Temporal Information Annotation: Crowd *vs.* Experts

Tommaso Caselli[∗], **Rachele Sprugnoli, Oana Inel**[∗♣]
[∗]The Network Institute Vrije Universiteit Amsterdam; FBK and University of Trento; [♣]IBM Nederland, CAS Benelux
De Boelelaan 1105 1081 HV Amsterdam; Via Sommarive, 18 38122 Povo (TN); Johan Huizingalaan 765, 1066 VH Amsterdam
{t.caselli;oana.inel}@vu.nl[∗], sprugnoli@fbk.eu

Abstract

This paper describes two sets of crowdsourcing experiments on temporal information annotation conducted on two languages, i.e., English and Italian. The first experiment, launched on the CrowdFlower platform, was aimed at classifying temporal relations given target entities. The second one, relying on the CrowdTruth metric, consisted in two subtasks: one devoted to the recognition of events and temporal expressions and one to the detection and classification of temporal relations. The outcomes of the experiments suggest a valuable use of crowdsourcing annotations also for a complex task like Temporal Processing.

Keywords: crowdsourcing, temporal processing, corpus creation

1. Introduction

This paper reports on a set of crowdsourcing experiments on temporal information annotation whose goal is to compare the results of experts with those of the crowd for the annotation of temporal expressions, events and temporal relations. The results will allow us to gain better insights on the Temporal Processing task by suggesting changes in the current annotation practices.

The TimeML Annotation Guidelines (Pustejovsky et al., 2003a) and the TempEval evaluation campaigns[1] have greatly contributed to the development of TimeML-compliant annotation schemes in different languages and the setting of evaluation procedures for systems against public benchmark data.

Reviewing the performance of systems shows that Temporal Processing is not a trivial task, especially when dealing with temporal relations. In Table 1 we report the results of the best systems for Italian and English with respect to four tasks: a.) Temporal Expressions (TIMEXes) Identification; b.) Event Extraction; c.) Temporal Relation Identification and Classification from raw text (TLINKs raw); and d.) Temporal Relation Classification given Gold entities (TLINKs Gold). The figures have been extracted from TempEval-3 for English (UzZaman et al., 2013) and EVENTI for Italian (Caselli et al., 2014).

Although a direct comparison among the test sets cannot be done, the difference in the systems' performance in the two languages can hardly be explained by making reference only to language specific issues. Furthermore, the systems' performance seems to be also affected by the language specific annotation guidelines and the quality of the annotated corpora. Concerning the annotation guidelines, the two languages: a.) share the same annotation philosophy (i.e., adherence to the superficial text form), the same set of markables and values; b.) are compliant with the ISO-TimeML standard; c.) have benefited from a continuous collaboration among the groups which developed them; and, finally, d.) have been annotated or revised by experts. Table 1 also shows the figures for the inter-annotator agreement (IAA) for each task together with the different measures used (Precision and Recall, the Dice coefficient, and the Kappa score).

Task	Lang	IAA	System	F1
TIMEXes	IT	P&R=0.95	HeidelTime1.8	0.709
	EN	P&R=0.83	HeidelTime-t	0.776
EVENTs	IT	P&R=0.86	FBK-B1	0.867
	EN	P&R=0.78	ATT-1	0.81
TLINK Raw	IT	Dice=0.86	FBK_C1	0.264
	EN	P&R=0.55	ClearTK-2	0.309
TLINK Gold	IT	K=0.88	FBK_D1	0.736
	EN	K=0.71	UTTime-1,4	0.564

Table 1: Inter-annotator agreement (IAA) for Italian (IT) and English (EN) together with system performance comparison (F1) for the two languages.

The remainder of the paper is organized as follows: in Section 2. we will shortly revise current state of the art in the use of crowdsourcing for Temporal Processing. Sections 3. and 4. will report on the two sets of experiments we have conducted: the first on crowdsourcing temporal relations given target entities, and the second on crowdsourcing temporal relations from raw text. Finally, Section 5. reports on insights from the experiments and provides directions for future work.

2. Related Works

Crowdsourcing has been extensively used for lots of tasks in NLP (e.g., evaluate quality of automatic translations, identify entailment pairs, among others). On the other hand, the use of crowdsourcing in the perspective of Temporal Processing has been mainly limited to studies which aim at assessing the difficulty of the task and the salience of linguistic and extralinguistic cues with a particular focus on the temporal relations rather than on all the subtasks involved as illustrated in Table 1 (Mani and Schiffman, 2005; Moeschler, 2000; Caselli and Prodanof, 2010). In Mani and Schiffman (2005), the authors developed an annotation experiment on ordering pairs of successively described events

[1] TempEval 2007 (Verhagen et al., 2007): `http://www.timeml.org/tempeval/`; TempEval 2010 (Verhagen et al., 2010): `http://www.timeml.org/tempeval2/`; TempEval 2013: `http://www.cs.york.ac.uk/semeval-2013/task1/`

in the past to assess how often the narrative convention is followed in a corpus of news. Six different temporal relations were selected, namely, *Entirely Before*, *Entirely After*, *Upto*, *Since*, *Equal* and *Unclear*. The initial IAA on the value of the temporal relations is 0.50 which is improved to 0.61, when reducing the distinction between *Entirely Before* - *Equal* and between *Entirely Before* and *Upto*. Similar results for IAA (i.e., 0.58) have been reported in Caselli and Prodanof (2010) for Italian.

