Validation Methodology for Expert-Annotated Datasets: Event Annotation Case Study

Oana Inel[1]
Delft University of Technology, The Netherlands
Vrije Universiteit Amsterdam, The Netherlands
o.inel@tudelft.nl, oana.inel@vu.nl

Lora Aroyo
Google Research, New York, US
loraa@google.com

Abstract

Event detection is still a difficult task due to the complexity and the ambiguity of such entities. On the one hand, we observe a low inter-annotator agreement among experts when annotating events, disregarding the multitude of existing annotation guidelines and their numerous revisions. On the other hand, event extraction systems have a lower measured performance in terms of F1-score compared to other types of entities such as people or locations. In this paper we study the consistency and completeness of expert-annotated datasets for events and time expressions. We propose a data-agnostic validation methodology of such datasets in terms of consistency and completeness. Furthermore, we combine the power of crowds and machines to correct and extend expert-annotated datasets of events. We show the benefit of using crowd-annotated events to train and evaluate a state-of-the-art event extraction system. Our results show that the crowd-annotated events increase the performance of the system by at least 5.3%.

2012 ACM Subject Classification Information systems → Crowdsourcing; Human-centered computing → Empirical studies in HCI; Computing methodologies → Machine learning

Keywords and phrases Crowdsourcing, Human-in-the-Loop, Event Extraction, Time Extraction

Digital Object Identifier 10.4230/OASIcs.LDK.2019.12

1 Introduction

Natural language processing (NLP) tasks span a large variety of applications [14], such as event extraction, temporal expressions extraction, named entity recognition, among others. While the performance of named entity recognition tools is constantly improving, the event extraction performance is still poor. On the one hand, events are vague and can have multiple perspectives, interpretations and granularities [16]. On the other hand, there is hardly a single, standardized way to represent events. Instead, we find a plethora of annotation guidelines, standards and datasets created, adapted and extended by human experts [33]. Although the annotation guidelines are aimed to ease the annotation task, the inter-annotator agreement values reported are still low, ranging between 0.78 and 0.87 [7, 33]. Current research [7, 33, 15] acknowledges the fact that expert-annotated datasets could be inconsistently annotated or could contain ambiguous labels, but there is no standardized way of measuring if they indeed contain inconsistent or incomplete annotations.

In the natural language processing field, crowdsourcing is extensively used as a mean of gathering fast and reliable annotations [29]. Although, typically, crowd annotations are evaluated against experts annotations by means of majority vote approaches, more recent

[1] Corresponding author

2nd Conference on Language, Data and Knowledge (LDK 2019).
Editors: Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski; Article No. 12; pp. 12:1–12:15
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

approaches focus on capturing the *inter-annotator disagreement* [1] and the creation of ambiguity-aware crowd-annotated datasets [12].

In this paper we present a data-agnostic validation methodology for expert annotated datasets. We investigate the degree of consistency and completeness of expert-annotated datasets and we propose an ambiguity-aware crowdsourcing approach to validate, correct and improve them. We apply this methodology on the expert annotated datasets of events and time expressions, namely TempEval-3 Gold (Gold) and TempEval-3 Platinum (Platinum), which were used in the TempEval-3 Time Annotation[2] task at SemEval 2013. To show the added value of employing crowd workers for providing event annotations, we use the crowd-annotated events to train and evaluate a state-of-the-art event extraction system which participated in the challenge. Therefore, we investigate the following research questions:

RQ1: *How reliable are expert-annotated datasets in terms of consistency and completeness?*

RQ2: *Can we improve the reliability of expert-annotated datasets in terms of consistency and completeness through crowdsourcing?*

To answer these research questions we make the following contributions:

- data-agnostic validation methodology of expert-annotated datasets in terms of consistency and completeness;
- 4,202 crowd-annotated English sentences from the TempEval-3 Gold and TempEval-3 Platinum datasets with events and 121 crowd-annotated sentences from the TempEval-3 Platinum dataset with time expressions;
- training and evaluation of a state-of-the-art system for event extraction with ambiguity-aware crowd-driven event annotations.

We make available the crowdsourcing annotation templates for all experiments, the scripts used for our validation methodology and the crowdsourcing results in the project repository[3].

The remainder of this paper is structured as follows. Section 2 reviews related work in the field of event extraction by focusing on automatic, crowdsourcing and human-in-the-loop approaches. Section 3 describes the dataset and Section 4 introduces our data-agnostic validation methodology. Section 5 presents the results of our data-agnostic validation methodology for measuring the consistency and completeness of expert-annotated datasets. Section 6 presents and discusses the results of our crowdsourcing experiments and the learning outcomes. Finally, Section 7 draws conclusions and introduces future work.

2 Related Work

We review related work on event and time expression detection in three main areas: automatic approaches (Section 2.1), crowdsourcing approaches (Section 2.2) and hybrid, human-in-the-loop approaches (Section 2.3). We focus on the identification of linguistic mentions of type event and time expression, as opposed to identifying named entities of type event and time.

2.1 Automatic Approaches

We review event and time expression detection systems that use domain-agnostic expert-annotated datasets for training and evaluation, such as datasets following the TimeML [26] specifications. This category includes the TempEval-3 dataset, that we use in the current research. We only focus on the detection of events and time expressions, without looking into event classification, time expression normalization or the relations between the two.

[2] https://www.cs.york.ac.uk/semeval-2013/task1/index.html

[3] https://github.com/CrowdTruth/Event-Extraction

For event extraction the majority of the participating systems in the TempEval-3 Time Annotation task used a supervised, knowledge-driven approach with various types of classifiers such as Conditional Random Fields (JUCSE) [20], Maximum Entropy (ATT and NavyTime) [18, 9] and Logistic Regression (ClearTK and KUL) [2, 19] and features such as morphological, semantic, lexical, among others. The TIPSem system [23], the best performing system in the previous challenge from the same series, outperformed all the participants with an F1-score of 82.89 compared to 81.05 of the ATT1 [18] system on identifying the event mention. To the best of our knowledge, the TIPSem [23] system and the CRF4TimeML [6] system (F1-score of 81.87) are currently the best performing systems trained on TimeML datasets.

