CATENA: CAusal and TEmporal relation extraction from NAtural language texts

Paramita Mirza
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
paramita@mpi-inf.mpg.de

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Abstract

We present CATENA, a sieve-based system to perform temporal and causal relation extraction and classification from English texts, exploiting the interaction between the temporal and the causal model. We evaluate the performance of each sieve, showing that the rule-based, the machine-learned and the reasoning components all contribute to achieving state-of-the-art performance on TempEval-3 and TimeBank-Dense data. Although causal relations are much sparser than temporal ones, the architecture and the selected features are mostly suitable to serve both tasks. The effects of the interaction between the temporal and the causal components, although limited, yield promising results and confirm the tight connection between the temporal and the causal dimension of texts.

1 Introduction

When the Greek government missed its 1.6 billion euro payment to the IMF as its bailout expired on 30 June 2015, people started to look for information, such as *What is going on? Why did it happen and what will happen next?* A compact summary that represents the development of a story over time, highlighting not only the temporal connections between events but also cause-effect chains, would be very beneficial for providing information that the readers need. Besides, this kind of knowledge, derived from structured information about events and their temporal-causal relations, could be used in a number of applications, from tools for automated generation of timelines to question answering and decision support systems.

While temporal relation classification is a well-studied task with a number of systems participating in the TempEval campaigns (Verhagen et al., 2010; UzZaman et al., 2013; Llorens et al., 2015), less attention has been devoted by the NLP community to the detection of causal links between events. Although recent attempts have tried to settle an annotation standard for causality inspired by TimeML (Mirza et al., 2014), the interactions between the temporal and the causal dimension of texts have been scarcely explored, especially from an empirical point of view. In this work, we face this challenge by presenting CATENA (CAusal and TEmporal relation extraction from NAtural language texts),[1] a multi-sieve architecture for the extraction and classification of both relation types from English documents, which are pre-annotated with *temporal entities*, namely *events* and *time expressions*.

2 Related Work

Our proposed approach for relation extraction is inspired by recent works on hybrid approaches for temporal relation extraction (D'Souza and Ng, 2013; Chambers et al., 2014). D'Souza and Ng (2013) introduce 437 hand-coded rules along with supervised classification models using lexical relation, semantic and discourse features. CAEVO, a CAscading EVent Ordering architecture by Chambers et al. (2014), combines rule-based and data-driven classifiers in a *sieve-based architecture* for temporal ordering. The classifiers (sieves) are ordered by their individual precision, and transitive closure is applied after each sieve to ensure consistent temporal graph.

[1]The system is made available at https://github.com/paramitamirza/CATENA.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 64–75, Osaka, Japan, December 11-17 2016.

