An Analysis of Causality between Events and its Relation to Temporal Information

Paramita Mirza
Fondazione Bruno Kessler,
University of Trento
Trento, Italy
paramita@fbk.eu

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

Abstract

In this work we present an annotation framework to capture causality between events, inspired by TimeML, and a language resource covering both temporal and causal relations. This data set is then used to build an automatic extraction system for causal signals and causal links between given event pairs. The evaluation and analysis of the system's performance provides an insight into explicit causality in text and the connection between temporal and causal relations.

1 Introduction

Causality is a concept that has been widely investigated from a philosophical, psychological and logical point of view, but how to model its recognition and representation in NLP-centered applications is still an open issue. However, information on causality could be beneficial to a number of natural language processing tasks such as question answering, text summarization, decision support, etc. The lack of information extraction systems focused on causality may depend also on the lack of unified annotation guidelines and standard benchmarks, which usually foster the comparison of different systems performances. Specific phenomena related to causality, such as causal arguments (Bonial et al., 2010), causal discourse relations (The PDTB Research Group, 2008) or causal relations between nominals (Girju et al., 2007), have been investigated, but no unified framework has been proposed to capture causal relations between events, as opposed to the existing TimeML standard for temporal relations (Pustejovsky et al., 2010).

The work presented in this paper copes with this issue by *i)* proposing an annotation framework to model causal relations between events and *ii)* detailing the development and the evaluation of a supervised system based on such framework.

We take advantage of the formalization work carried out for the TimeML standard, in which events, temporal relations and temporal signals have been carefully defined and annotated. We propose to model causal relations in a similar way to temporal relations, inheriting from TimeML the notion of event, relation and signal, even though our approach to causality is well rooted in the *force dynamic* model by Talmy (1985).

Besides, we focus our preliminary annotation on TimeBank (Pustejovsky et al., 2006), a corpus widely used by the research community working on temporal processing. This should possibly enable the adaptation of existing temporal processing systems to the analysis of causal information, given that we rely on well-known standards and data. On the other hand, this makes it easier for us to straightforwardly investigate the relation between temporal and causal information, given that a causing event should always take place *before* a resulting event.

2 Related Work

Research on the extraction of event relations has concerned both the analysis of the temporal ordering of events and the recognition of causality relations. However, the two research lines have progressed quite independently from each other. Recent works on temporal relations mostly revolve around the last

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2097–2106, Dublin, Ireland, August 23-29 2014.

TempEval-3[1] shared task on temporal and event processing. The task organizers released some data sets annotated with events, time expressions and temporal relations in TimeML format (Pustejovsky et al., 2003), mainly used for training and evaluation purposes. The results of TempEval-3 reported by UzZaman et al. (2013) show that, even though the performance of systems for extracting TimeML events and time expressions is quite good (>80% F-score), the overall performance of end-to-end event extraction pipelines is negatively affected by the poor performance of modules for temporal relation extraction. In fact, the state-of-the-art performance on the temporal relation extraction task yields only around 36% F-score (Bethard, 2013).

The problem of detecting causality between events is as challenging as recognizing their temporal order, but less analyzed from an NLP perspective. Besides, it has mostly focused on specific types of event pairs and causal expressions in text, and has failed to provide a global account of causal phenomena that can be captured with NLP techniques. SemEval-2007 Task 4 "Classification of Semantic Relations between Nominals" (Girju et al., 2007) gives access to a corpus containing nominal causal relations among others, as causality is one of the considered semantic relations in the task. Bethard et al. (2008) collected 1,000 conjoined event pairs connected by *and* from the Wall Street Journal corpus. The event pairs were annotated manually with both temporal (BEFORE, AFTER, NO-REL) and causal relations (CAUSE, NO-REL). They use 697 event pairs to train a classification model for causal relations, and use the rest for evaluating the system, which results in 37.4% F-score. Rink et al. (2010) perform textual graph classification using the same corpus, and make use of manually annotated temporal relation types as a feature to build a classification model for causal relations between events. This results in 57.9% F-score, 15% improvement in performance compared with the system without the additional feature of temporal relations.

