Annotating causality in the TempEval-3 corpus

Paramita Mirza
FBK, Trento, Italy
University of Trento
paramita@fbk.eu

Rachele Sprugnoli
FBK, Trento, Italy
University of Trento
sprugnoli@fbk.eu

Sara Tonelli
FBK, Trento, Italy
satonelli@fbk.eu

Manuela Speranza
FBK, Trento, Italy
manspera@fbk.eu

Abstract

While there is a wide consensus in the NLP community over the modeling of temporal relations between events, mainly based on Allen's temporal logic, the question on how to annotate other types of event relations, in particular causal ones, is still open. In this work, we present some annotation guidelines to capture causality between event pairs, partly inspired by TimeML. We then implement a rule-based algorithm to automatically identify explicit causal relations in the TempEval-3 corpus. Based on this annotation, we report some statistics on the behavior of causal cues in text and perform a preliminary investigation on the interaction between causal and temporal relations.

1 Introduction

The annotation of events and event relations in natural language texts has gained in recent years increasing attention, especially thanks to the development of TimeML annotation scheme (Pustejovsky et al., 2003), the release of TimeBank (Pustejovsky et al., 2006) and the organization of several evaluation campaigns devoted to automatic temporal processing (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013).

However, while there is a wide consensus in the NLP community over the modeling of temporal relations between events, mainly based on Allen's interval algebra (Allen, 1983), the question on how to model other types of event relations is still open. In particular, linguistic annotation of causal relations, which have been widely investigated from a philosophical and logical point of view, are still under debate. This leads, in turn, to the lack of a standard benchmark to evaluate causal relation extraction systems, making it difficult to compare systems performances, and to identify the state-of-the-art approach for this particular task.

Although several resources exist in which causality has been annotated, they cover only few aspects of causality and do not model it in a global way, comparable to what as been proposed for temporal relations in TimeML. See for instance the annotation of causal arguments in PropBank (Bonial et al., 2010) and of causal discourse relations in the Penn Discourse Treebank (The PDTB Research Group, 2008).

In this work, we propose annotation guidelines for causality inspired by TimeML, trying to take advantage of the clear definition of events, signals and relations proposed by Pustejovsky et al. (2003). Besides, as a preliminary investigation of causality in the TempEval-3 corpus, we perform an automatic analysis of causal signals and relations observed in the corpus. This work is a first step towards the annotation of the TempEval-3 corpus with causality, with the final goal of investigating the strict connection between temporal and causal relations. In fact, there is a temporal constraint in causality, i.e. the cause must occur BEFORE the effect. We believe that investigating this precondition on a corpus basis can contribute to improving the performance of temporal and causal relation extraction systems.

2 Existing resources on Causality

Several attempts have been made to annotate causal relations in texts. A common approach is to look for specific cue phrases like *because* or *since* or to look for verbs that contain a cause as part of their meaning, such as *break* (*cause to be broken*) or *kill* (*cause to die*) (Khoo et al., 2000; Sakaji et al., 2008; Girju et al., 2007). In PropBank (Bonial et al., 2010), causal relations are annotated in the form of predicate-argument relations, where ARGM-CAU is used to annotate "the reason for an action", for example: *"They* [PREDICATE ***moved***] *to London* [ARGM-CAU ***because of the baby***]*."*

Another scheme annotates causal relations between discourse arguments, in the framework of

Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language, pages 10–19,
Gothenburg, Sweden, April 26, 2014.

the Penn Discourse Treebank (PDTB). As opposed to PropBank, this kind of relations holds only between clauses and do not involve predicates and their arguments. In PDTB, the *Cause* relation type is classified as a subtype of CONTINGENCY.

Causal relations have also been annotated as relations between events in a restricted set of linguistic constructions (Bethard et al., 2008), between clauses in text from novels (Grivaz, 2010), or in noun-noun compounds (Girju et al., 2007).

Several types of annotation guidelines for causal relations have been presented, with varying degrees of reliability. One of the simpler approaches asks annotators to check whether the sentence they are reading can be paraphrased using a connective phrase such as *and as a result* or *and as a consequence* (Bethard et al., 2008).