A different approach has been followed by Ng and Khan (2012). The task was limited to annotating temporal relations between temporal expressions and events in the same sentence. Unfortunately, no direct results on the crowdsourced data are reported. The crowdsourced annotations have then been used to train an SVM model for classifying temporal relations between an event and a temporal expression in English, as defined in the Task C in TempEval-2 (Verhagen et al., 2010). The authors report an accuracy of only 67.4% for the TempEval training, of 65.2% for the crowdsourced data and, finally, of 71.7% when merging TempEval training and crowdsourced data. The low difference in performance between the crowdsourced and the TempEval training data suggests that the non-expert annotators were able to perform the task with a good level of accuracy comparable to that of experts for this task.

Major works on other subtasks of Temporal Processing mainly focused on event detection (Aroyo and Welty, 2012; Sprugnoli and Lenci, 2014). In Aroyo and Welty (2012) the focus of crowdsourcing is not on assessing the ability of the crowd to perform a specific task, i.e., event detection, but on disagreement as a "natural state" suggesting that event semantics are imprecise and varied. On the other hand, Sprugnoli and Lenci (2014) evaluated the ability of the crowd in detecting event nominals in Italian, pointing out the complexity of this task due to the presence of ambiguous patterns of polysemy.

3. Experiment 1: Crowdsourcing Temporal Relations with Given Entities

The first experiment focuses on the identification of temporal relations between verb pairs in Italian and English sentences extracted from the MultiSemCor corpus (Bentivogli and Pianta, 2005). The goals were to assess: a.) if crowd and experts have similar approaches in identifying and classifying temporal relations on given target elements; and b.) if syntax has a role in facilitating the identification of temporal relations in Italian and in English.

We extracted fifty aligned parallel sentences in the two languages from MultiSemCor and we selected two types of temporal relations following expert annotation subtasks:

- relations between the main event and its subordinated event (e.g., *So_[MAIN] che hai visto_[SUB.] Giovanni / I know_[MAIN] you've seen_[SUB.] John*);
- relations between two main events (e.g., *Giovanni bussò_[MAIN] ed entrò_[MAIN] / John knocked_[MAIN] and got_[MAIN] in*).

In the former case, events belonging to four different TimeML classes (PERCEPTION, REPORTING, I_ACTION, and I_STATE) were selected. As for the relations between main events, a random selection of event classes has been performed.

Two jobs, one for Italian and one for English, were built using the services of CrowdFlower,[2] with the same instructions and settings.

In each sentence, an expert applied the language specific annotation guidelines to identify the source and target verbs standing in a (possible) temporal relation. Source verbs were highlighted in green while the target verbs were highlighted in yellow. Contributors were asked to read the sentences and select the temporal relation between the word in yellow and the word in green choosing among 8 different values : `AFTER`, `BEFORE`, `INCLUDES`, `IS_INCLUDED`, `SIMULTANEOUS`, `NO_RELATION`, `OTHER`, and `DON'T KNOW`. A simple graphical visualization of temporal relations inspired by Allen's representations (Allen, 1991) was added to the instructions to improve the explanation of the temporal relation values. Finally, all temporal relations of both datasets have been annotated by an expert in order to evaluate the accuracy of the crowd judgments.

For each sentence, 5 different judgments were collected. We selected basic level contributors[3] and restricted the geographical location of contributors to Italy, USA and UK. We used the built-in quality control mechanism of CrowdFlower to distinguish between reliable and unreliable contributors. This mechanism is based on the presence of a set of work units with known answer in advance (i.e., a gold standard): reliable contributors are those who provide a correct answer for at least 70% of these units. All the others are automatically blocked and their judgements are discarded from the final results. In our tasks a gold standard of 5 sentences (i.e., 10% of the dataset), was added to both datasets.

3.1. Results and Discussion

108 contributors participated in the Italian job but only 13 passed the minimum level of accuracy required by CrowdFlower over the gold standard sentences. As for the English data only 13 out of 345 contributors were considered reliable. The results for accuracy and Fleiss kappa agreement score for Italian and English are reported in Table 2.

		ACCURACY	IAA
ITALIAN	Overall	70%	0.41
	Main-Main	72%	0.52
	Main-Sub	68%	0.25
ENGLISH	Overall	63%	0.32
	Main-Main	48%	0.24
	Main-Sub	79%	0.38

Table 2: Experiment 1: Results of accuracy and inter-annotator agreement on the Italian and English datasets.

As the overall accuracy in both languages shows, and in line with (Moeschler, 2000; Mani and Schiffman, 2005;

[2] `http://www.crowdflower.com/`

[3] Level 1 corresponds to "high performance contributors who account for 60% of monthly judgments and maintain a high level of accuracy across a basket of CrowdFlower jobs" `http://success.crowdflower.com/customer/portal/articles/1385183`.