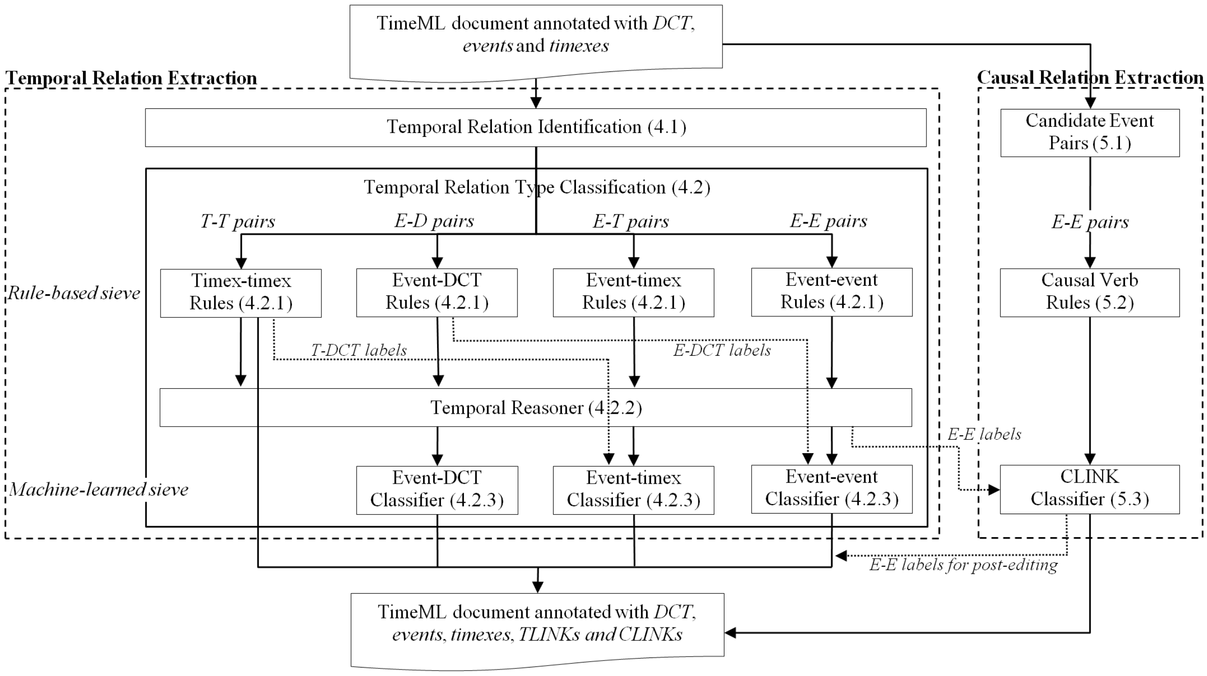

Figure 1: System architecture of CATENA

The problem of detecting causality between events is as challenging as recognizing their temporal order, but less analyzed from an NLP perspective. Besides, previous works have mostly focused on specific types of event pairs and causal expressions in text (Bethard et al., 2008; Do et al., 2011; Riaz and Girju, 2013). Several works, relying on corpus of parallel temporal and causal relations developed with specific connectives in mind (Bethard et al., 2008), have presented analyses on the interaction between temporal and causal relations (Bethard and Martin, 2008; Rink et al., 2010). Exploiting gold temporal labels as features for the causal relation classifier is shown to be beneficial. Mirza et al. (2014) presented some annotation guidelines to capture explicit causality between event pairs, inspired by TimeML. The resulting corpus, Causal-TimeBank, is then used to build supervised classification models for extracting causal relations (Mirza and Tonelli, 2014a). None of the above systems presents a hybrid approach in a sieve-based architecture to deal with this task. CATENA is at present the first integrated system available performing temporal and causal relation extraction.

3 System architecture

The CATENA system includes two main classification modules, one for temporal and the other for causal relations between events. As shown in Figure 1, they both take as input a document annotated with the so-called temporal entities according to TimeML guidelines (Pustejovsky et al., 2003), including the document creation time (DCT), events and time expressions (timexes). The output is the same document with temporal links (TLINKs) set between pairs of temporal entities, each assigned to one of the TimeML temporal relation types, such as BEFORE, INCLUDES or SIMULTANEOUS, which denotes the temporal ordering. The document is also annotated with causal relations (CLINKs) between event pairs.

The modules for temporal and causal relation classification rely both on a sieve-based architecture, in which the remaining unlabelled pairs – after running a rule-based component and/or a transitive reasoner – are fed into a supervised classifier. Although some steps can be run in parallel, the two modules interact, based on the assumption that the notion of causality is tightly connected with the temporal dimension and that information from one module can be used to improve or check the consistency of the other. In particular, (i) TLINK labels for event-event (E-E) pairs, resulting from the rule-based sieve + temporal reasoner modules, are used as features for the CLINK classifier; and (ii) CLINK labels (i.e. CLINK and CLINK-R) are used as a post-editing method for correcting the wrong labelled event pairs by the TLINK classifier. This step relies on a set of rules based on the temporal constraint of causality, i.e. (i) CLINK(e_1, e_2) $\rightarrow$ BEFORE(e_1, e_2) and (ii) CLINK-R(e_1, e_2) $\rightarrow$ AFTER(e_1, e_2). The modules for temporal and causal relation extraction are detailed in Section 4 and 5 respectively.

4 Temporal Relation Extraction System

The module for the extraction of temporal relations contains two main components, one for (i) *temporal relation identification*, which is based on a set of rules, and the other for (ii) *temporal relation type classification*, which is a combination of rule-based and supervised classification modules, with a temporal reasoning component in between. The three steps for temporal relation type classification are ordered based on their individual precisions. This mechanism allows the system to first label few links with high precision using rules, then to infer new links through the reasoner, and finally to increase recall through supervised classification, based on the output of the previous steps.

4.1 Temporal Relation Identification

All pairs of temporal entities satisfying one of the following rules, inspired by the TempEval-3 task description, are considered as having temporal links (TLINKs): (i) two main events of consecutive sentences, (ii) two events in the same sentence, (iii) an event and a timex in the same sentence, (iv) an event and a document creation time and (v) pairs of all possible timexes (including document creation time) linked with each other.[2] These pairs are then grouped together into four different groups: *timex-timex* (T-T), *event-DCT* (E-D), *event-timex* (E-T) and *event-event* (E-E).

4.2 Temporal Relation Type Classification

Our *sieve-based architecture* is inspired by CAEVO (Chambers et al., 2014), although we significantly reduce the system complexity as follows:

- We merge all rule-based classifiers into one sieve component (rule-based sieve), and all Support Vector Machine (SVM) classifiers in the machine-learned sieve.
- Instead of running transitive inference after each classifier, we run our *temporal reasoner* module on the output of the rule-based sieve, only once.

Furthermore, we use the output of the rule-based sieve (Section 4.2.1) as features for the machine-learned sieve (Section 4.2.3), specifically: (i) the timex-DCT link label proposed by the *timex-timex rules* are used as a feature in the *event-timex SVM*, and (ii) the event-DCT link label proposed by the *event-DCT rules* are used as a feature in the *event-event SVM*.

4.2.1 Temporal Rule-Based Sieve

The temporal rule-based sieve relies on specific hand-crafted rules designed for each type of temporal entity pairs, and takes as input the entity pairs identified in the previous step.

Timex-timex Rules For timex-timex relations, we take into account temporal expressions of types DATE and TIME, and determine the relation types based on their *normalized values*. For example, *"7 PM tonight"* (2015-12-12T19:00) IS_INCLUDED in *"today"* (2015-12-12).

Event-DCT Rules The rules for labelling E-D pairs are based on the *tense* and/or *aspect* of the event word. For example, for the event mention *"(had) fallen"*, which is in the past tense with perfective aspect, its relation with the DCT is labelled as BEFORE.

