Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation

Tommaso Caselli
RUG / Groningen, NL
t.caselli@rug.nl

Oana Inel
VU / Amsterdam, NL
oana.inel@vu.nl

Abstract

This paper describes a crowdsourcing experiment on the annotation of plot-like structures in English news articles. The CrowdTruth methodology and metrics have been applied to select valid annotations from the crowd. We further run an in-depth analysis of the annotated data by comparing it with available expert data. Our results show a valuable use of crowdsourcing annotations for such complex semantic tasks, and promote a new annotation approach that combines crowd and experts.

1 Introduction

Causal relations are a pervasive phenomenon in human activities, including narrative production. Causality is actually the main component of narratives, regardless of the mediums (novels, news articles, comments, micro-blogs, pictures, among others) and their fictional status (fictional vs. non-fictional narratives). In a narrative text, causal connections between events allow the story to progress, the actors to participate, and eventually reach a conclusion. Causality is responsible for logically connecting the events together in a meaningful way.

If we shift perspective, and look at narratives from the point of view of the producers rather than their structural properties, it is easy to observe how humans impose causal, or explanatory, relations among events that they perceive or are involved into. Humans have a great appetite for information and are in constant need to find explanations for the things they observe. We search the present for cues and evidence, merge and resolve information with what we already known (i.e., the past), and use this information to (try to) predict the future and make decisions. Explanatory relations and narrative strategies are one of the major cognitive tools we use to observe the world and, most importantly, to interpret it (Boyd, 2009; Gottschall, 2012). When reporting on an event in the world, or telling someone a personal experience, we do not merely describe what happens, i.e., we do not just list events in the order of occurrence [1], but we connect them in a set of coherent patterns, or, in other words, we give rise to *plot structures* (Bal, 1997). Plot structures express a form of reasoning about causal relations between events and states composing the narrative (Lehnert, 1981; Goyal et al., 2010; Mani, 2012).

The current stream of data and information is growing everyday and its size and complexity is such that humans may suffer from "information overload". To minimise such a problem, intelligent content management systems have been developed and they became more and more popular and used. Different methods and approaches have been developed to provide users with personalised and relevant information. However, most of this information is given in the form of full text documents that require the users to read them to identify (i.e., extract) the information. Automatic processing would be beneficial, especially if the results are presented as structured data based on narrative strategies. We follow, in this

[1] For a comparison, think about the Ancient Roman tradition of the *Annales*, concise historical records merely reporting events chronologically.

Proceedings of the Workshop on Events and Stories in the News, pages 44–54
Santa Fe, New Mexico, USA, August 20, 2018.

respect, the proposal of automatically generating storylines of events (Vossen et al., 2015; Caselli and Vossen, 2016).

This paper reports on a crowdsourcing experiment on the annotation of causal relations between pairs of events in news data.

The main contributions of this work are:

- an analysis of the crowdsourced data, in terms of parameters that may affect the annotation quality, time, and evaluation of the data using the CrowdTruth methodology (Aroyo and Welty, 2014; Aroyo and Welty, 2015);
- a comparison between experts and crowdsourced annotated data with respect to a publicly available reference benchmark corpus for storyline evaluation, the Event StoryLine Corpus (ESC) (Caselli and Vossen, 2017);
- the release of an enhanced version of the Event Storyline Corpus (ESC v1.2).

The remainder of the paper is organised as follows: Section 2 provides an overview of related work on the annotation of causal relations in different datasets, highlighting differences and commonalities with our contribution. Section 3 describes the dataset and the crowdsourcing experiment settings based on the CrowdTruth metrics. Section 4 reports on an in-depth analysis of the crowd data and its comparison with existing expert annotated data. Finally, Section 5 summarises our findings and suggests directions for future work.

2 Related Work

Causality can be broadly defined as the knowledge, or way of knowing, if an event, or a state of affairs, is responsible for *causing* another one. To avoid the intrinsic circularity of this definition, we can rephrase it in more generic terms such that causality establishes a connection between two processes, events, or states, whereby the first is (partly) responsible for the occurrence, or holding as true, of the second, and the second is (partly) dependent on the first.