The interaction between temporal and causal information, and the contribution of temporal information to the identification of causal links, are also one of the issues investigated in this paper. However, we aim at providing a more comprehensive account of how causal relations can be explicitly expressed in a text, and we do not limit our analysis to specific connectives.

Do et al. (2011) developed an evaluation corpus by collecting 20 news articles from CNN, allowing the detection of causality between *verb-verb*, *verb-noun*, and *noun-noun* triggered event pairs. Causality between event pairs is measured by taking into account Point-wise Mutual Information (PMI) between the cause and the effect. They also incorporate discourse information, specifically the connective types extracted from the Penn Discourse TreeBank (PDTB), and achieve a performance of 46.9% F-score. Unfortunately, the data set is not freely available, hence, comparing our work with theirs is not possible.

The most recent work of Riaz and Girju (2013) focuses on the identification of causal relations between verbal events. They rely on the unambiguous discourse markers *because* and *but* to automatically collect training instances of cause and non-cause event pairs, respectively. The result is a knowledge base of causal associations of verbs, which contains three classes of verb pairs: *strongly causal*, *ambiguous* and *strongly non-causal*.

The lack of a standard benchmark to evaluate systems for the extraction of causal relations between events makes it difficult to compare the performance of different systems, and to identify the state-of-the-art approach to this particular task. For this reason, we annotated TimeBank, a freely available corpus, with the aim of making it available to the research community for further evaluations.

3 Data annotation

In order to develop a classifier for the detection of causal relations between events, we first define annotation guidelines for explicit causality and then manually annotate a data set for training and testing.

3.1 Annotation scheme

Since one of the goals of this work is to investigate the interaction between temporal and causal information, we define an annotation scheme strongly inspired by the TimeML standard for events, time expressions and temporal relations. First, we inherit from TimeML the definition of events, which includes all types

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

of actions (punctual and durative) and states. Hence, we do not limit our annotation only to specific PoS such as verbal or nominal events.

Similar to the <TLINK> tag in TimeML for temporal relations, we introduce the <CLINK> tag to mark a causal relation between two events. Both TLINKs and CLINKs mark directional relations, i.e. they involve a source and a target event. However, while a list of relation types is part of the attributes for TLINKs (e.g. BEFORE, AFTER, INCLUDES, etc.), for CLINKs only one relation type is foreseen, going from a *source* (the cause, indicated with $_{S}$ in the examples) to a *target* (the effect, indicated with $_{T}$).

We also introduce the notion of causal signals through the <C-SIGNAL> tag. <SIGNAL>s have been introduced in TimeML to annotate temporal prepositions and other temporal connectives and subordinators. If a SIGNAL marks the presence of a temporal relation in a text, its ID is added to the attributes of such TLINK. In a similar way, C-SIGNALs are used to mark-up textual elements signalling the presence of causal relations, which include all causal uses of *prepositions* (e.g. because of, as a result of, due to), *conjunctions* (e.g. because, since, so that), *adverbial connectors* (e.g. so, therefore, thus) and *clause-integrated expressions* (e.g. the reason why, the result is, that is why). Also for CLINKs it is possible to assign a *c-signalID* attribute, in case a C-SIGNAL marks the causal relation between two events in text.

Concerning the notion of causality, it is particularly challenging to provide guidelines that clearly define how to identify it in text, since causality exists as a psychological tool for understanding the world independently of language and it is not necessarily grounded in text (van de Koot and Neeleman, 2012). There have been several attempts in the psychology field to model causality, including the counterfactual model (Lewis, 1973), the probabilistic contrast model (Cheng and Novick, 1991; Cheng and Novick, 1992) and the dynamics model (Wolff and Song, 2003; Wolff et al., 2005; Wolff, 2007), which is based on Talmy's force dynamic account of causality (Talmy, 1985; Talmy, 1988). We choose to lean our guidelines on the latter model, since it accounts also for different ways in which causal concepts are lexicalized.

Specifically, Wolff (2007) claims that causation covers three main types of causal concepts, i.e. CAUSE, ENABLE and PREVENT. These causal concepts are lexicalized through three types of verbs listed in Wolff and Song (2003): *i)* CAUSE-type verbs, e.g. *cause, prompt, force*; *ii)* ENABLE-type verbs, e.g. *allow, enable, help*; and *iii)* PREVENT-type verbs, e.g. *block, prevent, restrain*. These categories of causation and the corresponding verbs are taken into account in our guidelines (Tonelli et al., 2014).