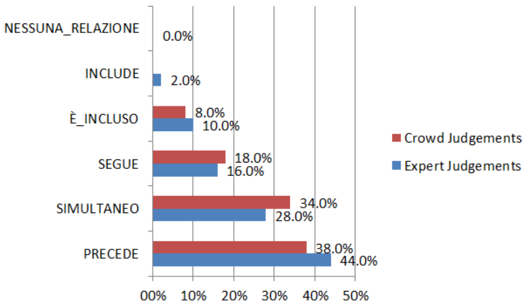


Figure 1: Answer distribution on the Italian dataset

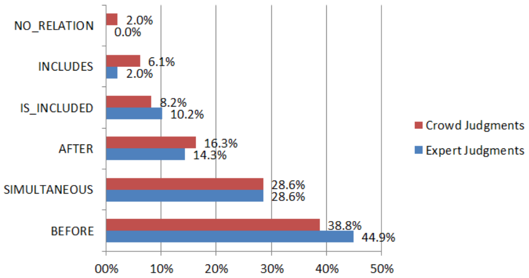


Figure 2: Answer distribution on the English dataset

Caselli and Prodanof, 2010), identifying and classifying temporal relations is a challenging task, regardless of the language in analysis. Further support is given by both global IAA scores (K=0.41 for Italian, K=0.32 for English). By observing the main sources of disagreements between crowd workers and the expert in both languages, the most frequent temporal values are SIMULTANEOUS (7 in Italian and 5 in English) and INCLUDES/IS_INCLUDED (6 in Italian and 7 in English). The English contributors provided also disagreements on the value BEFORE (4 cases) where the expert's judgment was SIMULTANEOUS and IS_INCLUDED. The distinction between SIMULTANEOUS and INCLUDES/IS_INCLUDED is very subtle, as SIMULTANEOUS can be though as a special case of INCLUDES/IS_INCLUDED. This suggests that average speakers have difficulties in identifying fine-grained values without specific training, detailed instructions and explicit markers of a temporal value.

When focusing on the two sets of temporal relations, we can observe that the Italian contributors obtained comparable results (72% for "Main-Main" relations and 68% for "Main-Sub"), while in English the contributors had a better accuracy for "Main-Sub" relations (79% *vs.* 48%, for "Main-Main"). The accuracy for the "Main-Sub" in both languages can be explained by the fact that main verbs function like temporal anchors for subordinated events, thus restricting possible temporal interpretations.

Finally, Figures 1 and 2 show the answer distribution of the expert annotator and non-expert contributors on the Italian and English datasets, respectively. It is interesting to remark that in both cases the answers given by non-expert contributors have the same general trend in terms of distributions of temporal relations values and that the NO_RELATION value is very rarely chosen (never in Italian and only 2% in English).

4. Experiment 2: Crowdsourcing Temporal Relations from Raw Text

The previous experiment suffers from two issues: a.) the events to be put in temporal relations were given by experts, thus, forcing the crowd to stick to annotation guidelines; and b.) the set of sentences in both languages is limited. Following the works of Soberon et al. (2013) and Inel et al. (2013), we designed an additional set of experiments in English and in Italian where the crowd is asked to perform the Temporal Processing task from raw texts: namely, we ask the crowd to identify event descriptions and temporal expressions, and, then, on top of these crowd annotated elements, the presence of temporal relations and their values. With respect to the previous experiment, we changed some parameters:

- 200 random sentences have been extracted from the English and Italian TimeBank corpora (Pustejovsky et al., 2003; Caselli et al., 2011), respectively;
- the use of the CrowdTruth metrics (Inel et al., 2014) rather than CrowdFlower internal quality control based on the gold standard data for cleaning the data from spammers and evaluating their quality.

With this second set of experiments, we aim to replicate a more realistic annotation scenario as the crowd workers will perform all subtasks involved in the temporal annotation of documents from raw text data.

4.1. The CrowdTruth Metric

The goal of the CrowdTruth methodology is a.) to distinguish the quality workers from the low-quality workers, and b.) to assess how well a given label (*e.g.*, an event, or a relation) is expressed by the input data. The first step in applying the CrowdTruth metrics on the given task is to translate the workers judgments into a worker annotation vector. This enables us to further compare the results by means of cosine similarity measures. The length of the vector depends on the number of possible answers in a question, while the number of such vectors depends on the number of questions contained in the task. If the worker selects a particular answer, its corresponding component would be marked with 1, and 0 otherwise. The worker annotation vectors component for this experiment can be exemplified as follows:

- *Event and Temporal Expression Detection*: for each unit we construct two vectors, i.e., a vector of events and a vector of temporal expressions, having the dimension equal to the total number of words in the sentence and the option "none", if no word in the sentence refers to an event or a temporal expression;
- *Temporal Relation Detection and Classification*: for each unit we construct a vector of dimension 5 with the following values: BEFORE, AFTER, SIMULTANEOUS, OVERLAPPING and NONE, if

there is no temporal relation expressed between the two word phrases.

Similarly, we compute a media unit vector which is the result of adding up all the workers annotation vectors for that unit. Next, we apply two worker metrics that differentiate between quality and low-quality workers by providing insights on (1) how close a worker performs compared to workers solving the same task and (2) how much a worker disagrees with the rest of the workers in the context of all units. These two metrics are also computed using the cosine similarity between (1) each two workers annotation vectors on a unit and (2) the annotation vectors of a worker and the aggregated annotations of the rest of the workers (but subtracting the worker vector). If the worker values are below a given threshold, the worker is marked as low-quality, i.e., a spammer, and his annotations are removed.