Another approach to annotate causal relations tries to combine linguistic tests with semantic reasoning tests. In Grivaz (2010), the linguistic paraphrasing suggested by Bethard et al. (2008) is augmented with rules that take into account other semantic constraints, for instance if the potential cause occurs before or after the potential effect.

3 Annotation of causal information

As part of a wider annotation effort aimed to annotate texts at the semantic level (Tonelli et al., 2014), we propose guidelines for the annotation of causal information. In particular, we define causal relations between events based on the TimeML definition of events (ISO TimeML Working Group, 2008), as including all types of actions (punctual and durative) and states. Syntactically, events can be realized by a wide range of linguistic expressions such as verbs, nouns (which can realize eventualities in different ways, for example through a nominalization process of a verb or by possessing an eventive meaning), and prepositional constructions.

Following TimeML, our annotation of events involved in causal relations includes the `polarity` attribute (see Section 3.3); in addition to this, we have defined the `factuality` and `certainty` event attributes, which are useful to infer information about actual causality between events.

Parallel to the TimeML tag <SIGNAL> as an indicator for temporal links, we have also introduced the notion of causal signals through the use of the <C-SIGNAL> tag.

3.1 C-SIGNAL

The <C-SIGNAL> tag is used to mark-up textual elements that indicate the presence of a causal relation (i.e. a CLINK, see 3.2). Such elements include all causal uses of:

- prepositions, e.g. *because of, on account of, as a result of, in response to, due to, from, by*;
- conjunctions, e.g. *because, since, so that, hence, thereby*;
- adverbial connectors, e.g. *as a result, so, therefore, thus*;
- clause-integrated expressions, e.g. *the result is, the reason why, that's why*.

The extent of C-SIGNALs corresponds to the whole expression, so multi-token extensions are allowed.

3.2 CLINK (Causal Relations)

For the annotation of causal relations between events, we use the <CLINK> tag, a directional one-to-one relation where the causing event is the *source* (the first argument, indicated as $_{S}$ in the examples) and the caused event is the *target* (the second argument, indicated as $_{T}$). The annotation of CLINKs includes the `c-signalID` attribute, whose value is the ID of the C-SIGNAL indicating the causal relation (if available).

A seminal research in cognitive psychology based on the force dynamics theory (Talmy, 1988) has shown that causation covers three main kinds of causal concepts (Wolff, 2007), which are CAUSE, ENABLE, and PREVENT, and that these causal concepts are lexicalized as verbs (Wolff and Song, 2003): (i) CAUSE-type verbs: *bribe, cause, compel, convince, drive, have, impel, incite, induce, influence, inspire, lead, move, persuade, prompt, push, force, get, make, rouse, send, set, spur, start, stimulate*; (ii) ENABLE-type verbs: *aid, allow, enable, help, leave, let, permit*; (iii) PREVENT-type verbs: *bar, block, constrain, deter, discourage, dissuade, hamper, hinder, hold, impede, keep, prevent, protect, restrain, restrict, save, stop*. CAUSE, ENABLE, and PREVENT categories of causation and the corresponding verbs are taken into account in our guidelines.

As causal relations are often not overtly expressed in text (Wolff et al., 2005), we restrict the annotation of CLINKs to the presence of an explicit

causal construction linking two events in the same sentence[1], as detailed below:

- **Basic constructions** for CAUSE, ENABLE and PREVENT categories of causation as shown in the following examples:
  *The purchase$_S$ **caused** the creation$_T$ of the current building*
  *The purchase$_S$ **enabled** the diversification$_T$ of their business*
  *The purchase$_S$ **prevented** a future transfer$_T$*

- Expressions containing **affect verbs**, such as *affect*, *influence*, *determine*, and *change*. They can be usually rephrased using *cause*, *enable*, or *prevent*:
  *Ogun ACN crisis$_S$ **affects** the launch$_T$ of the All Progressives Congress → Ogun ACN crisis **causes/enables/prevents** the launch of the All Progressives Congress*