For temporal expression extraction the best performance in terms of F1-score was 90.32, exhibited by both the NavyTime [9] and SUTime [10] systems. However, they both used a rule-based approach without actually making use of the training data. The next best performing systems on temporal expression extraction, with F1-scores above 0.90, were HeidelTime [31] and ClearTK [2], both using only expert-annotated data as training.

All the aforementioned systems have been evaluated on the TempEval-3 Platinum dataset, an expert-annotated corpus [32]. Although potential ambiguity and errors have been identified in this dataset in previous research [6, 33], the dataset has not been revised. As opposed to this approach, we also evaluate the performance of the ClearTK [2] system with ambiguity-aware crowd-driven event mentions.

2.2 Crowdsourcing Approaches

Crowdsourcing proved to be a reliable approach to gather large amounts of labeled data for many natural language processing tasks such as temporal event ordering [29], causal relation identification between events [5], event factuality [21], event validity [8], among others. As researched [1] showed, disagreement in crowdsourced annotations can be an indication of ambiguity, ambiguous classes of polysemy for event nominals were identified in [30] and ambiguous frames in [12]. In [7], the authors present a crowdsourcing approach for identifying events and time expressions in English and Italian sentences by asking the crowd to highlight phrases in the sentence that refer to events or time. A different approach was taken in [21], where the crowd had to validate one event, at a time, in a sentence. In all the aforementioned approaches, the annotations of the crowd were evaluated against expert annotations.

In this research we combine and extend the approaches proposed in [7] and [21] by asking the crowd to validate in each sentence a set of potential events and time expressions and highlight the missing ones. Moreover, before running the main crowdsourcing study, we run extensive small scale pilot experiments to identify the optimal crowdsourcing settings. Since events and, in a smaller proportion, time expressions are highly ambiguous mentions, we follow and apply the CrowdTruth disagreement-aware methodology [1], similarly to [12], to aggregate and evaluate the crowd annotations. These annotations are then evaluated against expert and also machine annotations. Furthermore, we use the crowd-annotated events as both training and evaluation data for a state-of-the-art event extraction system from the TempEval-3 challenge, namely ClearTK [2].

2.3 Hybrid and Human-in-the-loop Approaches

In NLP, hybrid human-machine approaches have been mainly envisioned on named entity extraction and typing [15] and named entity extraction and linking [11]. The human-machine hybrid NER system published in [3] focused on decomposing individual examples into either examples that can be labelled by automatic tools or by the crowd. Hybrid approaches for event and temporal expression extraction also focused on combining various machine learning approaches with human rules [25]. Although active learning approaches have been used for building event or temporal expression extraction systems [4, 22], the labels are still gathered by means of expert annotators instead of crowdsourcing. In [21], however, the authors use the crowd labels for training a supervised event extraction system.

Current hybrid approaches for event extraction focus on a predefined set of event types, while our approach is suitable for general events. Similarly to [21], we use the crowd-labelled events to train an existing state-of-the-art system for event extraction on the TempEval-3 corpus, but also to evaluate it.

3 Dataset

We focus our analysis on expert-annotated entities of type event and time expression in the TempEval-3 Gold (Gold) and TempEval-3 Platinum (Platinum) datasets from the SemEval 2013 task called TempEval-3 Time Annotation. The Platinum dataset was used to test the performance of the participating systems and the Gold dataset was used for the development of the systems. A detailed description of these two datasets can be found in [27, 28, 32].

We used the TimeML-CAT-Converter[4] and Stanford CoreNLP [24] to split the documents into sentences and tokens and to annotate the tokens with part-of-speech (POS[5]) tags and lemmas. In Table 1 we show the overview of the Gold (G) and Platinum (P) datasets (DS), *i.e.*, the number of documents, sentences, tokens, events and time expressions (times). The Gold dataset contains 256 documents which were split into 3,953 sentences and around 100k tokens and the Platinum dataset contains 20 documents, 273 sentences and around 7k tokens. The Gold dataset contains 3,604 events and 1,450 times, while the Platinum dataset contains 746 events and 138 times, and thus, 3.07 events and 1.27 times per sentence, on average.

Table 1 Overview of TempEval-3 - Gold (G) and TempEval-3 Platinum (P) Datasets (DS).

DS	# Doc	# Sent	# Tokens	#Ann Sent Events	#Ann Sent Times	# Events	# Times	Avg. #Events per Sent	Avg. #Times per Sent
G	256	3,953	≈ 100k	3,604	1,464	11,129	1,822	3.08	1.24
P	20	273	≈ 7k	243	106	746	138	3.06	1.30

Events and Times POS Tags Distribution: Similarly to [33], we looked at the POS tag distribution of events and time expressions in the Gold and Platinum datasets. In both datasets the majority of the events annotated are either verbs or nouns. Adjectives, adverbs and, in a smaller proportion, prepositions are also annotated as events. The Platinum dataset also contains 3 multi-token events composed of numerals. Regarding time expressions, around half of the annotated ones are composed of multiple tokens with various POS tags such as nouns, numbers, preposition, adverbs and adjectives.

Events and Times Tokens and Lemmas: Table 2 shows the number of distinct event and time tokens and lemmas by considering as well their POS tags. On average, in the Gold dataset an event token appears 3.86 times (between 1 and 993 times, *i.e.*, the token "said") while an event lemma appears around 5.94 times (between 1 and 1,154 times, *i.e.*, the lemma "say"). In the Platinum dataset an event token appears on average 1.38 times and an event lemma around 1.69 times. Regarding time expressions, tokens and lemmas appear on average 2.89 times in the Gold dataset and around 1.46 times in the Platinum dataset.

[4] `https://github.com/paramitamirza/TimeML-CAT-Converter`
[5] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

Table 2 Overview of Distinct Event and Time Tokens and Lemmas.