To determine how well an annotation is expressed in a unit, i.e., a sentence or a pair of word phrases, we compute the unit-annotation score, or clarity score, on the spam-filtered data. This metric is measured for each possible annotation on each unit as the cosine between the unit vector for that annotation and the media unit vector. A more detailed description of these metrics can be found in (Soberon et al., 2013; Aroyo and Welty, 2014).

4.2. Event and Temporal Expression Detection

We run an overall of seven different jobs, i.e., 3 jobs for English data and 4 jobs for Italian data, for event and temporal expressions detection. We provided the same instructions for English and Italian. Workers were allowed to select both single tokens and multi-token expressions and then had to decide if the identified word(s) was an event or a temporal expression. For each sentence, we collected a total of 15 judgments. Each worker was allowed to annotate a maximum of 10 sentences (e.g. 10 judgments). We used a basic definition of event as "something that has happened, is happening or will happen in the future". Temporal expressions were defined as words or phrases "expressing time".

As for the English data, a total of 372 workers from USA, UK, Australia and Canada participated in the experiments; 124 (33.33%) were identified as spammers on the basis of CrowdTruth metrics. For the Italian dataset, we collected judgments from 371 workers from Italy. By applying the CrowdTruth metrics, we identified 115 spammers (30.99%).

We further analyzed the data with the clarity score to compare the ability of the crowd(s) versus the experts: the higher is the clarity score, the more accurate and reliable are the crowd judgments. We will report the analysis by grouping the data per markable type, i.e., either event or temporal expression, and per language.

Concerning the annotation of events in English, 1296 tokens were judged as expressing an event, while in Italian only 1040 tokens were annotated. To compare the performance of the crowd(s) and the experts for this substask, we analyzed the number of overlapping tokens per clarity score thresholds. In Table 3 we report, for different clarity thresholds and for both languages, the number of tokens marked as events by the crowd(s) together with the overlap with the experts.

CLARITY	# CROWD EVENT TOKENS		CROWD-EXPERT OVERLAPPING EVENT TOKENS	
	EN	IT	EN	IT
≥0.2	1121	566	355 (31.66%)	342 (60.42%)
≥0.3	628	358	270 (42.99%)	251 (70.11%)
≥0.4	314	184	168 (53.50%)	145 (78.80%)
≥0.5	164	100	103 (62.80%)	80 (80%)
≥0.6	71	60	52 (73.23%)	51 (85%)

Table 3: Crowd *vs.* Expert: Event token annotation in English (EN) and Italian (IT)

With no threshold for clarity score, we identified 444 tokens (34.26%) which overlap expert annotation in the English data (TimeBank corpus), covering 84.25% of all events annotated by experts (527). On the other hand, for the Italian data, we identified 473 tokens which overlap with expert annotation (Ita-TimeBank corpus), covering only 53.87% of all event tokens annotated by the experts (878). With different clarity thresholds the number of annotated tokens by the crowd(s) get reduced (e.g. from 1,121 tokens with score ≥0.2 to 71 tokens with score ≥0.6 for English; from 566 tokens with score ≥0.2 to 60 tokens with score ≥0.6 for Italian) but the quality of the annotation improves, i.e., they are more reliable and in line with the expert data.

By analyzing mismatches in the event annotation between the crowd(s) and the experts we can observe that:

- with a threshold ≥ 0.3, 274 tokens in English are candidates of multi-token events such as noun phrases (*national callup*, *global embargo*), phrasal verbs (*fall apart*, *going up*), multiword expressions (*coup d'etat*), verbs accompanied by auxilliaries (*were offset*, *have fallen*), and copular constructions (*were lower*);

- with a threshold ≥ 0.3, 77 tokens in Italian are possible multi-token events. Similarly to English, we identified noun phrases (*raccolta diretta*, *sconfitta definitiva*), verbs accompanied by auxiliaries (*ha commentato*), multiword expressions (*5,000 metri*), and proper nouns (*Cross della Vallagarina*);

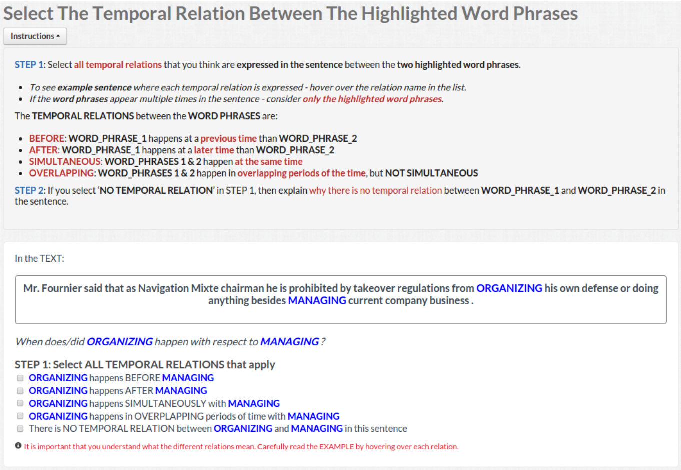
- the crowd annotation in English has identified 12 candidate event tokens which are missing in the expert data and has also provided annotations for 4 sentences which the experts did not annotate. The missing annotations are mainly nominal events (*trading*, *operations*) or verbs (*clobbered*, *cut*);

- the crowd annotation for Italian has identified 13 missing event tokens in the expert data. The missing annotations mainly correspond to named events (*Flushing Meadows*) and nominal events (*scadenza*, *cadute*).