- Expressions containing **link verbs**, such as *link*, *lead*, and *depend on*. They can usually be replaced only with *cause* and *enable*:
  *An earthquake$_T$ in North America was **linked** to a tsunami$_S$ in Japan → An earthquake in North America was **caused/enabled** by a tsunami in Japan*
  An earthquake in North America was **prevented by a tsunami in Japan*

- **Periphrastic causatives** are generally composed of a verb that takes an embedded clause or predicate as a complement; for example, in the sentence *The blast$_S$ caused the boat to heel$_T$ violently*, the verb (i.e. *caused*) expresses the notion of CAUSE while the embedded verb (i.e. *heel)* expresses a particular result. Note that the notion of CAUSE can be expressed by verbs belonging to the three categories previously mentioned (which are CAUSE-type verbs, ENABLE-type verbs and PREVENT-type verbs).

- Expressions containing **causative conjunctions and prepositions** as listed in Section 3.1. Causative conjunctions and prepositions are annotated as C-SIGNALs and their ID is to be reported in the `c-signalID` attribute of the CLINK.[2]

In some contexts, the coordinating conjunction *and* can imply causation; given the ambiguity of this construction and the fact that it is not an explicit causal construction, however, we do not annotate CLINKs between two events connected by *and*. Similarly, the temporal conjunctions *after* and *when* can also implicitly assert a causal relation but should not be annotated as C-SIGNALs and no CLINKs are to be created (temporal relations have to be created instead).

3.3 Polarity, factuality and certainty

The `polarity` attribute, present both in TimeML and in our guidelines, captures the grammatical category that distinguishes affirmative and negative events. Its values are NEG for events which are negated (for instance, the event *cause* in *Serotonin deficiency$_S$ may **not cause** depression$_T$*) and POS otherwise.

The annotation of `factuality` that we added to our guidelines is based on the situation to which an event refers. FACTUAL is used for *facts*, i.e. situations that have happened, COUNTERFACTUAL is used for *counterfacts*, i.e. situations that have no real counterpart as they did not take place, NON-FACTUAL is used for *possibilities*, i.e. speculative situations, such as future events, events for which it is not possible to determine whether they have happened, and general statements.

The `certainty` attribute expresses the binary distinction between certain (value CERTAIN) and uncertain (value UNCERTAIN) events. Uncertain events are typically marked in the text by the presence of modals or modal adverbs (e.g. *perhaps*, *maybe*) indicating possibility. In the sentence *Drinking$_S$ may **cause** memory loss$_T$*, the causal connector *cause* is an example of a NON-FACTUAL and UNCERTAIN event.

In the annotation algorithm presented in the following section, only the `polarity` attribute is taken into account, given that information about factuality and certainty of events is not annotated in the TempEval-3 corpus. In particular, at the time of the writing the algorithm considers only the polarity of causal verbal connectors, because this information is necessary to extract causal chains

[1] A typical example of implicit causal construction is represented by lexical causatives; for example, *kill* has the embedded meaning of causing someone to die (Huang, 2012). In the present guidelines, these cases are not included.

[2] The absence of a value for the `c-signalID` attribute means that the causal relation is encoded by a verb.

between events in a text. However, adding information on the polarity of the single events involved in the relations would make possible also the identification of positive and negative causes and effects.

4 Automatic annotation of explicit causality between events

In order to verify the soundness of our annotation framework for event causality, we implement some simple rules based on the categories and linguistic cues listed in Section 3. Our goal is two-fold: first, we want to check how accurate rule-based identification of (explicit) event causality can be. Second, we want to have an estimate of how frequently causality can be explicitly found in text.

The dataset we annotate has been released for the TempEval-3 shared task[3] on temporal and event processing. The TBAQ-cleaned corpus is the training set provided for the task, consisting of the TimeBank (Pustejovsky et al., 2006) and the AQUAINT corpora. It contains around 100K words in total, with 11K words annotated as events (UzZaman et al., 2013). We choose this corpus because gold events are already provided, and because it allows us to perform further analyses on the interaction between temporal and causal relations.