	Events		Times	
DS	Distinct Tokens	Distinct Lemmas	Distinct Tokens	Distinct Lemmas
Gold	2,883	1,871	630	623
Platinum	537	440	94	94

Sentences without Event and Time Annotations: As shown in Table 1, a fraction of the total amount of sentences contained in the two datasets do not contain annotated events, *i.e.*, 349 in Gold and 30 in Platinum, or time expressions, *i.e.*, 2,489 in Gold and 167 in Platinum.

4 Experimental Methodology

In this section we describe our data-agnostic validation methodology of expert-annotated datasets in terms of consistency and completeness. The goal of our experimental methodology is two-fold: *(1)* to measure the reliability of expert-annotated datasets for events and time expressions in terms of consistency and completeness and *(2)* to define an optimal crowdsourcing annotation template to improve the reliability of expert-annotated datasets for events and time expressions in terms of consistency and completeness. The two research questions defined in Section 1 and the following hypotheses guide our experimental methodology:

H1.1 (consistency): Tokens are annotated with different types across datasets.

H1.2 (consistency): Annotation guidelines for events are not used consistently.

H1.3 (completeness): Occurrences of the same previously annotated event tokens or time expression tokens are not annotated by experts.

H1.4 (completeness): Occurrences of the same previously annotated event lemmas or time expression lemmas are not annotated by experts.

H2.1 (reliability): Asking the crowd annotators to motivate their answer increases the reliability of their annotations.

H2.2 (reliability): Gathering event annotations from a large pool of crowd workers provides reliable results in terms of F1-score when compared to expert annotators.

H2.3 (reliability): Crowd-driven event annotations are a reliable way of improving the consistency and completeness of expert-annotated event datasets.

The first step of our methodology, described in Section 4.1, is guided by and extends previously published work on consistency and completeness analysis of expert-annotated datasets of named entities (location, organization, person and role) [15], of events in the TempEval-3 Gold, PropBank/NomBank and FactBank datasets [33] and of events and time expressions in all TempEval-3 datasets [6]. The second step of our methodology adapts the crowdsourcing approach proposed in [15] to improve, complete and correct expert-annotated datasets of events and time expressions. We derive the optimal crowdsourcing annotation template by experimenting with different annotation template independent variables, as described in Section 4.2. Finally, we train and evaluate the ClearTK [2] state-of-the-art event extraction system with crowd-annotated events, as described in Section 4.3.

4.1 Ground Truth Consistency and Completeness

We test hypotheses **H1.1-4** by performing a headroom measurement on the consistency and completeness of expert-annotated entities of type event and time in the TempEval-3 Gold and TempEval-3 Platinum datasets. For consistency (**H1.1-2**) we *(1)* check whether an entity span is annotated with different types across datasets and *(2)* review the experts' adherence to the annotation guidelines. For completeness (**H1.3-4**) we *(1)* verify for each event and time expression token and lemma the proportion in which it was annotated as an event or as a time expression and *(2)* inspect the sentences without annotated events or time expressions to verify whether they might contain missed mentions.

Table 3 Overview of Performed Pilot (P1 to P8) and Main (M1 & M2) Crowdsourcing Experiments.

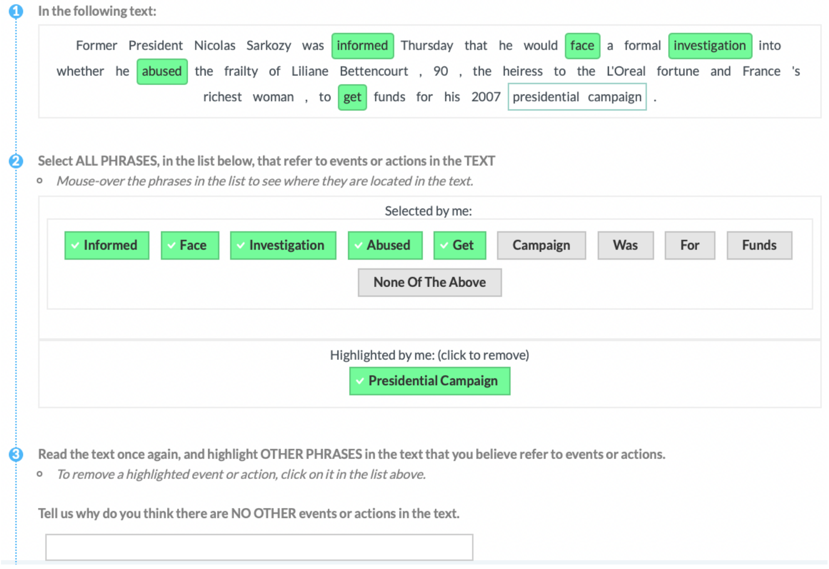
Exp.	Input Data				Crowdsourcing Template	
	#Sent	Entity Type	DS	Entity Values	Annotation Guidelines	Annotation Value
P1	50	Event Time	P	Experts (P) & Tools	Explicit Definition	Entities
P2	50	Event Time	P	Experts (P) & Tools	Explicit Definition	Entities + Motivation (NONE)
P3	50	Event Time	P	Experts (P) & Tools	Explicit Definition	Entities + Motivation (ALL)
P4	50	Event Time	P	Experts (P) & Tools	Implicit Definition	Entities
P5	50	Event Time	P	Experts (P) & Tools	Implicit Definition	Entities + Motivation (NONE)
P6	50	Event Time	P	Experts (P) & Tools	Implicit Definition	Entities + Motivation (ALL)
P7	50	Event Time	P	Experts (G&P) & Tools & Missing	Explicit Definition	Entities + Motivation (ALL)
P8	50	Event Time	P	Experts (G&P) & Tools & Missing	Explicit Definition	Entities + Motivation (ALL) + Highlight
M1	4,202	Event	G&P	Experts (G&P) & Tools & Missing	Explicit Definition	Events + Motivation (ALL) + Highlight
M2	121	Time	G&P	Experts (G&P) & Tools & Missing	Explicit Definition	Times + Motivation (ALL) + Highlight

4.2 Crowdsourcing Experiments

We further test **H1.3-4** through a series of pilot crowdsourcing experiments aiming to improve the ground truth datasets for events and time expressions. We start with a set of 16 *pilot experiments* (eight experiments for event annotation and eight for time expression annotation), P1 to P8 rows as shown in Table 3, in which we experiment with the input data that the crowd is requested to annotate and the design of the crowdsourcing template, similarly to [17]. The role of these pilot experiments is to *obtain the optimal annotation template design*, following **H2.1-2**. We run these experiments on the Figure Eight[6] platform, using level 2 workers from English-speaking countries, *i.e.*, UK, US, CAN and AUS, for each annotation we pay ¢3 (for annotation value without highlight functionality) or ¢4 (for annotation value with highlight functionality) and we ask 20 workers to annotate each sentence.