Event-timex Rules As for E-T pairs, we build a set of rules based on the temporal senses of some prepositions (Litkowski and Hargraves, 2006; Litkowski, 2014).[3] In particular we assign a label whenever a temporal preposition establishes a dependency path between an event (E) and a timex (T), in which T acts as the *temporal modifier* of E. For example, if T is introduced by a temporal prepositions expressing a STARTTIME sense such as *from* or *since*, the relation is labelled as BEGUN_BY.

[2]Note that this is not included in the enumerated possible TLINKs in the TempEval-3 task description.

[3]We took the list of temporal prepositions from `http://www.clres.com/db/classes/ClassTemporal.php`.

In the absence of a temporal preposition, T might simply be a temporal modifier of E, as exemplified in *"Police [confirmed]* $_{\mathrm{E}}$ *[Friday]* $_{\mathrm{T}}$ *that the body was found..."*. In this case, we assume that the E-T label is IS_INCLUDED. Moreover, sometimes events are modified by temporal expressions marking the starting time and ending time in a *duration pattern* such as '*between* TBEGIN *and* TEND' or '*from* TBEGIN *to/until* TEND'. We define additional rules as follows: (i) If T matches TBEGIN then E-T label is BEGUN_BY, and (ii) if T matches TEND then E-T label is ENDED_BY.

Event-event Rules E-E pairs are finally labelled following two sets of rules. The first set is based on the dependency path possibly existing between the first (e_1) and the second event (e_2), and the verb information encoded in e_1. For example, if e_2 is the *logical subject* of e_1 as in *"...the chain reaction [touched]* $_{e_1}$ *off by the [collapse]* $_{e_2}$ *of Lehman Brothers"*, e_1 and e_2 are connected by an AFTER relation.

The other set of rules is taken from CAEVO, including: (i) rules for linking a reporting event and another event syntactically dominated by the first, based on *tense* and *aspect*; and (ii) rules based on the role played by various tenses of English verbs in conveying temporal discourse (Reichenbach, 1947).

Further details on the implemented rules for the temporal rule-based sieve can be found in Appendix A.

4.2.2 Temporal Reasoner

Based on the output of the previous sieve, we run a transitive reasoner layer, similar to CAEVO, in order to infer new temporal links among candidate pairs. This alleviates the issue of high precision and low recall, typical of the rule-based sieve.

An annotated TimeML document can be mapped into a constraint problem according to how TLINKs are mapped into Allen relations (Allen, 1983). We apply the following mapping:

- $<$ and $>$ for BEFORE and AFTER
- o and o^{-1} for DURING and DURING_INV
- d and d^{-1} for IS_INCLUDED and INCLUDES
- s and s^{-1} for BEGINS and BEGUN_BY
- f and f^{-1} for ENDS and ENDED_BY

Once the documents are mapped into constraint problems, they are then processed by an automated temporal reasoner for computing their deductive closure, globally reasoning on them. We rely on the *Generic Qualitative Reasoner* (GQR) (Westphal et al., 2010), a fast solver for generic qualitative constraint problems, such as Allen constraint problems. The rationale of preferring GQR to other solutions, such as fast *Boolean Satisfiability Problem* (SAT) solvers, is due to its scalability, simplicity of use and efficient performances (Westphal and Wölfl, 2009).

4.2.3 Temporal Supervised Classifiers

We build three supervised classification models, one for event-DCT (E-D), one for event-timex (E-T) and one for event-event (E-E) pairs. We use LIBLINEAR (Fan et al., 2008) L2-loss linear SVM (default parameters), and one-vs-rest strategy for multi-class classification.

Tools and Resources Several external tools and resources are used to extract features from each temporal entity pair, including:

- *MorphoPro* (Pianta et al., 2008), to get PoS tags and phrase chunk for each token.
- *Mate tools* (Bjorkelund et al., 2010) to extract the dependency path between words.
- *WordNet similarity module*[4] to compute semantic similarity (Lin, 1998) between words.
- *Temporal signal lists* from Mirza and Tonelli (2014b), further expanded using the Paraphrase Database (Ganitkevitch et al., 2013), and manually clustered e.g. {*before*, *prior to*, *in advance of*}.

Feature Set We implemented a set of features, listed in Table 1, largely inspired by the best performing systems in TempEval-2 (Verhagen et al., 2010) and TempEval-3 (UzZaman et al., 2013) campaigns. We simplified the possible values of some features as follows:

[4]http://ws4jdemo.appspot.com/

Feature	TLINK E-D	TLINK E-T	TLINK E-E	CLINK E-E	Rep.	Description
Morphosyntactic information						
PoS	x	x	x	x	one-hot	Part-of-speech tags of e_1 and e_2.
phraseChunk	x	x	x	x	one-hot	Shallow phrase chunk of e_1 and e_2.
samePoS		x	x	x	binary	Whether e_1 and e_2 have the same PoS.
Textual context						
entityOrder		x			binary	Appearance order of e_1 and e_2 in the text.[5]
sentenceDistance		x	x	x	binary	`0` if e_1 and e_2 are in the same sentence, `1` otherwise.
entityDistance		x	x	x	binary	`0` if e_1 and e_2 are adjacent, `1` otherwise.
EVENT attributes						
class	x	x	x	x	one-hot	EVENT attributes as specified in TimeML.
tense	x	x	x	x	one-hot	
aspect	x	x	x	x	one-hot	
polarity	x	x	x	x	one-hot	
sameClass			x	x	binary	Whether e_1 and e_2 have the same EVENT attributes.
sameTenseAspect			x	x	binary	
samePolarity			x	x	binary	
TIMEX3 attributes						
type	x	x			one-hot	TIMEX3 attributes as specified in TimeML.
Dependency information						
dependencyPath			x	x	one-hot	Dependency path between e_1 and e_2.
isMainVerb	x	x	x	x	binary	Whether e_1/e_2 is the main verb of the sentence.
Temporal signals						
tempSignalTokens		x	x	x	one-hot	Tokens (cluster) of temporal signal around e_1 and e_2.
tempSignalPosition		x	x	x	one-hot	Temporal signal position w.r.t e_1/e_2 (BETWEEN, BEFORE, BEGIN, etc.)
tempSignalDependency		x	x	x	one-hot	Temporal signal dependency path between signal tokens and e_1/e_2.
Causal signals						
causSignalTokens				x	one-hot	Tokens (cluster) of causal signal around e_1 and e_2.
causSignalPosition				x	one-hot	Causal signal position w.r.t e_1/e_2 (BETWEEN, BEFORE, BEGIN, etc.)
causSignalDependency			x	x	one-hot	Causal signal dependency path between signal tokens and e_1/e_2.
Lexical semantic information						
wnSim			x	x	one-hot	WordNet similarity computed between the lemmas of e_1 and e_2.
TLINK labels from the rule-based sieve						
timex-DCT label		x			one-hot	The TLINK type of the e_2 (timex) and DCT pair (if any).
event-DCT label			x		one-hot	The TLINK types of the e_1/e_2 and DCT pairs (if any).

Table 1: Feature sets for TLINK classification of event-DCT (E-D), event-timex (E-T) and event-event (E-E) pairs, and for CLINK classifier (E-E pairs), with corresponding feature representation (Rep).

- *dependencyPath* We only consider a dependency path between an event pair if it describes coordination, subordination, subject or object relation.
- *signalTokens* Given a temporal signal, we do not include in the feature set the token but the *clusterID* of the cluster containing synonymous signals, e.g. $\{$*before*, *prior to*, *in advance of*$\}$.
- *wnSim* The value of WordNet similarity measure is discretized as follows: $sim \leq 0.0$, $0.0 < sim \leq 0.5$, $0.5 < sim \leq 1.0$ and $sim > 1.0$.

We exclude lexical features such as *token/lemma* of temporal entities from the feature set in order to increase the classifiers' robustness in dealing with completely new texts with different vocabularies. Instead, we include *WordNet similarity* in the feature set to capture the semantic relations between event words.

Label Simplification For training the classification models, we only consider 10 out of the 14 relation types defined in TimeML by collapsing some types, i.e., IBEFORE into BEFORE, IAFTER into AFTER, DURING and DURING_INV into SIMULTANEOUS, due to the sparse annotation of such labels in the datasets.

5 Causal Relation Extraction System

We propose the same hybrid approach combining rule-based and supervised classifiers for the identification of causal relations. However, while temporal order has a clear formalization in the NLP community, capturing causal relationships in natural language text is more challenging, for they can be expressed by different syntactic and semantic features and involve both situation-specific information and world knowledge. We adopt the notion of causality proposed in the annotation guidelines of the Causal-TimeBank (Mirza et al., 2014; Mirza and Tonelli, 2014a), which accounts for CAUSE, ENABLE and

[5]The order of e_1 and e_2 in E-E pairs is always according to the appearance order in the text, while in E-T pairs, e_2 is always a timex regardless of the appearance order.

PREVENT phenomena (Wolff, 2007; Wolff and Song, 2003) that are overtly expressed in text. In particular, we aim at assigning a causal link to pairs of events when: (i) the causal relation is expressed by *affect*, *link* and *causative verbs* (CAUSE-, ENABLE- and PREVENT-type verbs), hereinafter simply addressed as *causal verbs*; or (ii) the causal relation is marked by a *causal signal* (see e.g. footnote 6).

The two cases require different algorithms: while causal constructions containing causal verbs are quite straightforward to identify, causal signals are very ambiguous and can appear in different syntactic constructions.[6] Therefore, we tackle the first through a rule-based approach, while the second is best covered via supervision, taking advantage of the freely available Causal-TimeBank.

5.1 Causal Relation Identification

Similar to the temporal processing module, the first step towards causal relation classification is the identification of candidate event pairs. Given a document already annotated with events, we take into account every possible combination of events in a sentence in a forward manner as *candidate event pairs*. For example, if we have a sentence "e_1, triggered by e_2, cause them to e_3," the candidate event pairs are (e_1,e_2), (e_1,e_3) and (e_2,e_3). We also include as candidate event pairs the combination of each event in a sentence with events in the following one, to account for inter-sentential causality, under the simplifying assumption that causality may be expressed also between events in two consecutive sentences.

5.2 Causal Rule-Based Sieve

In the rule-based sieve, we classify causal constructions containing causal verbs. These show strong regularities: given a causal verb v, the first event e_1 is usually the *subject* of v and the second event e_2 is either the *object* or the *predicative complement* of v. Such relations between events and causal verbs are usually syntactically expressed, therefore our rules aim at identifying pairs of events being related to a causal verb in a causal construction by looking at their dependency paths.