The results for the temporal expression annotation subtask are illustrated in Table 4. For the English data, the crowd annotated 331 tokens: 158 (47.73%) are annotated also in the TimeBank corpus and correspond to more than 80% of all the temporal expression tokens annotated by the experts (197 tokens). Similar figures hold for the Italian data: the crowd has annotated a total of 231 tokens: 133 (57.57%) overlap with the Ita-TimeBank annotations and correspond to almost 70% of all temporal expression tokens identified

CLARITY	# CROWD TIMEX TOKENS		CROWD-EXPERT OVERLAPPING TIMEX TOKENS	
	EN	IT	EN	IT
≥0.2	205	162	156 (76.09%)	122 (75.30%)
≥0.3	185	134	152 (82.16%)	116 (86.56%)
≥0.4	162	111	136 (83.95%)	102 (91.89%)
≥0.5	139	92	123 (88.48%)	87 (94.56%)
≥0.6	97	73	88 (90.72%)	70 (95.89%)

Table 4: Crowd *vs.* Expert: Temporal Expression (TIMEX) token annotation in English (EN) and Italian (IT)



Select The Temporal Relation Between The Highlighted Word Phrases

STEP 1: Select all temporal relations that you think are expressed in the sentence between the two highlighted word phrases.

- *To see example sentence where each temporal relation is expressed - hover over the relation name in the list.*
- *If the word phrases appear multiple times in the sentence - consider only the highlighted word phrases.*

The TEMPORAL RELATIONS between the WORD PHRASES are:

- BEFORE: WORD_PHRASE_1 happens at a previous time than WORD_PHRASE_2
- AFTER: WORD_PHRASE_1 happens at a later time than WORD_PHRASE_2
- SIMULTANEOUS: WORD_PHRASES 1 & 2 happen at the same time
- OVERLAPPING: WORD_PHRASES 1 & 2 happen in overlapping periods of the time, but NOT SIMULTANEOUS

STEP 2: If you select 'NO TEMPORAL RELATION' in STEP 1, then explain why there is no temporal relation between WORD_PHRASE_1 and WORD_PHRASE_2 in the sentence.

In the TEXT:

Mr. Fournier said that as Navigation Mixte chairman he is prohibited by takeover regulations from ORGANIZING his own defense or doing anything besides MANAGING current company business .

When does/did ORGANIZING happen with respect to MANAGING ?

STEP 1: Select ALL TEMPORAL RELATIONS that apply

□ ORGANIZING happens BEFORE MANAGING
□ ORGANIZING happens AFTER MANAGING
□ ORGANIZING happens SIMULTANEOUSLY with MANAGING
□ ORGANIZING happens in OVERPLAPPING periods of time with MANAGING
□ There is NO TEMPORAL RELATION between ORGANIZING and MANAGING in this sentence

ℹ It is important that you understand what the different relations mean. Carefully read the EXAMPLE by hovering over each relation.

Figure 3: Instructions for the temporal relation task using crowd annotated events and temporal expressions (English case).

by the experts (193). In both languages, absolute (*2001*) and relative (*last year*, *lo scorso anno*) temporal expressions were correctly identified as well as points (*yesterday*, *ieri*) and durations (*nine months*, *tre giorni*) of different granularity. Concerning the mismatches between crowd(s) and experts, we can observe that the textual span of a temporal expression is the most affected dimension. In particular, workers in both languages tend not to include premodifiers (e.g., adjectives) and determiners (e.g., articles and demonstratives) in the extent of multi-token temporal expressions. The subset of wrong temporal expression tokens with respect to the expert data has commonalities in the two languages:

- signals of a temporal relations (e.g. *before*, *immediately*, *precedente*, *frattempo*) tends to be annotated as temporal expressions;
- an overgeneration of temporal expressions is due to the inclusion of words which have a fuzzy temporal value and which are not annotated because they cannot be normalized (e.g. *periodic*, *tra pochissimo*, *momento*);

In addition to this, we have observed that for the English data the crowd identified 9 temporal expressions which are missing from the expert annotation. This does not apply to the Italian data.

4.3. Temporal Relations Detection and Classification

The temporal relation subtasks, i.e., detection and classification, was run in a similar way with respect to the event and temporal expression ones. Nevertheless, one of the first issues was to decide which tokens or set of tokens from the previous annotation(s) was to be selected. Using only the clarity scores on the tokens would results in unnatural text spans which would have not respected the outcome of the first annotation round from the crowd. As a matter of fact, the crowd(s) was allowed to annotate both single token and multi-token expressions. We adopted a new approach: first, all annotated tokens and text spans were sorted from the shorter to the longest, i.e., from single tokens to multi-tokens with an increasing size. We then start comparing the span size (by means of the tokens' offset) and content in order to promote and select tokens and multi-tokens annotations. Every time that a multi-token markable included a single token one (or a shorter multi-token), we increased by 1 the raw annotation score from the crowd(s) for the single token (or the shorter multi-token). After this operation, we applied again the CrowdTruth metrics and obtained new clarity scores. We then applied basic filtering tuned for each language and markables: for English, we used a clarity score ≥ 0.18 for events and temporal expressions, while for Italian we used a clarity score ≥ 0.09 for events and a clarity score ≥ 0.18 for temporal expressions. On top of this, for each sentence and for each eligible markable we created pairs of *[event, event]* and *[event, time]* to be used in the temporal relation subtasks.