For each pilot experiment we used 50 sentences from the TempEval-3 Platinum (P) dataset as input data. The crowd needs to validate or add, through highlight, entities of type event or time expression. We vary the list of entities that the crowd needs to validate as follows. In the first six pilot experiments (P1-P6 in Table 3) the crowd was asked to validate only the entities annotated by the experts and returned by the systems participating in the

[6] `https://www.figure-eight.com`



1 In the following text:

Former President Nicolas Sarkozy was informed Thursday that he would face a formal investigation into whether he abused the frailty of Liliane Bettencourt , 90 , the heiress to the L'Oreal fortune and France 's richest woman , to get funds for his 2007 presidential campaign .

2 Select ALL PHRASES, in the list below, that refer to events or actions in the TEXT

- *Mouse-over the phrases in the list to see where they are located in the text.*

Selected by me:

✓ Informed ✓ Face ✓ Investigation ✓ Abused ✓ Get Campaign Was For Funds

None Of The Above

Highlighted by me: (click to remove)

✓ Presidential Campaign

3 Read the text once again, and highlight OTHER PHRASES in the text that you believe refer to events or actions.

- *To remove a highlighted event or action, click on it in the list above.*

Tell us why do you think there are NO OTHER events or actions in the text.

Figure 1 Screenshot of the Main Crowdsourcing Template (M1) to Validate and Highlight Events.

TempEval-3 task. In P7-P8, we expanded the list of entities to be validated with potentially missing entities such as *(1)* annotated entities in the Gold (G) and Platinum (P) datasets and *(2)* any other entity that was annotated in other sentence, but not in the current one.

As part of the crowdsourcing template design we experiment with the annotation guidelines and the annotation values. We request annotators to validate mentions that are both explicit (phrases that refer to events or actions, or temporal expressions) and implicit (phrases that refer to things happening in the past, present, or future, or that involve times, dates, durations, periods, etc.). For the annotation value, we experiment with four options: *(1)* validation of event or time entities, *(2)* validation of those entities with motivation (only when there is no valid entity), *(3)* validation of those entities with motivation (regardless of whether there are valid entities) and *(4)* validation of entities with motivation (regardless of whether there are valid entities) and highlight of potential missed entities.

Main Experiments. We evaluate the outcome of the pilot experiments against the expert annotations to derive the optimal crowdsourcing template in terms of performance (F1-score) to validate, correct and improve datasets for events and time expressions. We run the *main crowdsourcing experiments* on the entire dataset, with the optimal setup. The main crowdsourcing experiments (M1 an M2, the last two rows in Table 3) have the following setup: the input data consists of sentences and events or time expressions annotated by experts, participating systems in the TempEval-3 task and potentially missed events or time expressions; the crowdsourcing template uses explicit definitions and validation of entities with motivation (regardless of whether there are valid entities) and highlight of missed entities. Figure 1 shows the design of the crowdsourcing template for events. We run the *main experiments* on the Figure Eight platform, using level 2 workers from English-speaking countries. Each sentence is annotated by 15 workers and for each annotation we pay ¢4.

4.2.1 Crowd Annotation Aggregation

We aggregate and evaluate the crowd annotations using the CrowdTruth approach for open-ended tasks [13, 12]. First, we define the *worker vector*, *i.e.*, the decision of a worker over an input unit, *i.e.*, a sentence. The worker vector in our case is composed of all entities (either events or time expressions) to be validated or have been highlighted for a given sentence and the value *"none"* (capturing cases when there are no valid entities). Each component in the worker vector gets a value of 1 if the worker selected the entity as valid and 0, otherwise. The sum of all *worker vectors* for a given sentence results in the *sentence vector*. The worker and sentence vectors are then used to compute the following ambiguity-aware metrics:

- *entity-sentence score* ***(ESS)***: expresses the likelihood of each entity e (event of time expression) to be valid for the given sentence s; ESS is computed as the ratio of workers that picked the entity as valid over all the workers that annotated the sentence, weighted by the worker quality; the higher the ESS value, the more clear e is expressed in s;
- *sentence quality score* ***(SQS)***: expresses the workers agreement over one sentence s; SQS is computed as the average cosine similarity of all worker vectors for a sentence s, weighted by the worker quality and entity quality;
- *worker quality score* ***(WQS)***: expresses the overall agreement of one worker with the rest of the workers; WQS is computed using cosine similarity metrics, weighted by the sentence quality and entity quality;
- *entity quality score* ***(EQS)***: being an open-ended task, $EQS = 1$.

These ambiguity-aware metrics are mutually dependent (*i.e.*, they are computed in an iterative dynamic fashion), which means that each aforementioned quality metric depends on the values of the other two metrics. Thus, low quality workers can not decrease the quality of the sentences, and low quality sentences can not decrease the quality of the workers.

4.3 Training & Evaluating the ClearTK Event Extraction System

We used the crowd-annotated events to train and evaluate the ClearTK[7] [2] event extraction system reviewed in Section 2.1, that participated in the TempEval-3 challenge. The selection of the system was made purely based on the availability of the code to easily retrain and evaluate the models. ClearTK [2] uses BIO token chunking for event identification, using the following features: token text, stem, part-of-speech, the syntactic category of the token's parent in the constituency tree, the text of the first sibling of the token in the constituency tree and the preceding and following 3 tokens.