We take the list of 56 affect, link and causative verbs presented in Mirza et al. (2014) as the causal verb list. We further expand the list using the Paraphrase Database (Ganitkevitch et al., 2013) and original verbs as seeds, resulting in a total of 97 verbs. We then manually cluster the causal verbs sharing the same syntactic behaviour in groups and define a set of rules for each verb group, taking into account the possible existing dependency paths between v and e_1/e_2, as well as the *causal direction sense*[7] conveyed in v. Further details on the implemented rules for the causal rule-based sieve can be found in Appendix B.

5.3 Causal Supervised Classifier

In order to recognize and determine the causal direction of CLINKs that are signalled by a causal signal, we adopt a supervised approach. We build a classification model using LIBLINEAR (Fan et al., 2008) L2-loss linear SVM (default parameters), and one-vs-rest strategy for multi-class classification. The classifier has to label an event pair (e_1, e_2) with CLINK, CLINK-R or O for others.

We take as candidate event pairs only those in which the causal signal is connected via dependency path to either e_1 or e_2, or both. Besides, we exclude event pairs where the two events are directly connected through relations such as subject, object, coordinating or locative adverbial, because a causal relation usually does not hold in these cases.

Tools and Resources The same external tools and resources mentioned in Section 4.2.3 for building the temporal classifiers are used to extract features from each event pair. Additionally, we take the list of causal signals from the annotation guidelines presented in Mirza et al. (2014) as the *causal signal list*. Again we expand the list using the Paraphrase Database (Ganitkevitch et al., 2013), resulting in a total of 200 signals. We also manually cluster some signals together, e.g. {*therefore*, *thereby*, *hence*, *consequently*}, as we did for temporal signals.

[6] *"The building [collapsed]$_T$* ***because of*** *the [earthquake]$_S$"* vs *"**Because of** the [earthquake]$_S$ the building [collapsed]$_T$"*. S and T denote the *source* (cause) and *target* (effect) of the causal relation.

[7] For example, *result in* and *result from* have different senses affecting the causal direction, i.e. the causing event is the subject of *result in* and the object of *result from*.

Feature Set The implemented features are listed in Table 1. As shown in Figure 1, the event-event labels added by the rule-based sieve and the reasoner in the temporal relation extraction module are also used as features for the CLINK classifier.

6 Evaluation

The purpose of the evaluation is two-fold: (i) to evaluate the quality of extracted temporal and causal links separately; and (ii) to investigate the interaction between temporal and causal relation extraction systems in the integrated architecture.

6.1 Temporal and Causal Relation evaluation

We perform two evaluations, one following *TempEval-3* and the other *TimeBank-Dense* evaluation methodology.

Dataset For the evaluation of the temporal relation extraction module following TempEval-3, we use the same training and test data released for the shared task,[8] i.e. *TBAQ-cleaned* (cleaned and improved version of the TimeBank 1.2 and the AQUAINT corpora) and *TempEval-3-platinum*, respectively. The TimeBank 1.2 corpus contains 183 documents coming from a variety of news report, specifically from the ACE program and PropBank, while the AQUAINT corpus contains 73 news report documents and often referred to as the *Opinion corpus*. The TempEval-3-platinum corpus, containing 20 news articles, was annotated/reviewed by the TempEval-3 organizers.

The *TimeBank-Dense* corpus (Chambers et al., 2014) is created to address the sparsity issue in the existing TimeML corpora. The resulting corpus contains 12,715 temporal relations over 36 documents taken from TimeBank 1.2. For the TimeBank-Dense evaluation, we follow the experimental setup in Chambers et al. (2014), in which the TimeBank-Dense corpus is split into a 22 document training set, a 5 document development set and a 9 document test set.[9]

To evaluate the causal relation extraction module, we use the Causal-TimeBank corpus[10] (Mirza and Tonelli, 2014a) for training. For TimeBank-Dense evaluation, the test set is a subset of TimeBank, so we exclude the 9 test documents from Causal-TimeBank during training. For TempEval-3 evaluation, we manually annotated 20 TempEval-3-platinum documents with causal links following the annotation guidelines of the Causal-TimeBank.[11] Causal relations are much sparser than temporal ones, and we found only 26 CLINKs.

Label Adjustment Since the set of TLINK types used in the TimeBank-Dense corpus is slightly different from the one used in TempEval-3,[12] we map the relation types of TLINKs labelled by the rule-based sieve of CATENA (Section 4.2.1) as follows: (i) BEGINS, ENDED_BY → BEFORE, (ii) BEGUN_BY, ENDS → AFTER, and (iii) DURING, IDENTITY → SIMULTANEOUS. The set of labels for the TLINK classifiers (Section 4.2.3) is also adjusted accordingly following the labels in the TimeBank-Dense training data.

Evaluation Results In Table 2, we compare the performance of CATENA with the two best-performing systems participating in the *Task C* of TempEval-3 (relation annotation given gold entities) and *Task C 'relation type only'* (relation annotation given gold entities and related pairs). We also compare the results on the second task with the results of Laokulrat et al. (2015), who recently presented a state-of-the-art system for relation classification based on timegraphs and stacked learning. In CATENA, Task C 'relation type only' is performed by disabling the module for identifying temporal links described in Section 4.1.