Similarly to the previous subtasks, we provided the same instructions in both languages. We did not provide any particular definition of temporal relations nor a graphical visualization of them as in the first experiment thus leaving the workers relying on their own intuition. We simplified the set of temporal relations to four classes: `BEFORE`, `AFTER`, `SIMULTANEOUS` and `OVERLAPPING`. In addition to this, the crowd(s) could select for a `NO` temporal relation value. Each class of temporal relations was accompanied by a clarifying example. We avoided to use a value like `OTHER` as a temporal value: it was available in Experiment 1 but never selected by the crowds (see values in Figures 1 and 2). The relations we have selected can be thought as a coarse-grained set of temporal relations which could be easily understood by non-expert annotators. We allowed the workers to select as many temporal relation values as they thought were correct as a way to deal with ambiguous and vague cases. To guide the workers in the task, we highlighted each component of the *[event, event]* and *[event, time]* pairs in the target sentence. The annotation was obtained by collecting judgments with respect to a subset of basic questions asking the crowd when the target elements in a pair were happening one with respect to the other. In Figure 3 we report the English instructions.

We collected 12 judgments per pair and each worker was allowed to work on maximum 20 pairs. As for the English data, 3,547 total workers from USA, UK, Australia and Canada took part to the task, 947 workers (26.69%) were identified as spammers, and 18,881 annotations were retained as valid. For the Italian dataset, 2,503 workers from Italy took part to the task, 539 (21.53%) were classified

TLINK VALUES	TLINK VALUE FREQ.	AV. CLARITY SCORE
BEFORE	814	0.502
AFTER	355	0.258
SIMULT.	854	0.502
OVERLAP.	108	0.180
NO RELATION	7	0.031

Table 5: Crowd Annotation of Temporal Relations in English (EN)

TLINK VALUES	TLINK VALUE FREQ.	AV. CLARITY SCORE
BEFORE	579	0.382
AFTER	726	0.317
SIMULT.	517	0.458
OVERLAP.	176	0.230
NO RELATION	26	0.072

Table 6: Crowd Annotation of Temporal Relations in Italian (IT)

as spammers. Overall 18,710 judgments were considered valid. We present the analysis of the data separately per language.

4.4. English Temporal Relations

The English event and temporal expression subtask allowed us to create 2,019 pairs of markables. The temporal relation preference, i.e., how many times a value obtained the highest number of annotations, together with the average clarity score are reported in Table 5.

The figures in Table 5 show that the crowd has a tendency to identify temporal relations, i.e., to assume the presence of a temporal relation between the items in the pairs. Only 7 cases have received a preference for the absence of a temporal relation. Furthermore, two temporal values obtained the highest number of preferences, namely SIMULTANEOUS and BEFORE. This preference is also mirrored by the average values of the clarity scores associated to these temporal values which are the same, i.e., 0.502.

191 pairs received preference, i.e., same clarity score, for more than one temporal value: 165 pairs received 2 temporal values; 24 pairs 3 temporal values, and 2 pairs 4 temporal values. Focusing on the 165 pairs with double temporal values, we have observed that 128 pairs involved the SIMULTANEOUS value together with temporally close relations such as BEFORE, AFTER or OVERLAPPING. Only in 24 cases we have identified contradicting temporal values, i.e., BEFORE - AFTER, signaling difficult cases or errors in the annotations. In the rest of the cases, 13 pairs continue to exhibit closely related temporal relations such as AFTER - OVERLAPPING or BEFORE - OVERLAPPING suggesting that the identification of the correct temporal value is not a trivial task. Only 1 pair shows values OVERLAPPING - NO.

A comparison with the expert annotations has been conducted. We have identified only 242 pairs (11.98%) in common with the expert annotated data. Among them, 84 have a strict match with the expert annotations in terms of textual spans of the elements in the pairs (namely *[event, event]* pairs), while 158 have a partial match, i.e., either both markables or one of the markables in the pair differ with respect to the experts' annotation for the textual span. Furthermore, on this subset of 242 common temporal relations, we have investigated the agreement on the temporal values. Only in 76 cases the crowd and the experts agree (32 cases for perfect matches and 44 cases for partial matches) while disagreement occurs in 166 cases (52 for perfect matches and 114 for partial matches). It is interesting to notice that the majority of the disagreement with the experts for the temporal value concerns SIMULTANEOUS and OVERLAPPING with *[event, time]* pairs. We further investigated the relations between crowd annotations and expert data by analyzing in details 20 random sentences (10% of our dataset). In this subset of markable pairs, the experts annotated only 36 temporal relations while the crowd considered as valid 181 pairs[4]. We validated the 181 pairs with respect to two aspects: a.) correctness of the temporal relation, and b.) correctness of the temporal value (i.e., the value or values with maximum clarity score). It is important to point out that in this validation process we did not considered the application of the expert guidelines for the annotation but only if the temporal relation and the value annotated by the crowd were eligible or not. 119 out of 181 annotated temporal relations (65.74%) were considered valid and in 40 cases the temporal value assigned by crowd was considered incorrect by experts[5].