First, after gathering the crowd annotations for both the Gold and Platinum datasets, we apply the aggregation and evaluation metrics presented in Section 4.2.1. Second, we create multiple development (from Gold documents) and evaluation (from Platinum datasets) sets by splitting the crowd-annotated events based on their entity-sentence score, *i.e.*, for every entity-sentence score threshold between 0 and 1, with a step of 0.05. Therefore, we obtain 20 sets of development and evaluation datasets, each containing all the events with a score higher than the respective threshold. Finally, we perform the following four types of experiments to test hypothesis **H2.3**:

- train the system on expert-annotated events and test it on expert-annotated events
- train the system on expert-annotated events and test it on crowd-annotated events
- train the system on crowd-annotated events and test it on expert-annotated events
- train the system on crowd-annotated events and test it on crowd-annotated events

[7] `https://github.com/ClearTK/cleartk`

For all the aforementioned experiments we did not fine-tuned the model's parameters, but we used the ones that performed the best in the TempEval-3 event-extent extraction task.

5 Consistency and Completeness of Expert Annotations

In this section we inspect the consistency and completeness of expert-annotated event and time expression mentions in the TempEval-3 Gold and Platinum datasets, following the hypotheses **H1.1-4**. First, we measure the consistency of the expert-annotated mentions regarding the span of the mentions, the type of the annotated mentions and the adherence to the annotation guidelines in Section 5.1. Second, we measure the completeness of the expert-annotated events and times at the level of part-of-speech distribution and tokens and lemmas and we analyze the sentences without annotated events in Section 5.2.

5.1 Consistency of Expert Annotations

The events annotated by experts in the TempEval-3 Gold (Gold) dataset consist of a single token. Even when the event refers to a multi-token named event, such as "World War II" or "Hurricane Hugo", the experts only mark as event a single token, such as "war" or "hurricane". Interestingly, the TempEval-3 Platinum (Platinum) dataset contains multi-token events composed of numerals, such as "$ 250", "400 million". These events are *not consistent with the latest annotation guidelines* [28] (**H1.2**), since the events of type numeral should be removed. An inconsistency identified in [6] shows that the Platinum dataset contains the noun "season" annotated as event once, while in other sentences from the Gold dataset, it is annotated as a time expression. Furthermore, we observe that the token "tenure" is annotated as an event in the Gold dataset and as a time expression in the Platinum dataset. Therefore, besides a *mention type inconsistency*, we also see an *inconsistency across the training and the evaluation datasets*, proving **H1.1**. Another observation that we made is that overlapping mentions of both type event and time expression are not possible. For example, the word "election" was annotated as event in Platinum dataset, but in the Gold dataset is treated as a time expression, in the word phrase "election day".

5.2 Completeness of Expert Annotations

The completeness analysis follows the setup published in [33]. In the current research, we build on top of this analysis and extend it on a new dataset – TempEval-3 Platinum – and on a new entity type – time expression. Furthermore, we provide entity completeness statistics on the sentences without expert annotated events.

5.2.1 POS Tags Distribution

We analyze the distribution of POS tags (as returned by Stanford CoreNLP) across the events and times annotated by experts in the TempEval-3 Gold and Platinum datasets. For the events annotated by experts in the Platinum dataset, we see consistent observations with the ones published in Table 3 in [33]. Overall, in both datasets *verbs have the highest coverage as events* (63.29% in Gold and 54.43% in Platinum). However, there is still a *significant number of verbs that were not annotated as events*, such as the verbs "participate" or "follow". The *nouns annotated as events have a much lower coverage* (7.89% in Gold and 8.62% in Platinum). Interestingly, in the *Platinum dataset, the rate of verbs annotated as events is lower compared to the Gold dataset, but the rate of nouns annotated as events* *is higher than the Gold dataset.* Since, on average, not more than 1% of the total amount of adjectives, adverbs and prepositions were annotated as events by the experts in both datasets, we assume they might introduce ambiguity.

In both datasets, around 50% of all the annotated time expressions consists of single tokens of POS noun, numeral, adjective and adverb. While the rate of nouns and numerals annotated as times in the Platinum dataset is almost equal, in the Gold dataset, there are around 4 times more nouns annotated as time expressions compared to numerals. All the multi-token time expressions are combinations of tokens having at least a noun or a numeral.

5.2.2 Tokens and Lemmas

Table 4 presents the overview of the potential inconsistencies encountered in the expert-annotated events in the Platinum dataset, by looking at event tokens and lemmas across all (*ALL*) POS tags and per individual POS tag. As in the analysis performed in [33], we identify possible inconsistencies at the token level - *not all instances of an event are always annotated as events* (*e.g.*, the noun "apology" is annotated as event in 1 out of 6 cases, the verb "keep" is annotated as event in 1 case out of 9). This type of inconsistency appears for 74 distinct event tokens out of a total of 537 distinct event token - POS tag pairs (*i.e.*, 13.85% cases). Similarly, we also identify inconsistencies at the lemma level - *not all lemma instances of an event are always annotated as events* (*e.g.*, the noun "charge" is annotated as event in 1 out of 5 lemma-based occurrences, the verb "say" is annotated as event in 63 cases out of 65). There are 90 such distinct lemma-based inconsistency cases out of 440 unique pairs event lemma - POS tag (*i.e.*, 20.59% cases). The amount of *inconsistencies at the level of event lemma is higher than at the level of event token*, which means that only certain lemmas of a token are usually annotated as events by experts. Overall, the least amount of disagreement is seen for events that are either verbs or nouns.

Table 4 Event Inconsistencies at the Level of Event Tokens and Lemmas in TempEval-3 Platinum.