The evaluation shows that CATENA is the best performing system in both tasks, even if in Task C best precision and best recall are yielded by Bethard (2013) and Laokulrat et al. (2013), respectively. The recall drop (from .613 to .595) in Task C is because we remove the timex-timex pairs from the final

[8]Available at https://www.cs.york.ac.uk/semeval-2013/task1/index.php\%3Fid=data.html.

[9]Available at http://www.usna.edu/Users/cs/nchamber/caevo/.

[10]Available at http://hlt-nlp.fbk.eu/technologies/causal-timebank.

[11]Available at https://github.com/paramitamirza/CATENA/data/.

[12]Some relation types are not used, and the VAGUE relation introduced in the first TempEval task (Verhagen et al., 2007) is adopted to cope with ambiguous temporal relations, or to indicate pairs for which no clear temporal relation exists. The final set of TLINK types in TimeBank-Dense includes: BEFORE, AFTER, INCLUDES, IS_INCLUDED, SIMULTANEOUS and VAGUE.

	TempEval-3						TimeBank-Dense					
	Task C			Task C rel. type only				T-T	E-D	E-T	E-E	Overall
System	**P**	**R**	**F1**	**P**	**R**	**F1**	**System**	**F1**				**F1**
CATENA	.303	.595	**.402**	**.626**	**.613**	**.619**	**CATENA**	**.780**	.518	**.556**	.487	**.511**
Bethard (2013)	**.373**	.353	.363	-	-	-	CAEVO	.712	**.553**	.494	**.494**	.507
Laokulrat et al. (2013)	.152	**.656**	.247	.556	.574	.565						
Laokulrat et al. (2015)	-	-	-	.576	.579	.578						

Table 2: CATENA evaluated on Tempeval-3 data, compared with the two best participating systems according to UzZaman et al. (2013) and the system by Laokulrat et al. (2015) (left). CATENA is also compared with CAEVO on the TimeBank-Dense test set (right).

	CATENA						CAEVO		
	TempEval-3			TimeBank-Dense			TimeBank-Dense		
Sieve	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Temporal Relation Identification									
	.530	.954	.682	-	-	-	-	-	-
Temporal Relation Type Classification									
RB	.908	.127	.223	.727	.049	.092	-	-	-
RB + TR	.921	.163	.278	.713	.076	.138	-	-	-
ML	.610	.575	.592	.484	.471	.478	.458	.202	.280
RB + ML	.616	.595	.605	.495	.493	.494	.486	.240	.321
RB + TR + ML	**.626**	**.613**	**.619**	**.512**	**.510**	**.511**	.505	.328	.398
RB + TR + ML + AllVague	-	-	-	-	-	-	*.508*	*.506*	*.507*
Causal Relation Extraction									
RB	.917	.423	.579	-	-	-	-	-	-
ML	.429	.115	.182	-	-	-	-	-	-
RB + ML	**.737**	**.538**	**.622**	-	-	-	-	-	-

Table 3: Analysis of classifier performance per sieve. RB: rule-based sieve, ML: machine-learned sieve and TR: temporal reasoner.

annotated documents in order to avoid a relevant decrease in precision, since only very few of such pairs are annotated in the gold standard. The significant drop in precision shows the difficulty in matching annotators' decision to set TLINKs between entity pairs, although CATENA implements the instructions they had to follow in the annotation guidelines.

We also report in Table 2 the performance of CATENA in the TimeBank-Dense evaluation and compare it with CAEVO. We report only F1-score, since all possible links are labelled, yielding the same P and R values. We achieve a small improvement in the overall F1-score, i.e., .511 vs .507. If we consider the different entity pairs, CATENA performs best on timex-timex and event-timex relations, while CAEVO still achieves the best results on event-DCT and event-event pairs. One of the possible reasons for that is the lack of rules in CATENA to classify VAGUE TLINKs between E-E pairs, a relation type present only in TimeBank-Dense.

In order to measure the contribution of each component to the overall performance of CATENA, we also evaluate the performance of each sieve both in the temporal and in the causal module. Results are reported in Table 3, evaluated on both TempEval-3 and TimeBank-Dense test data. As expected, running a transitive closure module after the temporal rule-based sieve (RB + TR) results in improving recall, but the overall performance is still lacking (less than .30 F1-score).

Combining rule-based and machine-learned sieves (RB + ML) yields a slight improvement compared with enabling only the machine-learned sieve in the system (ML). Introducing the temporal reasoner module between the two sieves (RB + TR + ML) proves to be even more beneficial. This is especially evident in the TimeBank-Dense evaluation. The same phenomena are also observed by CAEVO; Table 3 (right) shows the related numbers reported in Chambers et al. (2014). Note that in CAEVO, the machine-learned sieves are not the last sieves, instead, the *AllVague* sieve is finally activated to label all remaining unlabelled pairs as VAGUE.

