of narrative models, i.e. explanatory patterns in news data, which can be used to identify both stereotypical and episodic narrations of seminal events, or topics, in news. One the innovative aspects is the connection with co-reference relations of events across documents, thus making the annotated data also useful for the development of cross-document summarization systems.

The corpus will be extended in the future by means of crowd-sourcing and by introducing annotations of climax events, i.e. the main events in the story. In parallel, we aim at developing more robust systems.

Acknowledgments

This work has been supported the NWO Spinoza Prize project "Understanding Language by Machines" (sub-track 3).

References

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy* 9(1):5–16.

Mieke Bal. 1997. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *In Proceedings of LREC 2012*. pages 333–338.

Steven Bethard, William J Corvey, Sara Klingenstein, and James H Martin. 2008. Building a corpus of temporal-causal structure. In *LREC*.

Andrea Bonomi and Alessandro Zucchi. 2001. *Tempo e linguaggio: introduzione alla semantica del tempo e dell'aspetto verbale*. Pearson Italia Spa.

Brian Boyd. 2009. *On the origin of stories*. Harvard University Press.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram corpus version 1.1. *Google Inc* .

Tommaso Caselli and Piek Vossen. 2016. The storyline annotation and representation scheme (star): A proposal. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*. Association for Computational Linguistics, Austin, Texas, pages 67–72. http://aclweb.org/anthology/W16-5708.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 501–506. http://www.aclweb.org/anthology/P14-2082.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ECB+ annotation of events and their coreference. Technical Report NWR-2014-1, VU University Amsterdam.

Agata Cybulska and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*. Reykjavik, Iceland.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. 2013. Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Sofia, Bulgaria, pages 31–36. http://www.aclweb.org/anthology/P13-4006.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*. Association for Computational Linguistics, Denver, Colorado, USA, pages 188–196. http://www.aclweb.org/anthology/W15-1622.

Antske Fokkens, Marieke van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. Gaf: A grounded annotation framework for events. In *Workshop on Events: Definition, Detection, Coreference, and Representation*. Association for Computational Linguistics, Atlanta, Georgia, pages 11–20. http://www.aclweb.org/anthology/W13-1202.

Jonathan Gottschall. 2012. *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Amit Goyal, Ellen Riloff, and Hal Daume III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, pages 77–86. http://www.aclweb.org/anthology/D10-1008.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics, Montréal, Canada, pages 356–364. http://www.aclweb.org/anthology/S12-1047.

Wendy G Lehnert. 1981. Plot units and narrative summarization. *Cognitive Science* 5(4):293–331.

Linguistic Data Consortium. 2005. ACE (Automatic Content Extraction) English annotation guidelines for entities.

Inderjeet Mani. 2012. Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies* 5(3):1–142.

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

Eleni Miltsakaki, Rashmi Prasad, Aravind K Joshi, and Bonnie L Webber. 2004. The penn discourse treebank. In *LREC*.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 2097–2106. http://www.aclweb.org/anthology/C14-1198.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 64–75. http://aclweb.org/anthology/C16-1007.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696* .

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*. Association for Computational Linguistics, San Diego, California, pages 51–61. http://www.aclweb.org/anthology/W16-1007.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*. ACM, pages 2–9.

Ian Niles and Adam Pease. 2003. Mapping wordnet to the sumo ontology. In *Proceedings of the ieee international knowledge engineering conference*. pages 23–26.

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)*. Association for Computational Linguistics, Austin, Texas, pages 47–56. http://aclweb.org/anthology/W16-5706.

James Pustejovsky, José Castao, Robert Ingria, Roser Saurì, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, pages 152–160. http://www.aclweb.org/anthology/W11-0419.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web* 37:132–151.