	Total Inconsistencies (%)		**Distinct Inconsistencies (%)**	
	Token	Lemma	Token	Lemma
ALL	287 (27.86%)	476 (39.04%)	74 (13.85%)	90 (20.59%)
VB	215 (28.25%)	388 (41.54%)	42 (11.26%)	53 (18.79%)
NN	66 (27.61%)	82 (32.15%)	27 (19.56%)	32 (24.24%)
JJ	5 (19.23%)	5 (19.23%)	4 (20.0%)	4 (20.0%)
RB	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Regarding time expressions, we observed that in the Platinum dataset year mentions such as "1953", "2010" are not annotated as time expressions by experts. Further, we looked into the multi-token time expressions and computed how many times a mention was missed. In the Platinum dataset, we found only two missed mentions, both at the level of token and lemma, while in the Gold dataset we found 91 missed mentions at the token level and 105 mentions at the lemma level. Overall, 46 time expression mentions were not always annotated out of 497 unique time expression tokens and 492 time expressions lemmas.

5.2.3 Sentences without Annotated Events

In Figure 2 and Figure 3 we plotted for each sentence without annotated events (in the TempEval-3 Gold dataset and respectively, in the TempEval-3 Platinum dataset) on the first *y axis* the number of tokens in each sentence (ordered) and on the second *y axis (1)* the total number of verb POS tags contained in the sentence and *(2)* the total number of event lemmas

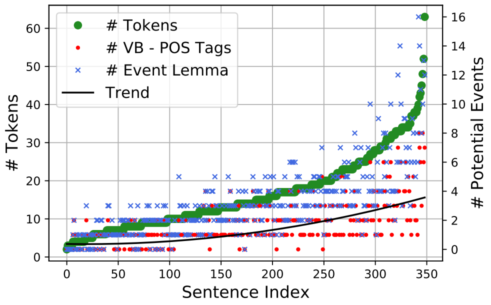


Figure 2 Overview of Potentially Missed Events in Sentences from the TempEval-3 Gold Dataset without Expert Event Annotations.

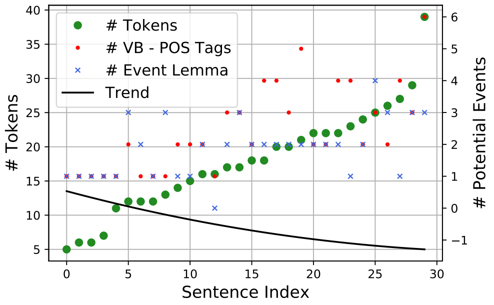


Figure 3 Overview of Potentially Missed Events in Sentences from the TempEval-3 Platinum Dataset without Expert Event Annotations.

that were annotated in other sentences, but not the current one. We observe a positive correlation between the number of verb POS tags contained in the sentences and the number of annotated event lemmas in other sentences, which means that many of the verbs in these sentences were actually tagged as events in other sentences. Even though the correlation does not seem as strong for the sentences in the TempEval-3 Platinum dataset (Figure 3, we believe this is due to the low number of sentences. Therefore, based on these observations and the ones presented in the previous subsections, we re-emphasize the incompleteness in the expert annotations, closely correlated to our hypotheses **H1.3-4**.

6 Results

In this section we report on the results[8] of the pilot and main *crowdsourcing experiments* in Section 6.1 and the results of employing the crowd-annotated events to train and evaluate an event extraction system in Section 6.2.

6.1 Crowdsourcing Experiments

In the 16 *crowdsourcing pilot experiments* we gathered in total 8,000 crowd annotations from a total of 134 unique workers. The total cost of these pilots was equal to $624. We start by evaluating the performance of the crowd in terms of precision (P), recall (R) and F1-score, in comparison with the expert annotations, in each pilot experiment. In Table 5 we see the overview of this analysis. To compare the crowd annotations with the expert annotations, we first applied the crowd aggregation metrics introduced in Section 4.2.1. As a result, each entity (either event or time expression) validated by the crowd gets an entity-sentence score (ESS) with values between 0 and 1, which shows the likelihood of that entity to be valid. First of all, we observe that the crowd performs better when they are provided with explicit definitions of the entities that they need to validate (see results for P1, P2, P3). Second, in alignment with our **H2.1** hypothesis and confirming it, we observe that when the crowd is asked to motivate their answers, their performance is improved (see results for P3 and P6).

As described in Section 4.2, in P7 and P8 we increased considerably the list of entities to be validated by the crowd. Furthermore, in P8 we also gave them the option to highlight

[8] `https://github.com/CrowdTruth/Event-Extraction`

Table 5 Crowd vs. Experts Performance Comparison on all Crowdsourcing Pilot Experiments.

	Events					Time Expressions				
	Thresh	P	R	F1-score	#TP	Thresh	P	R	F1-score	#TP
P1	0.35	0.84	0.93	0.89	152	0.60	0.71	0.86	0.78	50
P2	0.15	0.79	1.0	0.88	164	0.50	0.67	0.86	0.75	50
P3	0.50	0.83	0.98	**0.90**	161	0.60	0.76	0.84	**0.80**	49
P4	0.40	0.84	0.95	0.89	154	0.65	0.73	0.82	0.78	48
P5	0.35	0.80	0.98	0.88	159	0.65	0.80	0.72	0.76	42
P6	0.45	0.84	0.95	**0.89**	157	0.60	0.79	0.81	**0.80**	47
P7	0.45	0.75	0.95	0.84	156	0.65	0.75	0.83	0.78	48
P8	0.50	0.73	0.93	0.83	155	0.75	0.78	0.77	0.78	45

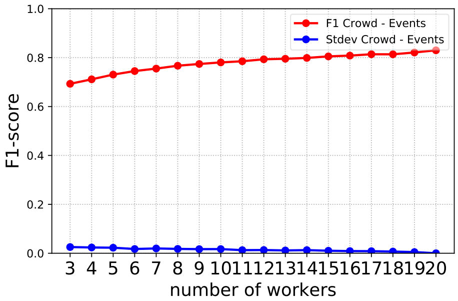


Figure 4 Events Crowd F1-score at the Best ESS Threshold for Various # Workers.