4.5. Italian Temporal Relations

As for the Italian data, and in line with results from the previous subtasks, the overall number of created pairs of markables is lower: 1,857. We report in Table 6 the temporal relation preference and the average clarity score.

Similarly to English, the Italian annotations show a tendency to identify temporal relations although the number of cases where no temporal relation is selected is higher (26 cases). Additional remarks concern the most frequent relations and the relationship with the average clarity score. In particular, we observe that AFTER is by far the most preferred relation (726 annotations) followed by BEFORE and SIMULTANEOUS (579 and 517 annotations, respectively). Nevertheless, the average clarity scores per relation, which can be interpreted as the crowd confidence on a specific value, provides different information. The highest average clarity score is associated with SIMULTANEOUS, followed by BEFORE and, only in third place, AFTER. This suggests that the Italian crowd was more confident when assigning SIMULTANEOUS or BEFORE values rather than AFTER, although this latter was the most frequently assigned value.

166 pairs received more than one temporal value: 149 pairs received 2 temporal values; 16 pairs 3 temporal values, and 1 pairs 4 temporal values. By analyzing in details the 149 pairs with 2 temporal values, we can observe that the data are more fragmented as 10 different combinations of values are available. However, when analyzing the data in terms of closely related temporal relations, we can group 92 pairs, where 38 involve the SIMULTANEOUS value with relations such as BEFORE, AFTER and OVERLAPPING and 54 involves the OVERLAPPING value with BEFORE and AFTER. Finally, 57 pairs exhibits contradicting values, with

[4] Note that this also includes possible duplicates in case of a different span of a target element.

[5] The validation has been conducted by one of the authors

43 pairs AFTER - BEFORE and 14 pairs involving NO temporal relations.

Comparing the crowd data with the Italian experts we have identified only 92 overlapping pairs: 32 have a strict match with the expert annotations and 60 only a partial match. When focusing on the temporal values, in 40 cases there is an agreement between the crowd and the expert data (15 pairs for perfect matches and 25 for partial matches), while disagreements amount to 52 cases (17 pairs for perfect matches and 35 for partial matches). Disagreements on the overlapping temporal relations (both partial and strict matches) involves mainly the distinction between SIMULTANEOUS and OVERLAPPING with pairs composed by *[event, time]*, where the crowd prefer to assign a SIMULTANEOUS relation and the experts assign OVERLAPPING.[6]

Similarly to the English data, we performed an analysis of 20 random sentences (10% of the data). Experts have annotated only 33 temporal relations while the crowd considered valid temporal relations for 295 pairs. The 295 pairs have been validated with the same method of the English dataset (first eligibility of a temporal relation and then correctness of the temporal value) and 110 of them resulted correct (37.28%). Only in 20 cases the expert and the crowd disagree on the temporal value.

4.6. Discussion

The first observation which emerges from this second set of experiments is that the English crowd and the Italian crowd have similar though different behaviors. In particular, in the crowd annotation of events and temporal expressions we have observed commonalities in terms of the text span of the markables and errors. For instance, for events the crowd tends to prefer "larger textual span" annotations than the experts by including participants (e.g. *held the stronghold Police arrested six Protestants*) or even assuming complex event representations (e.g. *driving under the influence of alcohol*). Differences in the text span of events should not be considered as real errors (i.e., wrong tokens marked as events) in the annotations but signal a more holistic understanding of events from the crowd(s) with respect to the analytic models preferred by the experts, in line with the results in (Aroyo and Welty, 2012).

For temporal expressions we can observe a more conservative approach (e.g. exclusion of articles and modifiers in general) and at the same time the inclusion of linguistic expressions which either: a.) do not trigger a temporal expressions but signal the presence of a temporal relation; or b.) are explicitly excluded from expert annotations because they cannot be modeled yet (e.g. *periodic*; *tra pochissimo*). We have observed that the crowd is able to identify missing annotations and that, at the same time, the use of clarity scores facilitate the filtering of the crowd annotated data thus allowing for the identification of reliable data. The annotation of events qualifies as a harder task than the annotation of temporal expressions as the results from Tables 3 and 4 show. The data also show a different behavior of the two crowds as far as the reliability of their annotations is concerned with the English crowd performing better than the Italian one.

The results for temporal relations continue to reproduce the commonalities and differences in the two crowds with the English crowd having a higher accuracy than the Italian one, as the analyses of the 10% of the sentences have shown. A conclusion we can drawn from the experiments is that the most difficult aspect of temporal annotation is the assignment of the correct temporal value rather than the identification of the basic elements and the existence or not of a temporal relation. Different explanations are at play to account for the discrepancy between the temporal relations identified by the crowd(s) and those identified by the experts. The annotation guidelines followed by the experts have affected the results. As for the English data, the differences between the crowd and the experts could be due to a conservative application of the annotation guidelines, i.e., limiting the annotation only to those cases where the expert annotators felt very confident. Further explanations can also be found in the method used to annotate the expert data. In particular, experts had to identify the pairs of elements which could stand in a temporal relations both at an intra-sentential level and at an inter-sentential level, while the crowd was presented with pre-selected pairs and only had to validate and select the best value(s). The annotation of the Italian temporal relations by the experts was limited to a specific subset (see (Caselli et al., 2014) for details) which have somehow limited the comparison between expert and crowd.