Causality has been subject of debates in different scientific communities. One of the most relevant aspect of this debate is the lack of a homogeneous theory of causality, and, most importantly, the availability of a plurality of perspectives on it. Providing an extensive and critical summary of this debate is out of the scope of this work, but, we will review relevant works in the areas of Linguistics and Natural Language Processing that contributed to shape this notion, its annotation in actual natural language data, and the development of automatic systems. We restrict this literature review to approaches in the news domain.

One of the distinguishing properties of causality in natural language, shared with other semantic relations such as meronymy and mereology, is granularity (Hobbs, 1985; Mulkar-Mehta et al., 2011). This allows humans to interactively play between coarse-grained and fine-grained levels of causality. Further studies (Talmy, 1976; Comrie, 1981; Girju and Moldovan, 2002) have investigated the variety of lexico and semantic constructions that can express causation in a natural language. At least for English, as well as other Indo-European languages, it is possible to differentiate the set of causative constructions into two big groups: i.) those expressing causality via **explicit** patterns; and ii.) those using **implicit** patterns. The difference between these two ways of expressing causality relies in the semantic transparency of the causative constructions. Explicit causative constructions are characterised by the presence of keywords such as causal connectives, adverbs or adjectives (e.g. *because (of)*, *with the results that*, *since*, *so*), causation verbs (e.g. *cause*, *bring about*, *kill*, *blacken*), and conditional constructions, among others. On the other hand, implicit causation can be expressed by complex nominals (e.g. *malaria*$_{NP1}$ *mosquitoes*$_{NP2}$, where NP2 is interpreted as causing NP1 [2]), verbs of implicit causation, and discourse structure.

The annotation of expressions of causality, or causal language, has received lots of attention which resulted in the realisation of different annotation schemes and initiatives. Computational lexicons, such

[2]This example is extracted from (Girju and Moldovan, 2002).

as WordNet (Miller, 1995), VerbNet (Schuler, 2005), PropBank (Kingsbury and Palmer, 2002), and FrameNet (Baker et al., 1998), were the first to encode this information at the level of lexical items or senses. For instance, WordNet encodes two relations, such as *causes* and *entailments*. VerbNet and PropBank include causative verbs. FrameNet represents causality through a variety of frames (e.g. CAUSATION, THWARTING) and roles (e.g. PURPOSE).

The Penn Discourse Treebank (PDTB) (Prasad et al., 2007) models causality as inference of discourse relations. Causality is a subclass of the contingency relation hierarchy, together with enablement and condition. The definition of causality we have used in the crowdsourcing experiments is strictly connected to that of contingency of the PDTB. However, we annotate relations between pairs of events rather than between discourse units.

Other initiatives concern three different annotation projects: CaTeRS (Mostafazadeh et al., 2016b), CATENA (Mirza and Tonelli, 2016), and BeCauSE 2.0 (Dunietz et al., 2017). The first two projects, based on the TimeML annotation scheme (Pustejovsky et al., 2003), annotate causality between pairs of events. CaTeRs adopts a commonsense reasoning perspective, rather than limiting the annotation to the presence of specific linguistic markers. The scheme adopts three values (`cause`, `enable`, and `prevent` (Wolff, 2007)) to be annotated as "true" with respect to the actual context of occurrence of the event pairs. The authors report a global Fleiss's κ score on all annotated relations (including also temporal relations) of $\kappa = 0.49$ without closure, and $\kappa = 0.51$ with closure. CATENA adopts a linguistic approach. The annotation of a causal relation is allowed only between pairs of events *in presence of* a non-discontinuous causal connective, i.e., limited to explicit relations. Finally, BeCauSE still addresses the annotation in terms of a linguistic approach, requiring the presence of a causal connective for the annotation to take place. The main difference with respect to CATENA and other initiatives concerns the fact that it annotates all constructions that express causality rather than restricting to a particular realisation (e.g. discourse relations, or TimeML events). The approach we have adopted in our crowdsourcing experiments follows CaTeRs as we have adopted a commonsense reasoning perspective. However, we have simplified the granularity of the values to one type only, cause, finding the three-way classification too fine grained for the crowd.