We assign a CLINK if, given two annotated events, there is an explicit causal construction linking them. Such construction can be expressed in one of the following ways:

1. Expressions containing **affect verbs** (*affect, influence, determine, change, etc.*), e.g. *Ogun ACN crisis* $_{S}$ ***influences*** *the launch* $_{T}$ *of the All Progressive Congress*.
2. Expressions containing **link verbs** (*link, lead, depend on, etc.*), e.g. *An earthquake* $_{T}$ *in North America was* ***linked*** *to a tsunami* $_{S}$ *in Japan*.
3. **Basic constructions involving causative verbs** of CAUSE, ENABLE and PREVENT type, e.g. *The purchase* $_{S}$ ***caused*** *the creation* $_{T}$ *of the current building*.
4. **Periphrastic constructions involving causative verbs** of CAUSE, ENABLE and PREVENT type, e.g. *The blast* $_{S}$ ***caused*** *the boat to heel* $_{T}$ *violently*. With "periphrastic" we mean constructions where a causative verb (*caused*) takes an embedded clause or predicate as a complement expressing a particular result (*heel*).
5. Expressions containing **CSIGNALs**, e.g. *Its shipments declined* $_{T}$ ***as a result of*** *a reduction* $_{S}$ *in inventories by service centers*.

We annotate both intra- and inter-sentential causal relations between events, provided that one of the above constructions is present. We do not annotate causal relations that are implicit and must be inferred by annotators, because they may be highly ambiguous and would probably affect inter-annotator agreement.

3.2 Corpus statistics

Based on the guidelines above, we manually annotated causality in the TimeBank corpus taken from TempEval-3, containing 183 documents with 6,811 annotated events in total.[2] We chose this corpus because gold events were already present, between which we could add causal links. Besides, one of our research goals is the analysis of the interaction between temporal and causal information, and TimeBank already presents full manual annotation of temporal information according to TimeML standard.

However, during annotation, we noticed that some events involved in causal relations were not annotated, probably because the corpus was originally built focusing on events involved in temporal relations. Therefore, we annotated also 137 new events, which led to around 56% increase in the number of annotated CLINKs.

The total number of annotated CSIGNALs is 171 and there are 318 CLINKs, much less than the number of TLINKs found in the corpus, which is 5,118. Besides, not all documents contain causality relations between events. From the total number of documents in TimeBank, only 109 (around 60%) of them contain explicit causal links and only 87 (around 47%) of them contain CSIGNALs. We also found that there is no temporal signal (marked by <SIGNAL> tag) annotated in TimeBank, which is unfortunate since it could help in disambiguating causal signals from temporal signals.

Annotation was performed using the CAT tool (Bartalesi Lenzi et al., 2012), a web-based application with a plugin to import annotated data in TimeML and add information on top of it. 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. The built corpus is then used as training and test data in the experiments for the classification of CSIGNALs and CLINKs, as described in Section 4. This preliminary analysis on the corpus, however, shows that explicit causal relations between events are less frequently found in texts than temporal ones. This may lead to data sparseness problems.

4 Experiments

Using the 183 documents from TimeBank manually enriched with causal information for training and testing, we implement two different classifiers: the first one is a CSIGNAL labeler, that takes in input information on events and temporal expressions as annotated in the original TimeBank, and classifies whether a token is part of a causal signal or not (Section 4.1). The second one is a CLINK classifier, which given an event pair detects whether they are connected by an explicit causal link (Section 4.2). Both experiments are carried out based on five-fold cross-validation. The overall approach is largely inspired by our existing framework for the classification of temporal relations (Mirza and Tonelli, 2014).