For causal relation extraction, the combination of rule-based and machine-learned sieves (RB + ML) achieves .622 F1-score in TempEval-3 evaluation, with the ML component contributing to increase

E-E pair	Sentence	TE3-gold	TE-label	CA-label	Post-editing
(e_{32}, e_{44})	*The [incident]* $_{e_{32}}$ *provoked an international [outcry]* $_{e_{44}}$*...*	-	SIMULTANEOUS	CLINK	BEFORE
(e_{32}, e_{45})	*The [incident]* $_{e_{32}}$ *provoked an international outcry and led to a major [deterioration]* $_{e_{45}}$ *in relations...*	-	AFTER	CLINK	BEFORE
(e_{18}, e_{19})	*...the [inspections]* $_{e_{18}}$ *were directly linked to the new law on NGOs and the targeted groups' [compliance]* $_{e_{19}}$ *with it.*	-	IS_INCLUDED	CLINK-R	AFTER
(e_4, e_6)	*A haze akin to volcanic fumes [cloaked]* $_{e_4}$ *the capital, causing convulsive [coughing]* $_{e_6}$ *and...*	INCLUDES	AFTER	CLINK	BEFORE

Table 4: Examples of E-E pairs in the TempEval-3-platinum dataset with gold annotated labels (TE3-gold), labelled by the temporal module (TE-label) and causal module (CA-label) of CATENA. These examples illustrate how TLINK post-editing using CLINK could improve the labelling quality.

the recall of the highly precise RB component. The low precision of the ML module is mostly due to dependency parsing mistakes and issues in disambiguating signals such as *from*, as in *"...passenger cars in China was on track to hit [400 million]* $_{\mathrm{T}}$ *by 2030, up* ***from*** *[90 million]* $_{\mathrm{S}}$ *now."* Unfortunately, from the total of 5 gold CLINKs in the 20 documents of the TimeBank-Dense test set, none is identified by CATENA.

6.2 Interaction between Temporal and Causal Relations

As shown in Figure 1, E-E labels returned by the temporal reasoner are used by the CLINK classifier as features, whose causal relations are then used to post-edit TLINK labels. We evaluate the impact of the first step through an ablation test, by removing TLINK types from the features used by the CLINK classifier. We only analyse the results of TempEval-3 evaluation, since there are no causal links recognized in the TimeBank-Dense test corpus. Without TLINK types, the F1-score drops from .622 to .571, with a significant recall drop from .538 to .462. This shows that temporal information is beneficial to the classification of causal relations between events, especially in terms of recall.

As for the evaluation of TLINK post-editing using CLINKs, the system identifies 19 causal links in the test set, which are passed to the temporal module. While 15 of them are already consistent with BEFORE/AFTER labels, 3 would add new correct TLINKs that are currently not annotated in the evaluation corpus, and were wrongly labelled by the temporal module of CATENA, as shown in Table 4. The fourth would add a BEFORE relation between *cloaked* and *coughing* in *"A haze akin to volcanic fumes [cloaked]* $_{\mathrm{S}}$ *the capital,* ***causing*** *convulsive [coughing]* $_{\mathrm{T}}$ *..."*. This relation is labelled as INCLUDES in the gold standard, but we believe that BEFORE would be correct as well.

7 Conclusions

We presented CATENA, a hybrid system for the extraction and classification of temporal and causal relations in text, which we make freely available to the research community. We adopt a sieve-based architecture both for the temporal and the causal module, integrating rule-based and machine learning components. The two modules were evaluated separately, showing that they achieve state-of-the-art performance on different tasks. Furthermore, the interaction between temporal and causal components, especially the benefits of passing information from one module to the other, was also analysed.

The system relies on the notion of events as defined in the TimeML standard, making it possible to easily put temporal and causal information in relation. Although the interplay between causality and temporality may seem obvious from a theoretical point of view, CATENA allows a systematic study and a quantification of this phenomenon. The presented approach would probably have more impact if implicit causality was also considered, which we did not take into account because it is not annotated in the Causal-TimeBank corpus. However, we plan to investigate this issue in the near future.

Acknowledgments

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404).

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November.

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT, Short Papers*, pages 177–180, Columbus, Ohio, June. Association for Computational Linguistics.

Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Anders Bjorkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

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.

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

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 918–927, Atlanta, Georgia, June. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT 2013*, pages 758–764, Atlanta, Georgia, June. ACL.

Natsuda Laokulrat, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. Uttime: Temporal relation classification using deep syntactic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 88–92, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Natsuda Laokulrat, Makoto Miwa, and Yoshimasa Tsuruoka. 2015. Stacking approach to temporal relation classification with temporal inference. *Journal of Natural Language Processing*, 22(3):171–196.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kenneth C. Litkowski and Orin Hargraves. 2006. Coverage and inheritance in the preposition project. In *3rd ACL-SIGSEM Workshop on Prepositions*.

Ken Litkowski. 2014. Pattern dictionary of english prepositions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1274–1283, Baltimore, Maryland, June. Association for Computational Linguistics.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. Semeval-2015 task 5: Qa tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado, June. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014a. 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*, pages 2097–2106, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014b. Classifying temporal relations with simple features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317, Gothenburg, Sweden, April. Association for Computational Linguistics.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden, April. Association for Computational Linguistics.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*.

H Reichenbach. 1947. *Elements of symbolic logic*. University of California Press, Berkeley, CA.

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 SIGDIAL 2013 Conference*, pages 21–30, Metz, France, August. Association for Computational Linguistics.

Bryan Rink, Cosmin Adrian Bejan, and Sanda M. Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *FLAIRS Conference*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.

Matthias Westphal and Stefan Wölfl. 2009. Qualitative CSP, finite CSP, and SAT: Comparing Methods for Qualitative Constraint-based Reasoning. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, IJCAI'09, pages 628–633, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Matthias Westphal, Stefan Wölfl, and Jason Jingshi Li. 2010. Restarts and nogood recording in qualitative constraint-based reasoning. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, pages 1093–1094.

Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332.

Phillip Wolff. 2007. Representing causation. *Journal of experimental psychology: General*, 136(1):82–111.

Appendix A Temporal Rule Set

tense	***aspect***	**E-D label**
PAST	PERFECTIVE	BEFORE
PRESENT	PROGRESSIVE	INCLUDES
PRESENT	PERFECTIVE_PROGRESSIVE	INCLUDES
FUTURE	*	AFTER

Table 5: E-D label rules based on *tense* and *aspect* of E.

tsense	**E-T label**
TIMEPOINT (e.g. *in*, *at*, *on*)	IS_INCLUDED
TIMEPRECEDING (e.g. *before*)	BEFORE
TIMEFOLLOWING (e.g. *after*)	AFTER
DURATION (e.g. *during*, *throughout*)	DURING
STARTTIME (e.g. *from*, *since*)	BEGUN_BY
ENDTIME (e.g. *until*)	ENDED_BY

Table 6: E-T label rules based on the sense of temporal preposition (*tsense*) introducing T.

dep	**e_1 *verb info***	**E-E label**	**Example**
LGS-PMOD	*	AFTER	"...reaction *[touched]* $_{e_1}$ off by the *[collapse]* $_{e_2}$ of..."
LOC-PMOD	*	IS_INCLUDED	"...enormous *[surge]* $_{e_1}$ in coal *[consumption]* $_{e_2}$..."
OPRD-IM/OPRD	aspectual verb for *initiation*	BEGINS	"...situation *[began]* $_{e_1}$ to *[relax]* $_{e_2}$ in..."
	aspectual verb for *culmination/termination*	ENDS	"...we 'd *[stop]* $_{e_1}$ *[bidding]* $_{e_2}$."
	aspectual verb for *continuation*	INCLUDES	"...industry 's growth *[continues]* $_{e_1}$ to *[slow]* $_{e_2}$."
	general verb, *aspect*=PERFECTIVE_PROGRESSIVE	SIMULTANEOUS	"...have been *[working]* $_{e_1}$ to *[develop]* $_{e_2}$ quantum..."
	general verb	BEFORE	"...consortium *[attempted]* $_{e_1}$ to *[block]* $_{e_2}$..."

Table 7: E-E label rules based on dependency path (*dep*) and verb information of e_1 (e_1 *verb info*).

Appendix B Causal Rule Set

v	***dep***$_1$	***dep***$_2$	***dir***	**E-E label**
AFFECT	(*)	OBJ		CLINK
LINK	(*)	OBJ/ADV-PMOD/DIR-PMOD/AMOD-PMOD	CLINK	CLINK
			CLINK-R	CLINK-R
CAUSE/ENABLE/PREVENT	(*)	OBJ/OPRD/OPRD-IM/ADV-PMOD		CLINK
		LGS-PMOD		CLINK-R
CAUSE-/ENABLE-/PREVENT-AMBIGUOUS	(*)	OPRD/OPRD-IM/ADV-PMOD		CLINK

Table 8: Causal verb rules for E-E pairs based on causal verb (v) category, dependency paths between v and e_1/e_2, and causal direction sense (*dir*). (*) denotes all possible dependency paths listed in Table 9.

Relation	**Path**	**Example**
between v and e_1	***dep1***	
e_1 is subject of v	SBJ	*The Pope's [visit]* $_{e_1}$ ***persuades*** $_v$ *Cubans...*
v is predicative complement of e_1	PRD-IM	*The [roundup]* $_{e_1}$ *was to* ***prevent*** $_v$ *them...*
v is modifier of e_1 (nominal)	NMOD	*An [agreement]* $_{e_1}$ *that* ***permits*** $_v$ *the Russian...*
v is apposition of e_1	APPO	*..., with the [crisis]* $_{e_1}$ ***triggered*** $_v$ *by...*
v is general adverbial of e_1	ADV	*The number [increased]* $_{e_1}$, ***prompting*** $_v$*...*
v is adverbial of purpose/reason of e_1	PRP-IM	*The major [allocated]* $_{e_1}$ *funds to* ***help*** $_v$*...*
between v and e_2	***dep2***	
e_2 is object of v	OBJ	*...have* ***provoked*** $_v$ *widespread [violence]* $_{e_2}$.
e_2 is logical subject of v (passive verb)	LGS-PMOD	*...**triggered*** $_v$ *by the [end]* $_{e_2}$ *of the...*
e_2 is predicative complement of v (raising/control verb)	OPRD	*...funds to* ***help*** $_v$ *[build]* $_{e_2}$ *a museum.*
	OPRD-IM	*...**persuades*** $_v$ *Cubans to [break]* $_{e_2}$ *loose.*
e_2 is general adverbial of v	ADV-PMOD	*...**protect*** $_v$ *them from unspecified [threats]* $_{e_2}$.
e_2 is adverbial of direction of v	DIR-PMOD	*...**lead to*** $_v$ *a [surge]* $_{e_2}$ *of inexpensive imports.*
e_2 is modifier of v (adjective or adverbial)	AMOD-PMOD	*...**related to*** $_v$ *[problems]* $_{e_2}$ *under a contract.*

Table 9: Dependency paths considered for setting a causal link between two events e_1 and e_2 when a causal verb v is present.