Our automatic annotation pipeline takes as input the TBAQ-cleaned corpus with gold annotated events and tries to automatically recognize whether there is a causal relation holding between them. The annotation algorithm performs the following steps in sequence:

1. The TBAQ-cleaned corpus is PoS-tagged and parsed using the Stanford dependency parser (de Marneffe and Manning, 2008).

2. The corpus is further analyzed with the *addDiscourse* tagger (Pitler and Nenkova, 2009), which automatically identifies explicit discourse connectives and their sense, i.e. EXPANSION, CONTINGENCY, COMPARISON and TEMPORAL. This is used to disambiguate causal connectives (e.g. we consider only the occurrences of *since* when it is a causal connective, meaning that it falls into CONTINGENCY class instead of TEMPORAL).

3. Given the list of *affect*, *link*, *causative* verbs (basic and periphrastic constructions) and *causal signals* listed in Sections 3.1 and 3.2, the algorithm looks for specific dependency constructions where the causal verb or signal is connected to two events, as annotated in the TBAQ-cleaned corpus.

4. If such dependencies are found, a CLINK is automatically set between the two events identifying the source ($_S$) and the target ($_T$) of the relation.

5. When a causal connector corresponds to an event, the algorithm uses the polarity of the event to assign a polarity to the causal link.

Specific approaches to detect when ambiguous connectors have a causal meaning are implemented, as in the case of *from* and *by*, where the algorithm looks for specific structures. For instance, in *"The building was* $\underline{damaged}_T$ ***by*** *the* $\underline{earthquake}_S$*"*, *by* is governed by a passive verb annotated as event.

Also the preposition *due to* is ambiguous as shown in the following sentences where it acts as a causal connector only in b):

a) *It had been* ***due to*** $\underline{expire}$ *Friday evening.*

b) *It* $\underline{cut}_T$ *the dividend* ***due to*** *its third-quarter* $\underline{loss}_S$ *of $992,000.*

The algorithm performs the disambiguation by checking the dependency structures: in sentence a) there is only one dependency relation *xcomp(due, expire)*, while in sentence b) the dependency relations are *xcomp(cut, due)* and *prep_to(due, loss)*. Besides, both *cut* and *loss* are annotated as events.

We are aware that this type of automatic annotation may be prone to errors because it takes into account only a limited list of causal connectors. Besides, it only partially accounts for possible ambiguities of causal cues and may suffer from parsing errors. However, this allows us to make some preliminary remarks on the amount of causal information found in the TempEval-3 corpus. Some statistics are reported in the following subsection.

4.1 Statistics of Automatic Annotation

Basic construction. In Table 1 we report some statistics on the non-periphrastic structures identified starting from verbs expressing the three categories of causation. Note that for the verbs *have*, *start*, *hold* and *keep*, even though they connect two events, we cannot say that there is always a causal relation between them, as exemplified in the following sentence taken from the corpus:

a) *Gen. Schwarzkopf secretly* $\underline{picked}_S$ *Saturday*

[3] http://www.cs.york.ac.uk/semeval-2013/task1/

*night as the optimal time to **start** the offensive$_{T}$.*
b) *On Tuesday, the National Abortion and Reproductive Rights Action League plans$_{S}$ to **hold** a news conference$_{T}$ to screen a TV advertisement.*

Types	Verbs	CLINK
CAUSE	have	1
	start	2
	cause	1
	compel	1
PREVENT	hold	1
	keep	3
	block	7
	prevent	1
ENABLE	-	-
Total		**17**

Table 1: Statistics of CLINKs with basic construction

Affect verbs. The algorithm does not annotate any causal relation containing affect verbs mostly because the majority of the 36 affect verb occurrences found in the corpus connect two elements that are not events, as in *"These big stocks greatly **influence** the Nasdaq Composite Index."*