4.1 Automatic Extraction of CSIGNALs

The task of recognizing CSIGNALs can be seen as a text chunking task, i.e. using a classifier to determine whether a token is part of a causal signal or not. Since the extent of causal signals can be expressed by multi-word expressions, we employ the IOB tagging convention to annotate the data, where each token can either be classified into B-CSIGNAL, I-CSIGNAL or O (for other). We build our classification model using the Support Vector Machine (SVM) implementation provided by YamCha[3], a generic, customizable, and open source text chunker. In order to provide the classifier a feature vector to learn from, we perform the two following steps:

1. Run the *TextPro* tool (Pianta et al., 2008) to get information on base NP chunking and whether a token is part of named entity or not.
2. Run *Stanford CoreNLP* tool[4] to get information on lemma, part-of-speech (PoS) tags and dependency relations between tokens.

In the end, the feature vector includes *token*, *lemma*, *PoS tag*, *NP chunking*, *dependency path*, and *several binary features*, indicating whether a token is: *i)* an event or part of a temporal expression,

[2]The annotated data set is available at `http://hlt.fbk.eu/technologies/causal-timebank`

[3]`http://chasen.org/~taku/software/yamcha/`

[4]`http://nlp.stanford.edu/software/corenlp.shtml`

according to gold TimeML annotation; *ii)* part of a named entity or not; and *iii)* part of a specific discourse connective type.

Dependency information is encoded as the dependency path between the current token and its *governor*. For example, in *"He fell because the chair was broken"*, there is a dependency relation *mark (broken, because)*, where *mark* indicates the presence of a finite clause subordinate to another clause (de Marneffe and Manning, 2008). Thus, we encode the dependency feature for the token *because* as *mark (broken)*. If the governor is an event, e.g. *broken* is annotated as an event, the dependency feature is represented as *mark (*EVENT*)* instead.

The mentioned binary features are introduced to exclude the corresponding token as a candidate token for a causal signal. In other words, if a token is part of a named entity or an event, it is very unlikely that it will be part of a causal signal. The same holds for all connective types that do not express causal relations, e.g. temporal or concessive ones. In order to obtain this information, we include in the feature vector the information about discourse connectives acquired using the *addDiscourse* tool (Pitler and Nenkova, 2009), which identifies connectives and assigns them to one of four semantic classes in the framework of the Penn Discourse Treebank (The PDTB Research Group, 2008): TEMPORAL, EXPANSION, CONTINGENCY and COMPARISON. Note that causality is part of the CONTINGENCY class.

System	Precision	Recall	F-score
Rule-based (baseline)	54.33%	40.35%	46.31%
Supervised chunking	91.03%	41.76%	57.26%

Table 1: Evaluation of CSIGNAL extraction system

Table 1 shows the performance of our classification model in a five-fold cross-validation setting, which yields a good precision but a poorer recall, summing up into 57.26% F-score. We also compare our supervised model with a baseline rule-based system, which labels as CSIGNALs all causal connectors listed in our annotation guidelines and those appearing in specific syntactic constructions. For instance, *from* and *by* are always labeled as CSIGNAL when they are governed by a passive verb annotated as event and govern another event, as in the sentence *"The building was* damaged $_{\text{T}}$ ***by*** *the* earthquake $_{\text{S}}$*."* Note that this is quite a strong baseline, since the rule-based algorithm embeds some of the intuitions on syntactic dependencies expressed also as features in the supervised approach.

4.2 Automatic Extraction of CLINKs

Similar to causal signal extraction, we approach the problem of detecting causal links between events as a supervised classification task. Given an ordered pair or events (e_1,e_2), the classifier has to decide whether there is a causal relation between them or not. However, since we also consider the directionality of the causal link, an event pair (e_1,e_2) is classified into 3 classes: CLINK (where e_1 is the source and e_2 is the target), CLINK-R (with the reverse order or source and target) or NO-REL. Again, we use YamCha to build the classifier. This time, a feature vector is built for each pair of events and not for each token as in the previous classification task.

As candidate event pairs, we take into account every possible combination of events in a sentence in a forward manner. For example, if we have e_1, e_2 and e_3 in a sentence (in this order), 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. This is necessary to account for inter-sentential causality, under the simplifying assumption that causality may occur only between events in two consecutive sentences.

We implement a number of features, some of which are computed independently based on either e_1 or e_2, e.g. lemma, PoS, while some others are pairwise features, which are computed based on both elements, e.g. dependency path, signals in between, etc. The implemented features are as follows:

String and grammatical features. The tokens and lemmas of e_1 and e_2, along with their PoS and a binary feature indicating whether e_1 and e_2 have the same PoS tags.

