Sourcing the Crowd for a Few Good Ones: Event Type Detection

Tommaso CASELLI[1] *Chu−Ren HUANG*[2]
(1) TrentoRise, Via Sommarive 18, Povo I-38123, Povo (TN) Italy
(2) Dept. of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hung Hom, Hong-Kong SAR
t.caselli@trentorise.eu, churen.huang@polyu.edu.hk

ABSTRACT
This paper reports a crowdsourcing experiment on the identification and classification of event types in Italian. The data collected show that the task is not trivial (360 trusted judgments collected *vs*. 475 untrsuted ones) but it has been shown to be linguistically felicitous. The overall accuracy of the annotation is 61.6%. A reliability threshold assigned to the workers allows us to indentify the sub-population who has the awareness to perform this complex task and the accuracy of this sub-population is raised to 93%. Our hypothesis is that although the initial crowdsourced data is necessarily noisy, it can yield high quality results if the sub-population of 'good' workers can be identified. In other words, crowdsourcing offers a solution to difficult annotation tasks as long as there is an effective way to identify the reliable workers.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE, L_2 (OPTIONAL, AND ON SAME PAGE)

Identificare Annotatori Affidabili: Riconoscimento di Tipi di Evento

Questo articolo descrive un esperimento di *crowdsourcing* per il riconoscimento e la classificazione dei tipi di evento in Italiano. I dati raccolti mostrano che il compito non è banale (360 giudizi affidabili *vs*. 475 giudizi non affidabili), ma dimostra di essere linguisticamente "felice". L'accuratezza globale della annotazione è del 61,6%. Una soglia di affidabilità assegnata ai lavoratori ci permette di identificare la sotto-popolazione che ha la consapevolezza di svolgere questo compito complesso la cui accuratezza arriva fino al 93%. La nostra ipotesi è che, sebbene i dati iniziali ottenuti tramite tecniche di *crowdsourcing* siano necessariamente rumorosi, dei risultati di buona qualità possono essere ottenuti se la sotto-popolazione di "buoni" lavoratori è identificabile. In altre parole, il *crowdsourcing* offre una soluzione per compiti di annotazione difficili finché vi è un modo efficace per identificare i lavoratori affidabili.

KEYWORDS: crowdsourcing, semantic annotation, event types, quality assessment.

KEYWORDS IN L_2: *crowdsourcing*, annotazione semantica, tipi di evento, valutazione della qualità.

Proceedings of COLING 2012: Posters, pages 1239–1248,
COLING 2012, Mumbai, December 2012.

1 Introduction

Many Natural Language Processing (NLP) systems are based on supervised learning approaches relying on large amounts of manually annotated training data collected by domain experts. Such annotation process is highly expensive both in terms of money and time. However, the absence of manually annotated Language Resources (LRs) makes supervised NLP systems subject to the so-called knowledge acquisition bottleneck. In recent years, in order to facilitate the development of LRs, two different approaches have been tackled. The first aims at automatically acquiring LRs, such as lexica, from large corpus data (Briscoe and Carroll, 1997; Korhonen et al., 2006, among others). The second investigates the exploitation of the Web 2.0 through the use of crowdsourcing techniques, i.e. by using non-expert annotators recruited on the Web. The crucial motivation of crowdsourcing is that when a simple linguistic task is performed by a population much larger than the sampling allowable by traditional experiments, interesting and hitherto unobserved distributional properties of human behaviors may emerge. In addition to this, for Language Technology, the additional motivation is that a web-based crowd can provide data for the construction of large-scale LRs in a faster, cheaper and still reliable way.
So far, annotation works conducted by means of crowsourcing techniques have focused on rather simple linguistic tasks, such as the evaluation of automatic translations (Callison-Burch, 2009), word sense disambiguation (Snow et al., 2008; Akkaya et al., 2010, Rumshinsky, 2011), textual entailment (Snow et.al., 2008; Wang and Callison-Burch, 2010), commonsense knowledge (Gordon et al., 2010), text alignment for machine translations (Ambati and Vogel, 2010) and speech transcriptions (Callison-Burch and Dredze, 2010) among others. Such choices are in line with the idea of using the "wisdom of the crowd" as the tasks can be simplified and presented to the workers as a sort of online game such that a large percentage of the population can be expected to perform the task reliably.
In this work we explore the untapped strength of crowdsourcing when the linguistic task is a complex and challenging one, trying to understand "how far can go the crowd?". As mentioned, the received wisdom is that when the tasks are complex, crowdsourced data may be too noisy to use. However, the noise may come in two ways. One possibility is that the data is noisy across the board. The other possibility is that the data is noisy from those who are not able to perform these tasks well but clean from those who perform well. The latter scenario seems promising since we learn from our experience that regardless of how difficult a task is, there will be someone good at it if a big enough population is searched. In other words, by sheer size, crowdsourcing should in principle be able to provide good quality data for more complex tasks difficult to obtain otherwise. The challenge is to separate the reliable crowd from the unreliable one.
In this paper, we study the complex task of event type classification and detection. The remainder of this work will be structured as follows: in Section 2 we will report the theoretical framework we have adopted for the analysis of the event type. In Section 3 the task of event type classification through crowdsourcing techniques will be described. Section 4 analyzes and comments on the results obtained. Finally, we reports on the conclusions and future work.