Link verbs. In total, we found 50 occurrences of link verbs in the corpus, but the algorithm identifies only 4 causal links. Similar to affect verbs, this is mainly due to the fact that two events are not found to be involved in the relation. For instance, the system associated only one CLINK to *link* (out of 12 occurrences of the verb) and no CLINKs to *depend* (which occurs 3 times). Most of the CLINKs identified are signaled by the verb *lead*; for example, *"Pol Pot is considered responsible for the radical policies$_{S}$ that **led** to the deaths$_{T}$ of as many as 1.7 million Cambodians."*

Periphrastic causative verbs. Overall, there are around 1K potential occurrences of periphrastic causative verbs in the corpus. However, the algorithm identifies only around 14% of them as part of a periphrastic construction, as shown in Table 2. This is because some verbs are often used in non-periphrastic structures, e.g. *make*, *have*, *get*, *keep* and *hold*. Among the 144 cases of periphrastic constructions, 41 causal links are found by our rules.

In Table 2, for each verb type, we report the list of verbs that appear in periphrastic constructions in the corpus, specifying the number of CLINKs identified by the system for each of them.

Some other CAUSE-type (*move, push, drive, influence, compel, spur*), PREVENT-type (*hold, save, impede, deter, discourage, dissuade, restrict*) and ENABLE-type (*aid*) verbs occur in the corpus but are not involved in periphrastic structures. Some others do not appear in the corpus at all (*bribe, impel, incite, induce, inspire, rouse, stimulate, hinder, restrain*).

Types	Verbs	Periphr.	CLINK	All
CAUSE	have	34	0	239
	make	6	2	125
	get	1	0	50
	lead	2	1	38
	send	5	1	34
	set	2	0	23
	start	1	0	22
	force	2	1	15
	cause	3	2	12
	prompt	3	2	6
	persuade	2	1	3
	convince	1	1	2
PREVENT	keep	1	1	58
	stop	3	0	24
	block	2	2	21
	protect	2	1	15
	prevent	6	2	12
	hamper	1	0	2
	bar	1	0	1
	constrain	1	0	1
ENABLE	help	31	13	45
	leave	2	2	45
	allow	22	3	39
	permit	2	1	6
	enable	4	2	5
	let	4	3	5
Total		**144**	**41**	**848**

Table 2: Statistics of periphrastic causative verbs

Causal signals. Similar to periphrastic causative verbs, out of around 1.2K potential causal connectors found in the corpus, only 194 are automatically recognized as actual causal signals after disambiguation, as detailed in Table 3. Based on these identified causal signals, the algorithm derives 111 CLINKs.

Even though the *addDiscourse* tool labels 11 occurrences of the adverbial connector *so* as having a causal meaning, our algorithm does not annotate any CLINKs for such connector. In most cases, it is because it acts as an inter-sentential connector, while we limit the annotation of CLINKs only to events occurring within the same sentence.

CLINKs polarity. Table 4 shows the distribution of the positive and negative polarity of the detected CLINKs.

Only two cases of negated CLINKs are automatically identified in the corpus. One example is the following: *"Director of the U.S. Federal Bureau of*

Types	C-SIGNALs	Causal	CLINK	All
prep.	because of	32	11	32
	on account of	0	0	0
	as a result of	13	9	13
	in response to	7	1	7
	due to	*2*	*1*	*6*
	from	*2*	*2*	*500*
	by	*23*	*24*	*465*
conj.	because	58	37	58
	since	*26*	*19*	*72*
	so that	5	4	5
adverbial	as a result	3	0	3
	so	*11*	*0*	*69*
	therefore	4	0	4
	thus	6	2	6
	hence	0	0	0
	thereby	1	0	1
	consequently	1	1	1
clausal	the result is	0	0	0
	the reason why	0	0	0
	that is why	0	0	0
Total		**194**	**111**	**1242**

Table 3: Statistics of causal signals in CLINKs

*Investigation (FBI) Louis Freeh said here Friday that U.S. air raid$_{\text{T}}$ on Afghanistan and Sudan is **not** directly **linked** with the probe$_{\text{S}}$ into the August 7 bombings in east Africa."*

Connector types		POS	NEG
Basic	CAUSE	5	0
	PREVENT	12	0
	ENABLE	-	-
Affect verbs		-	-
Link verbs		3	1
Periphrastic	CAUSE	10	1
	PREVENT	6	0
	ENABLE	24	0
Total		**60**	**2**

Table 4: Statistics of CLINKs' polarity

CLINKs vs TLINKs. In total, the algorithm identifies 173 CLINKs in the TBAQ-cleaned corpus, while the total number of TLINKs between pairs of events is around 5.2K. For each detected CLINK between an event pair, we identify the underlying temporal relations (TLINKs) if any. We found that from the total of CLINKs extracted, around 33% of them have an underlying TLINK, as detailed in Table 5. Most of them are CLINKs signaled by causal signals.