Other works have addressed causality in the broader context of automatically learning narrative structures, or plot-like structures, using unsupervised methods. A notable work in this area is the Narrative Event Chains (Chambers and Jurafsky, 2008). Narrative chains are partially ordered sequences of events related to a common protagonist, i.e., sequences of verbs sharing a common actor, identified through typed dependencies, obtained from a large corpus collection. Narrative chains do not model causality directly, but they assume that narratives, such as news articles, are coherent structures. This means that if a sequence of verbs shares a coreferring argument, then these verbs must be connected by the discourse structure. One of the main criticism of this approach is that the chains express more co-occurrence relations rather than actual narrative relations, and, in some cases, may result in non-coherent chains of events.

Crowdsourcing of causal relations has received less attention than other natural language processing tasks, such as event extraction and factuality assessment (Lee et al., 2015), temporal information extraction (Caselli et al., 2016; Snow et al., 2008), word sense disambiguation (Jurgens, 2013; Akkaya et al., 2010), among others. To study narratives, (Hu and Broniatowski, 2017) proposed a crowdsourcing approach to represent a text, which is split into smaller text snippets, as a causal network. The crowd workers were asked to draw links between text snippets that are related through a causal relation, in an external tool. In a similar way, creative writing crowdsourcing tasks have been developed (Mostafazadeh et al., 2016a) to build a corpus of commonsense stories containing causal and temporal relations between everyday events. Other initiatives have annotated causal relations between propositions (Sukhareva et al., 2016), among other context-sensitive semantic verb relations, i.e., co-reference, temporal, entailment. The crowd workers had an observed agreement of 71.8%, where Krippendorff's α was equal to 0.32 on a very limited set (i.e., there were only between 2%-6% of causal relations in the entire dataset). In this work, we specifically focus on identifying loose causal relations between events in a large variety of topics using simplified crowdsourcing instructions.

3 Crowdsourcing Causal Relations between Events

As already stated, we follow a commonsense reasoning approach to annotate causality. Furthermore, our goal is to approximate the annotation of plot-like structures rather than strict causal relations between pairs of linguistic items. In the rest of the paper, we use causal relations and plot-like relations as synonym terms. Thus, causality is naively used to refer to the broader notion of contingent relations. This choice is also dictated by a desire to be as much ecological as possible with respect to the crowd in the process of data collection. In our vision, ecology is declined in two ways: i.) avoid to bias the crowd with lengthy and complex task instructions (including examples); ii.) collect a diversity of judgements assuming multi-faceted versions of ground truth data, i.e., there is no such a thing as absolute right or wrong, but varieties of truths. In the remainder of this section, we describe the dataset (Section 3.1), the crowdsourcing annotation template (Section 3.2), the quality metrics used to evaluate the crowdsourced data (Section 3.3) and the crowdsourcing experiments performed (Section 3.4). The data and the crowd annotations are publicly available. [3]

3.1 Dataset

The experimental dataset covers 22 topics from the Event StoryLine Corpus v1.0 (ESC v1.0) (Caselli and Vossen, 2017). The ESC corpus contains expert annotations that cover a high range of entities and relations such as: actors, locations, temporal expressions, events, temporal relations, event coreference relations, and plot-like relations between pairs of events. The plot relations in the ESC data are marked with a <PLOT_LINK> tag, and broadly correspond to contingent relations between pairs of events. The annotation of these links is based on relatively simple annotation guidelines, instructing the annotators in the identification of the eligible pairs of events and associated relation (i.e., relation directionality). The inter-annotator agreement for <PLOT_LINK> has been calculated using the Dice coefficient and equals 0.638.