Textual context. The sentence distance and event distance of e_1 and e_2. Sentence distance measures

how far e_1 and e_2 are from each other in terms of sentences, i.e. 0 if they are in the same sentence. The event distance corresponds to the number of events occurring between e_1 and e_2 (i.e. if they are adjacent, the distance is 0).

Event attributes. Event attributes as specified in TimeML annotation, which consist of *class*, *tense*, *aspect* and *polarity*. Events being a noun, adjective and preposition do not have tense and aspect attributes in TimeML. Therefore, we retrieve this information by extracting the tense and aspect of the verbs that govern them, based on their dependency relation. We also include four binary features representing whether e_1 and e_2 have the same event attributes or not. These features, especially the *tense* and *aspect* one, are very relevant for detecting causality. For instance, if e_1 is in the future tense and e_2 in the past tense, there cannot be a causal relation connecting e_1 (as source) and e_2 (as target or result).

Dependency information. We include as features *i)* the dependency path that exists between e_1 and e_2, *ii)* the type of causative verb connecting them (if any) and *iii)* binary features indicating whether e_1/e_2 is the *root* of the sentence. This information is based on the collapsed representation of dependency relations provided by the parsing module of Stanford CoreNLP. Consider the sentence *"Profit from coal fell* $_{\text{T}}$ *to $41 million from $58 million, partly* ***because of*** *a miners' strike* $_{\text{S}}$*."* Based on the collapsed typed dependencies, we would obtain a direct relation between *fell* and *strike*, which is *prep_because_of (fell, strike)*. This information combined with the classification of *because of* as a causal signal would straightforwardly identify the relation connecting the two events as causal.

Causal signals. We take into account the annotated CSIGNALs connecting two candidate events. We look for causal signals occurring between e_1 and e_2, or before e_1. We also include the position of the signals (*between* or *before*) as feature, since it is crucial to determine the direction of the causality of a given ordered event pair. This is particularly evident if you consider the position of causal signals in the following examples: *i)* "The building collapsed $_{\text{T}}$ **because of** the earthquake $_{\text{S}}$" vs. *ii)* "**Because of** the earthquake $_{\text{S}}$ the building collapsed $_{\text{T}}$." This feature is also very relevant in connection with the *Textual context*, since two events being in two different sentences are linked by an explicit causal relation only in specific cases, for instance if there is a CSIGNAL in between, typically at the beginning of the second sentence. Note that in case of several CSIGNALs occurring between e_1 and e_2, we take the closest CSIGNAL to e_2, as in the sentence "The building was damaged $_{\text{S}}$ **by** the earthquake , **thus**, people moved $_{\text{T}}$ away". The dependency path between the causal signal and e_1/e_2 is also important to determine the correct involved events in the causal relations. For instance, in the sentence "They decided $_{\text{T}}$ to move **because of** the earthquake $_{\text{S}}$", the involved event is *decided* instead of *move*.

Temporal relations (TLINKs). Rink et al. (2010) showed that including temporal relation information in detecting causal links results in improving classification performance. Nevertheless, they only analyze this phenomenon when causality is expressed by the conjunction *and*. We decided to include this information in the feature set by specifying the temporal relation type connecting e_1 and e_2, if any, to see whether TLINKs help in improving causality detection also in a more comprehensive setting.

We evaluate our approach in a five-fold cross-validation setting, and we compare the performance of our classifier with a baseline rule-based system. This relies on an algorithm that, given a term t belonging to *affect*, *link*, *causative* verbs (basic and periphrastic constructions) or *causal signals* (as listed in the annotation guidelines), looks for specific dependency constructions where t is connected to two events. If such dependencies are found, a CLINK is automatically set between the two events identifying the source and the target of the relation. Further details on the baseline system and its evaluation can be found in Mirza et al. (2014).

In our experimental setting, we evaluate two versions of the CLINK classifier: the first includes as features the *gold annotated* CSIGNALs in the classification model, while the second takes in input the CSIGNALs *automatically annotated* by the classifier described in Section 4.1. We also evaluate the contribution of *dependency*, *CSIGNAL* and *TLINK* features by excluding each of them from the classification model.