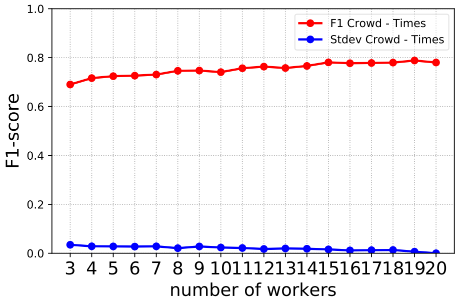


Figure 5 Times Crowd F1-score at the Best ESS Threshold for Various # Workers.

potentially missing entities, *i.e.*, entities that are not found in the validation list. However, the crowd still performs well when compared to the experts. Even though the overall F1-score slightly dropped, the total number of true positive entities remains almost the same. The drop in F1-score is due to the fact that the crowd finds more relevant entities than the ones annotated by experts. Thus, we hypothesize that this is a viable and reliable way of gathering missing entities and correct the expert inconsistencies. Therefore, based on these observations, we ran the *main experiment* using the P8 setup.

Next, we focused on understanding what would be the optimal number of crowd annotations needed per sentence, at the best performing *ESS* threshold for the crowd. For each number of workers between 3 and 20, we averaged their F1-score for a total of 100 runs, by randomly generating sets of [3:20] workers. In Figure 4 and Figure 5 we plot both the average F1-score and the standard deviation (stdev) among all the runs for the pilot experiment P8, for events and respectively, time expressions. In both cases, we observe that around 15 workers the F1-score of the crowd stabilizes and the stdev is negligible. Furthermore, this observation aligns with our **H2.2** hypothesis which says that enough annotations from the crowd provides reliable results when compared to experts.

In the *main experiments* we gathered 63,030 crowd annotations from 160 unique workers and the total cost of the experiments was $3,112, by running the setup of P8 with 15 workers, on the entire set of sentences. In order to see how the crowd compares to the expert annotations, we again performed the evaluation of the crowd entities for every entity-sentence score threshold. Thus, for time expressions we got the best performing F1-score of 0.70 at thresholds between [0.65 and 0.90] and for events we got the best performing F1-score of

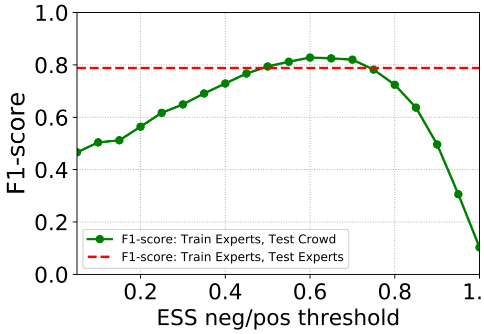


Figure 6 ClearTK F1-score when Trained on Expert Events and Tested on Crowd Events.

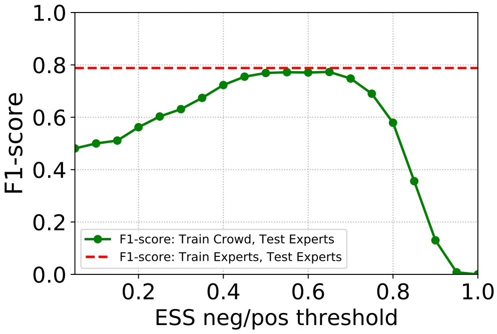


Figure 7 ClearTK F1-score when Trained on Crowd Events and Tested on Expert Events.

0.81 at a threshold of 0.60. Overall, we see that these results are consistent with the ones in the pilot experiments, even though the scale is much larger. Therefore, we acknowledge that the crowd is able to provide *consistent event and time expression annotations.*

Table 6 ClearTK F1-score when Trained on Crowd Events and Tested with Crowd Events.

Crowd ESS Threshold		Test								
		0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
Train	0.30	**0.824**	**0.806**	0.783	0.75	0.721	0.697	0.669	0.649	0.623
	0.35	0.797	**0.798**	**0.798**	0.786	0.764	0.744	0.72	0.699	0.674
	0.40	0.766	0.783	**0.797**	**0.799**	0.791	0.778	0.765	0.745	0.72
	0.45	0.738	0.769	0.797	**0.818**	**0.823**	0.81	0.802	0.79	0.768
	0.50	0.71	0.747	0.779	0.814	**0.828**	**0.827**	**0.829**	0.815	0.796
	0.55	0.687	0.727	0.761	0.799	0.821	**0.826**	**0.83**	0.819	0.804
	0.60	0.658	0.698	0.735	0.776	0.802	0.816	**0.826**	0.824	0.819
	0.65	0.639	0.681	0.721	0.764	0.79	0.807	0.820	**0.822**	0.819
	0.70	0.596	0.638	0.673	0.716	0.747	0.771	0.791	0.800	**0.805**

6.2 Training and Evaluating with Crowd Events

We report on the results of the ClearTK event extraction systems, when trained and evaluated on crowd-annotated events. It is important to acknowledge that for training purposes we used the systems' parameters that performed the best in the TempEval-3 task, and we did not fine-tuned them to better fit our training data.

In Figure 6 we plotted the F1-score of the system when trained on expert events and evaluated on crowd events, for every event-sentence score (*ESS*) threshold. We can observe that between the *ESS* thresholds [0.5:0.75] the system performs much better than when it is evaluated on the expert events. The measured F1-score of the ClearTK system in the TempEval-3 task was 0.788, while the maximum achieved F1-score when evaluated on crowd events reaches values of around 0.83. However, when we train the system on crowd events and we test it on expert events, the performance achieved by the system is only almost as good (0.77) as the reported F1-score of 0.788. This happens due to the fact that the crowd annotates events in a more consistent way, while experts, according to Section 5, are missing potentially valid annotations. Finally, in Table 6 we show the results of both training and evaluating the ClearTK system on crowd events, for each ESS threshold between [0.30:0.70]. The results clearly indicate that the crowd event annotations are a reliable and consistent way of providing event annotations (correlated to **H2.3**) - the crowd performs the best when trained and evaluated at similar ESS thresholds. Furthermore, we observe that while for training the best performing threshold could vary between [0.50:0.60], for testing the threshold of 0.60 seems to provide the best and most consistent F1-scores, up to 0.830.