ESC consists of 22 topics, for a total of 281 news articles. We extracted 1,204 annotated sentences containing at least two expert annotated events. Following the approach of the ESC corpus, we have excluded events belonging to the following classes from the event pairs: ASPECTUAL, REPORTING, CAUSATIVE, and GENERIC. These classes actually represent sets of event mentions which cannot give rise to a plot-like structure, or a contingent relation. For instance, on the one hand, in the case of a REPORTING event (e.g. *say*, *report*), a plot-like relation holds with respect to the actual content of what is "reported" rather than between the marker of the presence of a reporting event. On the other hand, CAUSATIVE events (e.g. *cause*, *sparkle*, *trigger*) have been excluded as they are interpreted as explicit markers of a causal relation. The actual plot relation holds between their arguments. An overview of the dataset is shown in Table 1. The ESC dataset contains 2,290 manually annotated <PLOT_LINK> relations between event pairs. This set of relations is then expanded to 5,684 pairs when using coreference relations. As for the manually annotated pairs, only 1,571 out of 2,290 (68.6%) occur in the same sentence. In the 1,204 sentences that we selected in our experiments, there are only 1,540 expert annotated event pairs, that are further used in our analysis (Section 4).

Table 1: Dataset Overview

#Topics	#Doc	#Sent	#Event Pairs	# Expert Annotation ESC v1.0	# Expert Pairs ESC v1.0 in Our Experiments
22	281	1,204	7,778	1,571	1,540

3.2 Crowdsourcing Annotation Template

We ran the crowdsourcing experiments on the Figure Eight[4] platform, formerly known as CrowdFlower. Figure 1 shows the annotation template used to gather crowd annotations on causal relations between

[3] https://github.com/CrowdTruth/Crowdsourcing-StoryLines

[4] https://www.figure-eight.com

event pairs. The annotation template uses simplified instructions that can all be seen in Figure 1, i.e., we did not provide detailed instructions or annotation guidelines, nor examples. In short, the workers were given a sentence and a list of expressions to validate. An expression consists of one of the statements $Event_A$ `causes` $Event_B$ or $Event_A$ `is caused by` $Event_B$, where $Event_A$ (E_A) and $Event_B$ (E_B) appear in the sentence. For example, the sentence shown in Figure 1 contains three events: *warns*, *bombs* and *war*. Taking all possible combinations of these three events, the crowd is asked to validate the following expressions: *warns* `causes` *bombs*, *warns* `is caused by` *bombs*, *warns* `causes` *war*, *warns* `is caused by` *war*, *bombs* `causes` *war*, *bombs* `is caused by` *war*. To help workers identify the position of the two events composing each expression in the sentence, the events are highlighted in the sentence when hovering over the given expression. For instance, in Figure 1 we hover over the expression *warns* `is caused by` *bombs* (grey background), and therefore, the events *warns* and *bombs* are highlighted in blue in the sentence. The workers were allowed to choose as many expressions as they considered valid. In case no valid expression was found for the given sentence, the crowd workers were asked to motivate their answer in a text field.

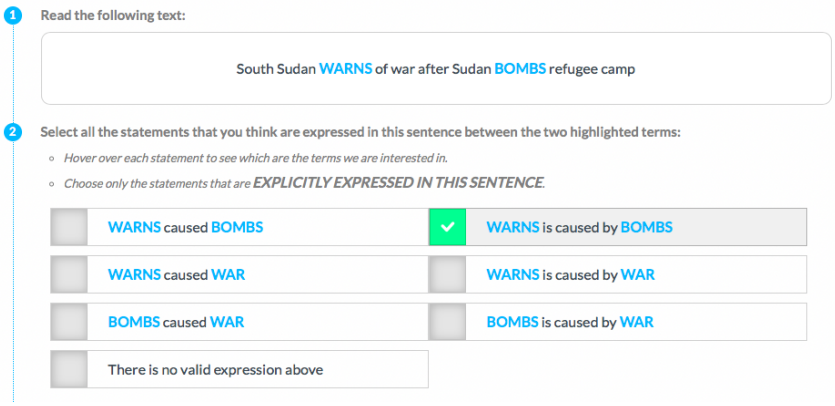

Figure 1: Screenshot of the Crowdsourcing Template to Annotate Causal Relations between Events.