Evaluation results are reported in Table 2. We observe that the baseline is always outperformed by the other classifiers. CSIGNAL is the most important feature, with a particularly high impact on recall. The

intuition behind this result is that, if a CSIGNAL is present, it is a strong indicator of a causal relation being present in the surrounding context. This is similar to what Derczynski and Gaizauskas (2012) report for temporal information, showing that temporal signals provide useful information in TLINK classification. Dependency information contributes to the performance of the classifier, but is less relevant than TLINK information. A more detailed analysis of the relation between temporal and causal information is reported in the following section. The significantly decreasing recall of the classifier using the automatic extracted CSIGNALs as features is most probably caused by the low recall of the CSIGNAL extraction system.

System	Precision	Recall	F-score
Rule-based (baseline)	36.79%	12.26%	18.40%
Supervised classification (with gold CSIGNALs)	74.67%	35.22%	47.86%
- without dependency feature	65.77%	30.82%	41.97%
- without CSIGNAL feature	57.53%	13.21%	21.48%
- without TLINK feature	61.59%	29.25%	39.66%
Supervised classification (with automatic CSIGNALs)	67.29%	22.64%	33.88%

Table 2: Performance of CLINK extraction system

5 Discussion

We further analyse the output of the automatic extraction systems, in order to understand some phenomena triggering the results.

5.1 Recognizing CSIGNALs

When we manually inspect the output of the CSIGNAL extraction system, we find that the false positives are actually the causal signals that annotators missed in the corpus, and not ambiguous connectives. The system surprisingly yields better precision than human annotation, finding new correct signals.

The recall, however, suffers most probably from data sparseness. It is possible that during the cross-validation experiments some splits do not have enough data to learn from, recalling that only around 47% of the documents contain annotated CSIGNALs. Furthermore, 20% of the false negative cases are due to classifier's mistakes in detecting the causal signal *by*, which is highly ambiguous. Our assumption with the rule-based system that "*by* is likely to be a causal signal when it is used to modify a passive verb" is too restrictive, since *by* can convey a causal meaning even if the target event is not in the passive voice, as in the example *"The embargo is meant to cripple* $_{\text{T}}$ *Iraq* ***by*** *cutting* $_{\text{S}}$ *off its exports of oil and imports of food and military supplies."*

Another ambiguous causal signal that the classifier fails to detect is the conjunction *and*. We believe that more training data, and perhaps more lexical information on the tokens connected by the conjunction *and*, are needed for the classifier to be able to disambiguate them.

5.2 Detecting CLINKs

We found that most of the mistakes done by the classifier, as well as by the rule-based system, are caused by the dependency parser output that tends to establish a dependency relation between a causative verb or causal signal and the closest verb. For example, in the sentence *"StatesWest Airlines withdrew* $_{\text{T}}$ *its offer to acquire Mesa Airlines* ***because*** *the Farmington carrier did not respond* $_{\text{S}}$ *to its offer"*, the dependency parser identify *because* as the mark of *acquire* instead of *withdrew*.

Moreover, also for this task data sparseness is definitely an issue. One possible solution would be to annotate more data, for instance the AQUAINT data set used for TempEval-3 competition (UzZaman et al., 2013). Another possibility would be to automatically generate additional data from the Penn Discourse TreeBank corpus, where causality is one of the discourse relations annotated between argument pairs. However, a further processing step would be needed to identify inside the argument spans the events between which a relation holds, which may introduce some errors.

Regarding the directionality of causal relations, the classifier is generally quite precise. 112 out of 150 CLINKs detected by the classifier actually match a causal relation present in the gold annotated data. Only 8 of them have been classified with the wrong direction. We believe that using the TLINK types as features contributes to this good performance in disambiguating causality direction (CLINK vs. CLINK-R).

5.3 Interaction between temporal and causal information

We provide in Table 3 some statistics on the overlaps between causal links and temporal relation types from the gold data. The *Others* class in the table includes SIMULTANEOUS, IS_INCLUDED, BEGUN_BY and DURING_INV relations. These counts were obtained by overlapping the temporal information in TimeBank with the causal information manually added for our experiments. In total, only 32% of the gold causal links have the underlying temporal relations. Note that the annotators could not see the temporal links already present in the data, therefore they were not biased by TLINKs when assessing causal links.