5. Conclusion and Future Work

In this paper we reported on two sets of experiments on temporal annotation aiming at identifying new insights on such a complex task using crowdsourcing. The results from the first experiment, classifying temporal relations given target entities, have confirmed that this is a difficult task for humans. Syntactic information may help in the identification of the correct temporal value but deciding among fine-grained temporal relations is not easy. On the other hand, machine performance with target entities given obtains good results (F1 0.564 for English and F1 0.736 for Italian) suggesting that detailed annotation guidelines can contribute to the performance of automatic tools but are often difficult to follow.

The second experiment has shown that: a.) the English and the Italian crowds have different levels of accuracy on the four subtasks we have experimented with (i.e., event ad temporal expression detection, and temporal relations detection and classification); b.) by means of the CrowdTruth clarity score we can select between reliable and not reliable annotations and, most importantly, analyze cases of disagreement between the crowd annotations not only in terms of wrong/correct but also in terms of complex/simple; and c.) temporal relation identification is a feasible task while major issues are (again) related to the classification of the temporal relations.

As Cassidy et al. (2014) have shown, one of the problems of expert annotated corpora is the limited amount of available temporal relations which affects the evaluation of system performances. The second experiment we have run has

[6] Following the expert annotation guidelines the value is IS_INCLUDED.

shown that the crowd can be reliably used to identify the presence or absence of a temporal relation between given markables rather than classify temporal relations. Furthermore, current annotation schemes should simplify the subset of possible temporal relations by using more coarse-grained value in order to be “more natural” with respect to the ability of humans in identifying the correct temporal values.

Finally, the outcomes of the experiments suggest that problems in the Temporal Processing task are not related to the task definition but rather in the amount (and quality) of the annotated data. The fact that both crowds have correctly identified much more temporal relations than those available in expert corpora suggests that crowdsourcing could be a viable solution to increase the amount of available data in these corpora. The granularity of the temporal relations is another issue which needs to be solved. More coarse-grained values must be used, even if the annotation is conducted by experts, as this will mirror in a better way the way we, as humans, deal and process temporal information. Data from the crowd could also be used to identify complex examples: as we have observed, disagreements does not necessarily corresponds to errors. These difficult cases can be further used to better evaluate system performance.

6. Acknowledgements

One of the authors wants to thanks the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3) for partially supporting this work.

7. Bibliographical References

Allen, J. F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355.

Aroyo, L. and Welty, C. (2012). Harnessing Disagreement for Event Semantics. In *Proceedings of DeRiVE 2012 Workshop, ISWC.*

Aroyo, L. and Welty, C. (2014). The Three Sides of CrowdTruth. *Journal of Human Computation*, 1:31–34.

Bentivogli, L. and Pianta, E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(3):247–261.

Caselli, T. and Prodanof, I. (2010). Robust Temporal Processing: from Model to System. *Special issue: Natural Language Processing and its Applications*, 46:29–40.

Caselli, T., Lenzi, V. B., Sprugnoli, R., Pianta, E., and Prodanof, I. (2011). Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop.*

Caselli, T., Sprugnoli, R., Speranza, M., and Monachini, M. (2014). EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*. Pisa University Press.

Cassidy, T., McDowell, B., Chambers, N., and Bethard, S. (2014). An Annotation Framework for Dense Event Ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, Maryland, June.

Inel, O., Aroyo, L., Welty, C., and Sips, R.-J. (2013). Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures. In *Proceedings of DeRiVE 2013 Workshop, ISWC.*

Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., and Sips, R.-J. (2014). CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In *The Semantic Web–ISWC 2014*, pages 486–504. Springer.

Mani, I. and Schiffman, B. (2005). Temporally anchoring and ordering events in news. *Time and Event Recognition in Natural Language.*

Moeschler, J. (2000). Le modèle des inférences directionelles. *Cahiers des Linguistique Françaises*, 22:57–100.

Ng, J.-P. and Kan, M.-Y. (2012). Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations. In *Proceedings of COLING 2012.*

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003). The TimeBank corpus. 2003:40.

Pustejovsky, J., Castao, J., Ingria, R., Saur, R., Gaizauskas, R., Setzer, A., and Katz, G. (2003a). TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Fifth International Workshop on Computational Semantics (IWCS-5).*

Soberon, G., Aroyo, L., Welty, C., Inel, O., Overmeen, M., and Lin, H. (2013). Content and Behaviour Based Metrics for Crowd Truth. In *CrowdSem.*

Sprugnoli, R. and Lenci, A. (2014). Crowdsourcing for the Identification of Event Nominals: an Experiment. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, June.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations.*

Verhagen, M., Sauri, R., Caselli, T., and Pustejovsky, J. (2010). SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th international workshop on semantic evaluation.*