3.3 Crowdsourcing Quality Metrics

The task of extracting causal relations between events is prone to disagreement, diverse perspectives, and interpretations due to: i.) the inherent ambiguity of natural language; and ii.) the difficult nature of dealing with events and causality. To address and consider these aspects, we chose to evaluate the quality of the crowdsourced data by using the assumption behind the CrowdTruth disagreement-aware methodology (Aroyo and Welty, 2014; Aroyo and Welty, 2015): ambiguity is reflected in all crowdsourcing components (i.e., units, workers, annotations) and the ambiguity of each component influences the other components. For our usecase, a unit represents a sentence, the workers are the contributors from the Figure Eight platform, and the annotations are statements of type E_A `causes/is caused by` E_B, where E_A and E_B appear in the sentence, and the value "*NONE*", from which the workers can choose, as described in Section 3.2. A worker judgement is composed of such validated statements.

In this work, we followed and applied the CrowdTruth methodology and metrics as suggested in (Dumitrache et al., 2018). For our use case, the identification of causal relations between events, each worker's judgement is translated into a binary worker vector, $WorkerVec$, which has a length equal to n+1, where n is the total number of causal relation statements to choose from and the last component refers to the value "*NONE*". Each causal relation component that was picked by the worker gets a value of 1, and 0 otherwise. The $WorkerVec$ of all workers that annotated the same sentence s are summed up to compute the sentence vector, $SentVec$. These two vectors are then used to compute the quality score for each sentence, worker and causal relation, in particular:

- *unit quality score* (UQS): represents the degree of agreement among the workers that annotated the sentence s, *i.e.*, the lower the score, the less clear the sentence. UQS is computed as the average cosine similarity between all $WorkerVec$ for s, weighted by the worker quality (WQS).

- *worker quality* (WQS): represents the degree of a worker's agreement with the rest of the workers on the specific task. WQS of worker i is computed as the product of 2 cosine similarity metrics - the worker-worker agreement WWA (a pair-wise agreement between every two workers) and the worker-sentence agreement WSA (the agreement of a worker with all the workers that annotated the same sentence); the two worker metrics are weighted by the unit quality score UQS; thus, the annotations of the workers with lower quality score will weight less in the final output.
- *sentence - causal relation score* ($SCausalRel$): represents the likelihood of the causal relation r to be expressed in sentence s. $SCausalRel$ is computed as the ratio of the number of workers that picked the causal relation r over all workers that annotated the sentence, weighted by WQS.

Using these preliminaries, the CrowdTruth metrics model the inter-dependency between the three main components of the crowdsourcing experiments - units (sentences), workers and causal relation statements. The aforementioned quality metrics are computed in a dynamic fashion, iteratively, until the results are stable. As a result of this process, the final crowd annotations, the $SCausalRel$, are weighted by the quality of the workers that annotated the given unit. The reason for choosing the CrowdTruth approach to weight the annotations of the workers rather than those provided by the platform (in our case Figure Eight) is that the trust values of the crowdsourcing platform does not account for the ambiguity of the data that is annotated.

3.4 Crowdsourcing Data Collection

In total, we ran two crowdsourcing experiments, as show in Table 2 - a pilot experiment $TrialEventPairs$ on 4 topics and a main experiment $6EventPairs$ on all 22 topics. We ran the pilot experiment, $TrialEventPairs$, to identify the optimal settings in terms of number of event pairs to be shown at the same time to the workers. Figure 2 shows the distribution of UQS for each set of sentences containing between [1, 28] event pairs. Besides the distribution of UQS, the plot also shows the mean UQS value, the median UQS value and the number of sentences containing the given number of event pairs. There is a clear pattern between the increase of event pairs (X axis) and the decrease of the *UQS*. This suggests that the amount of event pairs influences the overall quality of the sentences and consequently, the performance of the workers on identifying causal relations between events. Given that for sentences containing more than 6 event pairs the mean UQS drops below 0.4 in most cases, we identified 6 event pairs as the optimal number. Therefore, in the main experiment ($6EventPairs$), the crowd needs to validate a maximum of 12 causal relation statements (2 for each pair of events).