For causative verbs, the *BEFORE* relation is the only underlying temporal relation type, with the exception of one *SIMULTANEOUS* relation.

As for C-SIGNALs, the distribution of temporal relation types is less homogeneous, as shown in Table 6. In most of the cases, the underlying temporal relation is *BEFORE*. In few cases, CLINKs signaled by the connector *because* overlap with an *AFTER* relation, as in *"But some analysts questioned$_{\text{T}}$ how much of an impact the retirement package will have, **because** few jobs will end$_{\text{S}}$ up being eliminated."*

Connector types		CLINK	TLINK
Basic	CAUSE	5	2
	PREVENT	12	0
	ENABLE	-	-
Affect verbs		-	-
Link verbs		4	1
Periphrastic	CAUSE	11	1
	PREVENT	6	0
	ENABLE	24	0
C-SIGNALs		111	54
Total		**173**	**58**

Table 5: Statistics of CLINKs' overlapping with TLINKs

In some cases, CLINKs signaled by the connector *since* match with a *BEGINS* relation. This shows that *since* expresses merely a temporal and not a causal link. As it has been discussed before, the connector *since* is highly ambiguous and the CLINK has been wrongly assigned because of a disambiguation mistake of the addDiscourse tool.

5 Evaluation

We perform two types of evaluation. The first is a qualitative one, and is carried out by manually inspecting the 173 CLINKs that have been automatically annotated. The second is a quantitative evaluation, and is performed by comparing the automatic annotated data with a gold standard corpus of 100 documents taken from TimeBank.

5.1 Qualitative Evaluation

The automatically annotated CLINKs have been manually checked in order to measure the precision of the adopted procedure. Out of 173 annotated CLINKs, 105 were correctly identified obtaining a precision of 0.61.

Details on precision calculated on the different types of categories and linguistic cues defined in Section 3.2 are provided in Table 7. Statistics show that performances vary widely depending on the category and linguistic cue taken into consideration. In particular, relations expressing causation of PREVENT type prove to be extremely difficult to be correctly detected with a rule-based approach: the algorithm precision is 0.25 for basic constructions and 0.17 for periphrastic constructions.

During the manual evaluation, two main types

C-SIGNALs	BEFORE	AFTER	IS_INCLUDED	BEGINS	others
because of	5	-	-	-	-
as a result of	2	-	-	-	-
in response to	1	-	-	-	-
due to	1	-	-	-	-
by	11	-	1	2	3
because	14	2	1	-	1
since	4	1	-	3	-
so that	1	-	-	-	-
thus	1	-	-	-	-
Total	**40**	**3**	**2**	**5**	**4**

Table 6: Statistics of CLINKs triggered by C-SIGNALs overlapping with TLINKs

Connector types		Extracted	Correct	P
Basic	CAUSE	5	3	0.60
	PREVENT	12	3	0.25
	ENABLE	0	n.a.	n.a.
Affect Verbs		0	n.a.	n.a.
Link Verbs		4	3	0.75
Periphrastic	CAUSE	11	8	0.73
	PREVENT	6	1	0.17
	ENABLE	24	17	0.71
C-SIGNALs		111	70	0.63
Total		**173**	**105**	**0.61**

Table 7: Precision of automatically annotated CLINKs

of mistakes have been observed: the wrong identification of events involved in CLINKs and the annotation of sentences that do not contain causal relations.