2 Event Types: theoretical background

The event type, lexical aspect or *aktionsaart*, is a lexical category and represents the intrinsic temporal structure associated with eventualities. Though strictly interconnected, the notion of event type is not to be confused with that of (viewpoint) aspect, which, on the other hand, is a grammatical category and contributes to the description of an eventuality as being bounded or unbounded.

The event type is commonly associated with verbs since the range of linguistic tests elaborated so far in literature are based on syntactic criteria with the aim of identifying homogeneous classes. As Moens (1987) points out, what is needed as a starting point in an aspectual classification of verbs are tests based on co-occurrence possibilities of the verb with certain adverbial expressions or with the progressive and perfect aspect. However, we want to depart from this perspective, and we claim that the event type applies to all eventualities, independently of their linguistic realizations. This means that event nouns, like "*assemblea*" [meeting], can be associated with a specific event type value.
Vendler's (1967) seminal work proposed four main classes of event types, namely *states*, *activities*, *accomplishments* and *achievements*. Each of these classes can be described in terms of three basic sematic features such as [+/- homogeneous], [+/- durative] and [+/- dynamic]. For clarity's sake, one example per class is provided below.

1 The door is closed [*state*];

2 John ran. [*activity*]

3 John closed the door [*accomplishment*]

4 John died [*achievement*]

In this work we depart from the original approach proposed by Vendler and adopt a different theoretical backgound following Pustejovsky's proposals (1991; 1995). With respect to previous studies based on semantic primitives (Vendler, 1967; Dowty, 1970; Lakoff, 1970 among others), the theoretical model adopted assumes:

- the existence of a complex subeventual structure for predicates which provides a template for verbal decomposition and lexical semantics;
- that adverbial modification is described in terms of scope assignement on the event structure; and
- that semantic arguments within an event structure expression can be mapped on argument structure in a predictable and systematic way.

Vendler's classes are thus reorganized from four to three basic event type values, namely: *state*, *process* and *transition* and defined as follows:

- State: a single event which is evaluated relative to no other event (Example 1);
- Process: a sequence of events which identify the same semantic expression (Example 2);
- Transition: an event which identifies a semantic expression which can be evaluated only relative to its opponent (Examples 3 and 4).

For the current study, we will not enter into the details of the phenomenon of event composition, which accounts for the interaction of the basic event types with syntactic constituents and grammatical categories to form derived event representations (e.g.: the fact that a transition event occurring at the progressive viewpoint is to be re-classified as a process event).

3 Crowdsourcing the identification of event type in context

Our goal is the identification and classification of the actual event types of predicates. In this work, we concentrated on verbs, but we are aiming at extending the work to all predicative elements, including nominals and adjectives.
Recognizing the event type of a verb in context is not a trivial task (Klavans and Chodorow 1992). Recently, Zarcone and Lenci (2009) have conducted an experiment on the identification and classification of verb event types in Italian by using three expert annotators. They report results on classification accuracy ranging between 44% to 73%.
As for our experiment we collected a subset of 100 sentences from a 2,000 sentence corpus automatically extracted from La Repubblica (Baroni et al., 2004), a large corpus of Italian newspaper articles containing more that 300 million tokens. The 2,000 sentence corpus has been created by selecting the 20 most frequent verbs in the corpus La Repubblica which satisfy the following criteria:

- they must belong to WordNet semantic class of *motion* or *change*; and
- they must belong to at least one of the following semantic types in the SIMPLE/CLIPS Ontology (Ruimy et al., 2003): *change of location*, *move*, *cause change of location* and *cause motion*.