Each unit, which is composed of a sentence and a set of causal relation statements, was annotated by 15 workers and each annotation was paid 2¢. The workers were categorized as level 2 according to Figure Eight, i.e., a smaller group of more experienced, higher accuracy contributors. For the $TrialEventPairs$ experiment we gathered 3,360 annotations from a total of 157 unique workers and for the $6EventPairs$ experiment we gathered 27,675 annotations from a total of 697 unique workers. We split our input units in batches of around 50 units, i.e., we were publishing jobs of around 50 units at a time. In each job, the workers were allowed to annotate as many units as they wanted, with a maximum limit of 20 units per job. In total, in the $TrailEventPairs$ experiment the workers annotated between 1 and 75 units, with an average of 21 units per worker and in the $6EventPairs$ experiments, the workers annotated between 1 and 457 units, with an average of 40 units per worker. The total cost of the two experiments was 756$.

4 A Comparison with Experts

We ran a set of comparative analyses between the data collected through this crowdsourcing experiment and the annotations of <PLOT_LINK> in the ESC v1.0 dataset. Given that the CrowdTruth metrics allow us to estimate the quality of the annotated data, expressed by the *SCausalRel* score, we can use the different thresholds as corresponding to different qualities of the crowd annotated data. The usefulness of a comparison with expert data is in this case two-folded: i.) it provides additional evaluation of the crowd data which complements the CrowdTruth measures; ii.) it allows us to gain more insights on the

Table 2: Overview of crowdsourcing experiments to derive optimal annotation settings and template

Type	Exp.	Input Data				Crowdsourcing Template	
		#Topics	#Sent.	#Units	#Event Pairs	Annotations	Max # of Annotations
Pilot	*TrialEventPairs*	4	217	224	1,477	E_A causes E_B E_A caused by E_B NONE	57
Main	*6EventPairs*	22	1,204	1,845	7,778	E_A causes E_B E_A caused by E_B NONE	13

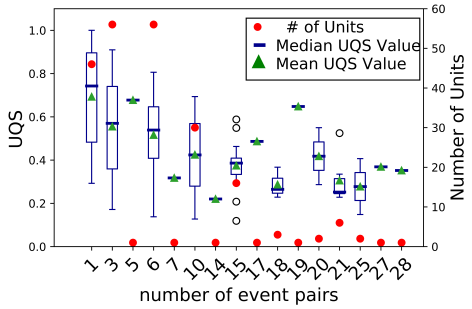

Figure 2: Distribution of UQS for any number of event pairs in the $TrialEventPairs$ pilot experiment.

differences between experts and crowd (in annotation behaviour), and to identify a reliability threshold for directly using the crowdsourced data as Gold Standard, or for integrating them with expert data.

The analysis was conducted as follows: first, we excluded all units that were marked as "*NONE*", regardless of the *SCausalRel* score. This allows us to access a large set of events pairs. In case there is actually no relation among the pairs, the *SCausalRel* score will be either very low or equal to zero, if no worker has annotated it. The *SCausalRel* score ranges between 0 and 1, where 1 expresses perfect agreement among all crowd annotators. After this, we have generated different thresholds, starting from 1 and lowering the score by a 0.1 point at a time, up to 0.5, a value signalling a 50% agreement among all workers. In this way, we can compare crowd data of different agreement with the expert data. We have used standard Precision (P), Recall (R), and F1-score (F1), assuming the expert data as a Gold Standard.

Table 3: Comparing Experts and Crowd: Causal Relation Identification

Threshold	P	R	F1	# Crowd Relations
1.0	0.923	0.007	0.015	13
0.9	0.764	0.086	0.155	174
0.8	0.670	0.191	0.297	440
0.7	0.547	0.298	0.386	838
0.6	0.424	0.424	0.424	1,540
0.5	0.316	0.546	0.401	2,654

Table 3 illustrates the results of the overall evaluation, i.e., the ability of the crowd to identify both the event pairs that stand in a causal relation and the directionality of the causal relation. As the figures show, there is a clear pattern: the lower the threshold, the higher the number of relations annotated by the crowd. Lower thresholds actually correspond to higher disagreement among the workers, pointing out differences in the interpretation of the sentences, as well as signalling the complexity of the task. In this case, the differences may concern the actual pair, the relation directionality, or both. We can also observe that lower thresholds correspond to an improvement of Recall (i.e. higher matching with experts), at the

Table 4: Comparing Experts and Crowd: Event Pairs Detection (only)