The assignment of a wrong source or a wrong target to a CLINK is primarily caused by the dependency parser output that tends to establish a connection between a causal verb or signal and the closest previous verb. For example, in the sentence *"StatesWest Airlines said it* $\underline{withdrew}_{\mathrm{T}}$ *its offer to* $\underline{acquire}$ *Mesa Airlines* ***because*** *the Farmington carrier did not* $\underline{respond}_{\mathrm{S}}$ *to its offer"*, the CLINK is annotated between *respond* and *acquire* instead of between *respond* and *withdrew*. On the other hand, dependency structure is very effective in identifying cases where one event is the consequence or the cause of multiple events, as in *"The president offered to* $\underline{offset}_{\mathrm{T}}$ *Jordan's costs* ***because*** *40% of its exports* $\underline{go}_{\mathrm{S}}$ *to Iraq and 90% of its oil* $\underline{comes}_{\mathrm{S}}$ *from there."* In this case, the algorithm annotates a causal link between *go* and *offset*, and also between *comes* and *offset*.

The annotation of CLINKs in sentences not containing causal relations is strongly related to the ambiguous nature of many verbs, prepositions and conjunctions, which encode a causal meaning or express a causal relation only in some specific contexts. For instance, many mistakes are due to the erroneous disambiguation of the conjunction *since*. According to the addDiscourse tool, *since* is a causal connector in around one third of the cases, as in *"For now, though, that would be a theoretical advantage* ***since*** *the authorities have admitted they have no idea where Kopp is."* However, there are many cases where the outcome of the tool is not perfect, as in *"****Since*** *then, 427 fugitives have been taken into custody or located, 133 of them as a result of citizen assistance, the FBI said"*, where *since* acts as a temporal conjunction.

5.2 Quantitative Evaluation

In order to perform also a quantitative evaluation of our automatic annotation, we manually annotated 100 documents taken from the TimeBank corpus according to the annotation guidelines discussed before. We then used this data set as a gold standard.

The agreement reached by two annotators on a subset of 5 documents is 0.844 Dice's coefficient on C-SIGNALS (micro-average over markables) and of 0.73 on CLINKS.

We found that there are several cases where the algorithm failed to recognize causal links due to events that were originally not annotated in TimeBank. Therefore, as we proceed with the manual annotation, we also annotated missing events that are involved in causal relations. Table 8 shows that, in creating the gold standard, we annotated 61 new events. As a result, we have around 52% increase in the number of CLINKs. Nevertheless, explicit causal relations between events are by far less frequent than temporal ones, with an average of 1.4 relations per document.

If we compare the coverage of automatic annotation with the gold standard data (without newly added events, to be fair), we observe that automatic annotation covers around 76% of C-SIGNALs and only around 55% of CLINKs. This is due to the limitation of the algorithm that only considers a

Annotation	EVENT	C-SIGNAL	CLINK
manual	3933	78	144
manual-w/o new events	3872	78	95
automatic	3872	59	52

Table 8: Statistics of causality annotation in manual versus automatic annotation

	precision	recall	F1-score
C-SIGNAL	0.64	0.49	0.55
CLINK	0.42	0.23	0.30

Table 9: Automatic annotation performance

small list of causal connectors. Some examples of manually annotated causal signals that are not in the list used by the algorithm include *due mostly to*, *thanks in part to* and *in punishment for*.

Finally, we evaluate the performance of the algorithm for automatic annotation (shown in Table 9) by computing precision, recall and F1 on gold standard data without newly added events. We observe that our rule-based approach is too rigid to capture the causal information present in the data. In particular, it suffers from low recall as regards CLINKs. We believe that this issue may be alleviated by adopting a supervised approach, where the list of verbs and causal signals would be included in a larger feature set, considering among others the events' position, their PoS tags, the dependency path between the two events, etc.

6 Conclusions

In this paper, we presented our guidelines for annotating causality between events. We further tried to automatically identify in TempEval-3 corpus the types of causal relations described in the guidelines by implementing some simple rules based on causal cues and dependency structures.