For each verb a set of 100 random sentences has been collected. The verbs are: ARRIVARE [arrive], TORNARE [come back], PASSARE [pass/go], ENTRARE [enter], USCIRE [exit/leave], SEGUIRE [follow], CORRERE [run], INCONTRARE [meet], SALIRE [climb/rise/go up], MUOVERE [move/raise], TIRARE [throw/pull], PARTIRE [leave/go/depart], SUPERARE [overcome/get over], CADERE [fall], GIRARE [turn/spin/rotate], ALZARE [raise/get up/turn up], SALTARE [jump], VIAGGIARE [travel], CONDURRE [lead], PROCEDERE [proceed/go on].
The subcorpus of 100 sentences was uploaded on the Crowdflower platform (CF[1]) with the task name "Classify the verbs". Following the basic philosophy of crowdsourcing, we have tried to keep the annotation task for the workers as simple as possible. Thus, we have simplified the definitions of Pustejovsky's basic event types in a way that the workers could easily understand them. The participants were asked to "classify the verbs according to their meaning". In particular, we have focused the explanation of the task on the idea that each verb meaning could be grouped into a class which corresponds to one of Pustejovsky's event type. The annotators were presented with the following definitions:

- State: the verb describes a condition of something or someone;
- Process: the verb describes/reports that a certain action has taken place, is taking place or will take place;
- Transition: the verb describes/reports that a certain action has taken place or will take place and as a consequence of the occurrence of this action there has been a change of state in the world.

The definitions were accompanied by a number of examples which aimed at clarifying the task. For each example we provided a paraphrasis of the verb meaning which tried to match the

[1] http://crowdflower.com/

event type definition and the associated event type. For clarity's sake we report one example of the instruction below. The verb which the workers have to assign the event type class is in bold.

5 Marco **arrivo'** al negozio.
[Marco arrived at the shop.]
Verb meaning: Marco has moved form a place to another and now he's at the shop.
Event type: TRANSITION

The experiment was set along the following parameters: a.) each worker could analyze a maximum of 20 sentences; and b.) each sentence could receive a maximum of 10 judgments from the workers. As for this latter aspect, we considered 10 judgments per sentence as a good top threshold for validating the annotation quality of the final answers following Snow et al. (2008)'s analysis.

3.1 Quality control: Gold Standard and worker recruitment

One of the central issues in crowdsourcing is the quality control of the data. In order to filter non reliable workers and possible spammers, we adopted two strategies. The first strategy exploits the "Gold Standard" functionality of the CF platform. 15 random sentences were annotated by an expert with respect to their event type. The Gold Standard will help us in assuring that the worker' answers are correct with respect to the instructions. The second strategy is to rely on altruism instead of monetary reward in recruiting to discourage spammers. For this task, we did not offer any compensation and recruited our workers by means of a campaign on social networks such as Facebook and Twitter.

On the basis of the answers to the Gold Standard, each worker receives a reliability score. This reliability score is useful for evaluating the annotation of subsequent data, i.e. non Gold Standard items, since it allows to filter out those instances with low values, thus excluding them from the final data set.

4 Evaluation

Our purpose in the evaluation is twofold: on the one hand, we are interested in determining if crowdsourcing can be used to obtain high quality information for complex semantic tasks or if there is a limit over which expert annotation is required, and, on the other hand, we are interested in understanding what is the level of awareness of average speakers when involved in the identification of complex linguistic phenomena like event types.

4.1 Reliability of the crowd

In Table 1 we report the aggregated results. The experiment was available on the Web through CF for a period of two weeks starting on Feb. 29th this year. 46 people took part in the experiment providing a total of 835 judgements. Each sentence received at least one judgement.

The first result is the difference between trusted and untrusted judgments. By computing the judgement percentage per judgement, more than 56% (475 out of 835) of the judgments expressed have been considered as not reliable according to the Gold Standard filter, thus providing a first cue on the complexity of this task. On the other hand, it is interesting to notice

Analytics	Results
number of judgments	835
number of trusted judgments	360
number of untrusted judgments	475
judgments on gold standard data	67
average trusted judgments per sentence	3.82
number of participants	46
overall accuracy	61.6%
overall accuracy of gold standard	53%
accuracy of gold of trusted workers	93%

Table 1: Overall breakdown of the experiment.

that: a.) the accuracy of the trusted workers on the Gold Standard data is surprisingly high (93%); and b.) the overall accuracy is 61.6%, which qualify the data as reliable, although noisy. These figures allow a first important generalization: although the task is complex and the possibility of reducing its complexity are limited due to the task itself (i.e. event type detection), it is still doable and it is possible to identify a relatively high number of reliable workers. Further data in support of this analysis can be obtained by observing the distribution of the selected verbs among the three classes. Provided the verbs'characteristics, the classes of *Transition* and *Process* are by far the most selected event types (48 and 42 assignments out of 100 sentences, respectively), while the *State* class is very low (only 10 assignements of of 100).
As a pre-test for determining the worker's qualification, an initial reliability score of 1.0 is assigned to each worker and it is reduced by 0.25 for each wrong answer to the Gold Standard items. The final reliable judgments provided by the CF platform can be grouped along four main clusters on the basis of this score. Table 2 reports the figures.