Threshold	P	R	F1	# Crowd Pairs	# FPs	# Unique FPs	%Correct FP
1.0	0.923	0.007	0.015	13	1	1	100%
0.9	0.787	0.088	0.159	174	37	36	77.77%
0.8	0.695	0.198	0.309	440	134	97	82.75%
0.7	0.586	0.314	0.409	827	342	208	63.38 %
0.6	0.480	0.453	0.466	1,456	757	415	56.41%
0.5	0.390	0.589	0.469	2,328	1,420	663	49.65%

cost of Precision. However, this level of analysis is too coarse grained. For instance, at a 0.6 *SCausalRel* score threshold, Precision and Recall are the same, and both of them are below 50%. On the one hand, this signals that the diversity (of the crowd) is valuable in identifying more relations than the experts. On the other hand, it does not tell us much about the quality of the data. We basically know that 60% of the times, the workers agree on the presence of a relation, and that they have identified much more relations than the experts. Although this is in line with our annotation approach based on commonsense reasoning (people, with diverse personal experiences, identify a larger set of likely relations), we do not know if the extra relations with respect to the experts are valid or not.

We thus conducted two additional analyses on the crowd data by inspecting, separately, the event pairs alone, and then, the relation directionality. This provides a better assessment of the quality of the crowd data as well as which sub-task is harder: the event pair identification or the relation directionality.

Table 4 reports on the results for the event pairs identification subtask. The mismatch between the number of crowd relations in Table 3 and that of the crowd pairs in Table 4 is due to the fact that in some cases both directionality values (i.e. `causes` and `is caused by`) have the same *SCausalRel*, or the *SCausalRel* is in the same threshold range, thus increasing the number of relations, especially for lower thresholds. The values for P, R and F1 are in line with those of the global evaluation (see Table 3). In this case, we have extended the analysis by manually inspecting 20% of the False Positives for each threshold, with the exclusion of threshold 1.0. The analysis shows that, until a threshold of 0.6, the majority of False Positives are actually valid pairs that were missed by the experts. As lower thresholds subsume all pairs from higher ones, the manual validation of the False Positives shows that it is possible to identify an optimal threshold for the crowd data, that in this case corresponds to 0.7, where 63.38% of the event pairs are actually valid. We have also analysed the non-valid cases. We have identified two reasons for the errors: i.) either the event pair is not valid in the actual context of occurrence (see example 1); or ii.) the event pair is genuinely wrong (see example 2).

1. *A powerful earthquake [. . .],* ***killing*** *at least five people and injuring dozens in a region devastated by the* ***quake****-triggered tsunami of 2004.* [ESC v1.0, 37_1, sentence 3]

2. *During the escape, Arcade Joseph Comeaux , Jr . [. . .] took them hostage and forced them to* ***drive*** *to Baytown, Texas, where he* ***restrained*** *the officers in the back of the van [. . .].* [ESC v1.0, 3_4, sentence 3]

In example 1, the event **quake** took place in 2004, a different (and distant) time period with respect to the actual **killing** in the sentence. Interestingly, we observe that such context dependent errors compose the majority of invalid False Positive up to 0.7. At 0.6 and 0.5, we have observed an increase of errors (or better disagreements) like example 2 where, rather than a misinterpretation of the context, it is the presence of the causal/explanatory relation itself that is in doubt or not valid. In this latter case, if we use a commonsense-based trigger question like "*why were the officers* ***restrained****?*", it is very unlikely to answer "*Because the escapee drove them.* A more suitable answer would be "*Because the escapee took the officers hostage.*

Finally, concerning the directionality of the relations, we measured the observed agreement of the pairs that both the experts and the crowd have annotated, using the same thresholds. Agreement ranges

between 0.973 for threshold 1.0 and to 0.909 for threshold 0.5. At 0.7, we observed a score of 0.945. The trend is somehow parallel to the pairs detection, although these values signal an almost perfect "agreement" with the experts.