7 Conclusion and Future Work

In this paper we proposed a data-agnostic validation methodology for expert-annotated datasets and we showed its application on the case of events and, to some extent, time expressions. We propose a set of analytics to measure the consistency and completeness of such datasets and a crowdsourcing approach to mitigate these problems. We conducted extensive pilot crowdsourcing experiments and we derived the optimal setup to gather event and time expression annotations based on them. We showed that the crowd-annotated events are a reliable dataset to train and evaluate state-of-the-art event extraction systems. Furthermore, we showed that the performance of such systems can be improved by at least 5.3% when both trained and evaluated on crowd data.

As part of future work we plan to use the crowd-annotated events for *(1)* training and evaluating a larger range of state-of-the-art event extraction systems, as well as *(2)* running more extensive experiments such as fine-tunning the learning parameters based on the crowd-training data and using different crowd event thresholds. Furthermore, we plan to investigate the impact that ambiguous events have in training and evaluating event extraction tools. Finally, we plan to replicate the experiment with time expressions and investigate the added value of gathering crowd annotations for this mention type.

References

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

2 S. Bethard. ClearTK-TimeML: A minimalist approach to TempEval 2013. In ** SEM, Volume 2: SemEval 2013*, volume 2, pages 10–14, 2013.

3 K. Braunschweig, M. Thiele, J. Eberius, and W. Lehner. Enhancing named entity extraction by effectively incorporating the crowd. *BTW Workshop*, 2013.

4 K. Cao, X. Li, M. Fan, and R. Grishman. Improving event detection with active learning. In *International Conference Recent Advances in Natural Language Processing*, pages 72–77, 2015.

5 T. Caselli and O. Inel. Crowdsourcing StoryLines: Harnessing the Crowd for Causal Relation Annotation. In *Proceedings of the Workshop Events and Stories in the News*, 2018.

6 T. Caselli and R. Morante. Systems' Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task. In *LREC*, 2018.

7 T. Caselli, R. Sprugnoli, and O. Inel. Temporal Information Annotation: Crowd vs. Experts. In *LREC*, 2016.

8 A. Ceroni, U. Gadiraju, and M. Fisichella. Justevents: A crowdsourced corpus for event validation with strict temporal constraints. In *ECIR*, pages 484–492, 2017.

9 N. Chambers. NavyTime: Event and time ordering from raw text. Technical report, Naval Academy Annapolis MD, 2013.

10 A. Chang and C. D. Manning. SUTime: Evaluation in tempeval-3. In ** SEM, Volume 2: SemEval 2013*, volume 2, pages 78–82, 2013.

11 G. Demartini. Hybrid human–machine information systems: Challenges and opportunities. *Computer Networks*, 90:5–13, 2015.

12 A. Dumitrache, L. Aroyo, and C. Welty. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *HCOMP 2018*, pages 12–20, 2018.

13 A. Dumitrache, O. Inel, L. Aroyo, B. Timmermans, and C. Welty. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. *arXiv preprint arXiv:1808.06080*, 2018.

14 A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *ESWC Conference*, pages 351–366, 2013.

15 O. Inel and L. Aroyo. Harnessing diversity in crowds and machines for better NER performance. In *European Semantic Web Conference*, pages 289–304, 2017.

16 O. Inel, L. Aroyo, C. Welty, and R.-J. Sips. Domain-independent quality measures for crowd truth disagreement. *DeRiVE Workshop*, page 2, 2013.

17 O. Inel, G. Haralabopoulos, D. Li, C. Van Gysel, Z. Szlávik, E. Simperl, E. Kanoulas, and L. Aroyo. Studying Topical Relevance with Evidence-based Crowdsourcing. In *CIKM*, pages 1253–1262. ACM, 2018.

18 H. Jung and A. Stent. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In **SEM, Volume 2: SemEval 2013*, volume 2, pages 20–24, 2013.

19 O. Kolomiyets and M.-F. Moens. KUL: Data-driven approach to temporal parsing of newswire articles. In ** SEM, Volume 2: SemEval 2013*, volume 2, pages 83–87, 2013.

20 A. K. Kolya, A. Kundu, R. Gupta, A. Ekbal, and S. Bandyopadhyay. JU_CSE: A CRF based approach to annotation of temporal expression, event and temporal relations. In **SEM, Volume 2: SemEval 2013*, volume 2, 2013.

21 K. Lee, Y. Artzi, Y. Choi, and L. Zettlemoyer. Event detection and factuality assessment with non-expert supervision. In *EMNLP*, pages 1643–1648, 2015.

22 S. Liao and R. Grishman. Using prediction from sentential scope to build a pseudo co-testing learner for event extraction. In *IJCNLP*, pages 714–722, 2011.

23 H. Llorens, E. Saquete, and B. Navarro. TIPsem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *SemEval*, 2010.

24 C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics System Demonstrations*, pages 55–60, 2014.

25 C. Min, M. Srikanth, and A. Fowler. LCC-TE: a hybrid approach to temporal relation identification in news text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 219–222, 2007.

26 J. Pustejovsky, R. Knippen, J. Littman, and R. Saurí. Temporal and event information in natural language text. *Language resources and evaluation*, 39(2):123–164, 2005.

27 J. Pustejovsky, J. Littman, R. Saurí, and M. Verhagen. TimeBank 1.2. *Linguistic Data Consortium*, 40, 2006.

28 R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. TimeML annotation guidelines. *Version*, 1(1):31, 2006.

29 R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263, 2008.

30 R. Sprugnoli and A. Lenci. Crowdsourcing for the identification of event nominals: an experiment. In *LREC*, pages 1949–1955, 2014.

31 J. Strötgen, J. Zell, and M. Gertz. Heideltime: Tuning english and developing spanish resources for tempeval-3. In ** SEM, Volume 2: SemEval 2013*, volume 2, pages 15–19, 2013.

32 N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, and J. Pustejovsky. SemEval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In ** SEM, Volume 2: SemEval 2013*, pages 1–9, 2013.

33 C. Van Son, O. Inel, R. Morante, L. Aroyo, and P. Vossen. Resource Interoperability for Sustainable Benchmarking: The Case of Events. In *LREC*, 2018.