	BEFORE	AFTER	IBEFORE	IAFTER	Others	Total
CLINK	15	5	0	0	4	24
CLINK-R	1	67	0	3	8	79

Table 3: Statistics of CLINKs overlapping with TLINKs

The data confirm our intuition that temporal information is a strong constraint when detecting causal relations, with the BEFORE class having the most overlaps with CLINK and AFTER with CLINK-R. This is in line with the outcome of our feature analysis reported in Table 2, suggesting that feeding temporal information into a causal relation classifier yields an improvement in performance. However, the converse would be less effective, since the occurrences of explicit causal relations are by far less frequent than temporal ones. Besides, we found that the few cases where CLINKs overlap with AFTER relation are not due to annotation mistakes, as in the example *"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."*

Finally, the performance achieved by our system in causal relation extraction (with gold C-SIGNALs) is 47.86% F-score, which is better than the performance of the state-of-the-art temporal relation extraction system with 36.26% (Bethard, 2013). This probably depends on the fact that extracting CLINKs is a simpler task compared with TLINK extraction: in the first case 3 classes are considered, while temporal relation types are classified into 14 classes.

6 Conclusions

In this paper, we presented a framework for annotating causal signals and causal relations between events. Besides, we implemented and evaluated two supervised systems, one classifying C-SIGNALs and the other CLINKs.

With the first task, we showed that while recognizing unambiguous causal signals is very trivial, ambiguous signals such as *by* and *and* are very difficult to identify because they occur in diverse syntactic constructions. We definitely need more data to learn from, and perhaps use more lexical information on the words connected by such causal signals as features. The knowledge base of causal associations between verbs developed by Riaz and Girju (2013) may be a useful resource to provide such information, and we will explore this possibility in the future.

We found that the low recall achieved by the CLINK classifier is probably affected by wrong dependencies identified by the Stanford parser. In the future, we would like to test also the C&C tool (Curran et al., 2007) to extract dependency relations, since it has a better coverage of long-range dependencies. We have also shown that causal signals are very important in detecting explicit causal links holding between two events. Finally, we showed that temporal relation types help in disambiguating the direction of causality, i.e. to determine the source and target event. However, the converse may not hold, since the causal links in the data set are very sparse, and only 2% of the total TLINKs overlap with CLINKs.

Acknowledgements

The research leading to this paper was partially supported by the European Union's 7th Framework Programme via the NewsReader Project (ICT-316404). We also thank Rachele Sprugnoli and Manuela Speranza for their contribution in defining the annotation guidelines.

References

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

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.

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.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. PropBank Annotation Guidelines, Version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

Patricia W. Cheng and Laura R. Novick. 1991. Causes versus enabling conditions. *Cognition*, 40(1-2):83 – 120.

Patricia W. Cheng and Laura R. Novick. 1992. Covariation in natural causal induction. *Psychological Review*, 99(2):365–382.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically Motivated Large-Scale NLP with C&C and Boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 33–36, Prague, Czech Republic, June. Association for Computational Linguistics.

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.

Leon Derczynski and Robert J. Gaizauskas. 2012. Using Signals to Improve Automatic Classification of Temporal Relations. *CoRR*, abs/1203.5055.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Stroudsburg, PA, USA. 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.

David Lewis. 1973. Causation. *The Journal of Philosophy*, 70(17):pp. 556–567.

Paramita Mirza and Sara Tonelli. 2014. 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 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association.

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, 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)*.

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

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings o the Fifth International Workshop on Interoperable Semantic Annotation*.

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 *Proceedings of the Twenty-Third International FLAIRS Conference*.

Leonard Talmy. 1985. Force dynamics in language and thought. *Chicago Linguistic Society*, 21:293–337.

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. Technical Report NWR-2014-2, Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2013/01/NWR-2014-2.pdf.

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 *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

H. van de Koot and A. Neeleman, 2012. *The Theta System: Argument Structure at the Interface*, chapter The Linguistic Expression of Causation, pages 20 – 51. Oxford University Press: Oxford.

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.