In a manual revision of the annotated causal links, we observe that the algorithm obtains a precision of 0.61, with some issues related to the class of PREVENT verbs. Some mistakes are introduced by the tools used for parsing and for disambiguating causal signals, which in turn impact on our annotation algorithm. Another issue, more related to recall, is that in the TBAQ-cleaned corpus not all events are annotated, because it focuses originally on events involved in temporal relations. Therefore, the number of causal relations identified automatically would be higher if we did not take into account this constraint.

From the statistics presented in Section 4.1, we can observe that widely used verbs such as *have* or *keep* express causality relations only in few cases. The same holds for affect verbs, which are never found in the corpus with a causal meaning, and for link verbs. This shows that the main sense of causal verbs usually reported in the literature is usually the non-causal one.

Recognizing CLINKs based on causal signals is more straightforward, probably because very frequent ones such as *because of* and *as a result* are not ambiguous. Others, such as *by*, can be identified based on specific syntactic constructions.

As for the polarity of CLINKs, which is a very important feature to discriminate between actual and negated causal relations, this phenomenon is not very frequent (only 2 cases) and can be easily identified through dependency relations.

We chose to automatically annotate TBAQ-cleaned corpus because one of our goals was to investigate how TLINKs and CLINKs interact. However, this preliminary study shows that there are only few overlaps between the two relations, again with C-SIGNALs being more informative than causal verbs. This may be biased by the fact that, according to our annotation guidelines, only explicit causal relations are annotated. Introducing also the implicit cases would probably increase the overlap between TLINKs and CLINKs, because annotator would be allowed to capture the temporal constrains existing in causal relations even if the are not overtly expressed.

In the near future, we will complete the manual annotation of TempEval-3 corpus with causal information in order to have enough data for training a supervised system, in which we will incorporate the lessons learnt with this first analysis. We will also investigate the integration of the proposed guidelines into the Grounded Annotation Format (Fokkens et al., 2013), a formal framework for capturing semantic information related to events and participants at a conceptual level.

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, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines, December. http://www.ldc.upenn.edu/Catalog/docs/LDC2011T03/propbank/english-propbank.pdf.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

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*, pages 11–20, Atlanta, Georgia, June. Association for Computational Linguistics.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic, June. Association for Computational Linguistics.

Cécile Grivaz. 2010. Human Judgements on Causation in French Texts. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Li-szu Agnes Huang. 2012. The Effectiveness of a Corpus-based Instruction in Deepening EFL Learners' Knowledge of Periphrastic Causatives. *TESOL Journal*, 6:83–108.

ISO TimeML Working Group. 2008. ISO TC37 draft international standard DIS 24617-1, August 14. http://semantic-annotation.uvt.nl/ISO-TimeML-08-13-2008-vankiyong.pdf.

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *In Proceedings of 38th Annual Meeting of the ACL, Hong Kong, 2000*, pages 336–343.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky, J. Castano, R. Ingria, Roser Saurí, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics*.

James Pustejovsky, Jessica Littman, Roser Saurí, and Marc Verhagen. 2006. Timebank 1.2 documentation. Technical report, Brandeis University, April.

Hiroki Sakaji, Satoshi Sekine, and Shigeru Masuyama. 2008. Extracting causal knowledge using clue phrases and syntactic patterns. In *Proceedings of the 7th International Conference on Practical Aspects of Knowledge Management*, PAKM '08, pages 111–122, Berlin, Heidelberg. Springer-Verlag.

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100.

The PDTB Research Group. 2008. The PDTB 2.0. Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

Sara Tonelli, Rachele Sprugnoli, and Manuela Speranza. 2014. NewsReader Guidelines for Annotation at Document Level, Extension of Deliverable D3.1. Technical Report NWR-2014-2, Fondazione Bruno Kessler. https://docs.google.com/viewer?url=http%3A%2F%2Fwww.newsreader-project.eu%2Ffiles%2F2013%2F01%2FNWR-2014-2.pdf.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating events, time expressions, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

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 Saurí, 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.

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

Phillip Wolff, Bianca Klettke, Tatyana Ventura, and Grace Song. 2005. Expressing causation in english and other languages. *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin*, pages 29–48.

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