5 Conclusion and Future Work

This paper has reported on a crowdsourcing task for identifying causal relations between pairs of events. We adopted a loose definition of causality, that is best represented by contingent relations. By means of a pilot experiment we could identify the best amount of events to present to the workers in order to obtain as much as possible reliable annotations. We used the CrowdTruth metrics both to evaluate the quality of the annotated data and to weight the quality of the workers based on their overall agreement with the rest of the workers. This has allowed us to access diverse annotations, using "disagreement" as an extra source of information rather than to decide what is right or wrong. Finally, we have converted the crowd data in the same format of ESC v1.0 and generated flexible Gold Standard data, either by merging the crowd data per threshold to the experts or by using only the crowd data. We call this new resource ESC v1.2 and make it publicly available.

Natural languages have an extremely varied set of devices (i.e. granularity) to express relations among concepts, also for causal/contingent relations. Such relations are in most cases not explicitly marked in the sentence/text. As a further insight from the analysis of the crowd-expert pairs only (i.e. Table 4), we can observe that the causal relation task has different levels of complexity for the crowd. In particular, it appears that the identification of valid pairs of events is a harder task than the identification of the relation directionality.

The combined comparison with expert data has helped us to gain more insights on the differences in annotations between these two approaches. There is a general tendency for crowd workers to provide more valid annotations than experts, confirming previous studies (Caselli et al., 2016). At the same time, we can exploit the *SCausalRel* score to identify reliability thresholds of the annotated data. The differences in quality should not be considered as errors but rather as proxies for the complexity of the task and of the actual data in analysis. This calls for the development of new annotation procedures. We should reconsider using experts to generate annotations from scratch, and thus risking of making the generation of new datasets an infeasible task due to money, time, and effort. On the other hand, we should embrace the ability and diversity of the crowd to perform complex semantic tasks and promote a new allegiance between crowd and experts. As our results have shown, even at a threshold of 0.5, there is still a lot of valid information (in our case 49.65%) that should not be discarded and this is when we should employ experts. As lower thresholds signal also more complex data, experts should be employed in revising these data. This will result in richer, better, and possibly less biased datasets to be used as benchmarks for NLP systems.

As future work, we are planning to extend the ESC corpus with newly annotated data by applying the "crowd-experts-in-the-loop" approach in two directions. The first aims at collecting more data, and therefore, to allow the development or adaptation of NLP systems for storyline extraction. The second goal aims at extending the annotations in languages other than English, thus giving rise to a multilingual version of the ESC dataset.

References

Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 195–203. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Journal of Human Computation*, 1:31–34.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the*

17th international conference on Computational linguistics-Volume 1, pages 86–90. Association for Computational Linguistics.

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

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

Tommaso Caselli and Piek Vossen. 2016. The storyline annotation and representation scheme (star): A proposal. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 67–72.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada, August. Association for Computational Linguistics.

Tommaso Caselli, Rachele Sprugnoli, Oana Inel, et al. 2016. Temporal information annotation: Crowd vs. experts. In *LREC*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. *Proceedings of ACL-08: HLT*, pages 789–797.

Bernard Comrie. 1981. Causative constructions in language universals and linguistic typology.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. *arXiv preprint arXiv:1805.00270*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Roxana Girju and Dan Moldovan. 2002. Mining answers for causation questions. In *Proc. The AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 67–82.

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

Amit Goyal, Ellen Riloff, and Hal Daume III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA, October. Association for Computational Linguistics.

JR Hobbs. 1985. Granularity (pp. 432–435). In *9th International Joint Conference on Artificial Intelligence, Los Angeles*, pages 18–23.

Dian Hu and David A Broniatowski. 2017. Measuring perceived causal relationships between narrative events with a crowdsourcing application on mturk. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 349–355. Springer.

David Jurgens. 2013. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648.

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

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

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

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75.

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

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Rutu Mulkar-Mehta, Jerry Hobbs, and Eduard Hovy. 2011. Granularity in natural language discourse. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 360–364. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. 2016. Crowdsourcing a large dataset of domain-specific context-sensitive semantic verb relations. In *LREC*.

Leonard Talmy. 1976. Semantic causative types. *The grammar of causative constructions*, pages 43–116.

Piek Vossen, Tommaso Caselli, and Yiota Kontzopoulou. 2015. Storylines for structuring massive streams of news. In *Proceedings of the First Workshop on Computing News Storylines*, pages 40–49.

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