Group	# sentences	Reliability score
Cluster 1	43	1.0
Cluster 2	24	0.95 -0.7
Cluster 3	21	0.67 - 0.52
Cluster 4	11	0.5 - 0.33

Table 2: Reliability clusters of the trusted judgments.

A manual analysis of the data has shown that there is no error in the assignment of the event type for the items belonging to the first two clusters, i.e. reliability ranging from 1.0 to 0.7. On the other hand, in the last two clusters, i.e. reliability ranging from 0.67 to 0.33, we have identified 11 wrong answers. The distribution of the mistakes appears to be balanced between the two groups as there are 5 mistakes in Cluster 3 and 6 in Cluster 4. Nevertheless, by observing the corresponding percentages, it clearly appears that the items in the last group, Cluster 4, are those with the highest error rate and, thus, the least reliable (54.5% error rate in Cluster 4 *vs*. 23.8% error rate in Cluster 3). This suggests that the assignment of the event type cannot be determined only on the basis of a majority voting of the reliable workers and that not all the data provided by the workers for this specific task can be used as they are. Although the CF system assigns the event types to non Gold Standard items on the basis of a majority vote among the judgments of the trusted workers, the reliability score plays a much more important role in identifying those clusters of data which are problematic. As a consequence for

the development of LRs for complex linguistic information, such as the identification of event types, the results of this experiment provide some insights. The first is that, in principle, no linguistic task is too complex to be performed by non-experts, even though the amount of noisy data is expected to be higher than for easy tasks. In addition to this, reliability scores are more important that majority voting thus providing support to the development of well-balanced but small Gold Standards whose main purpose is the identification of those clusters of data which are more "prone" to contain errors and for which expert post-processing is required. As for our data, we propose to to set the reliability threshold to 0.7.
Finally, it appears that the correct class can be identified with a minimum of three/four judgments from reliable workers, as reported in Table 1 where the average number of trusted judgments per sentence is 3.82.

4.2 Awareness of the crowd

On the basis of the results, we can perform a further analysis on the awareness of the average speakers on the phenomenon of event type **identification and classification**. The analysis we report in this section is preliminary, although in line with what described in Zarcone and Lenci (2009). Although, average speakers seem to understand the notion of event type, the identification and classification of this property in the actual linguistic context is not trivial. As already stated, the fact that we have collected more untrusted judgements than trusted ones is a direct proof of this fact.
A further element of analysis on this aspect is provided by the the agreement on the correct class (i.e. majority voting). We have restricted the analysis to the Gold Standard items. The figures ranges between 43% to 88%. It is interesting to observe that the highest percentages of agreement are on those cases which express in a more clearcut way the event type. When facing more complex cases, including also instances of event type shifting, the percentages tend to split on all three possible classes with small differences.

Finally, it is interesting to observe that the results we have obtained are in line with those of Zarcone and Lenci (2009). As already stated, Zarcone and Lenci (2009) obtained an agreement on event type identification and classification ranging from 44% to 73%. In our experiment we have obtained an agreement per class ranging from 43% to 88%. One of the most interesting aspect is that they have used three expert annotators while we have used naive ones. These data support our conclusions on the awareness of the speaker with respect to the event types.

Conclusion and future work

This paper has explored the possibility of using crowdsourcing techniques to collect data for the identification and classification of event types in context. The most characteristic feature of this work with respect to previous studies is the difficulty of the task which is proposed to the non-expert annotators through a crowdsourcing platform.
The results collected provide empirical support to the claim that the identification and classification of event type is not easy (360 trusted judgments *vs*. 475 untrusted judgments) but, at the same time, it suggests that crowdsourcing techniques can be applied also to collect complex semantic information. As a matter of fact, we have obtained an overall accuracy of 61.6% which can be considered a good threshold for such a complex semantic task, with a top accuracy of 93% on Gold Standard data from trusted workers.
The data collected cannot be used as they are but require an expert post-processing analysis. However, the expert post-processing can be reduced to a subset of the data, in particular to those

which are below a certain reliability threshold. As for the event type identification, we claim that such a reliabiliy threshold can be put at 0.7. In this way, the development of annotated corpora both for testing and training can be facilitated with useful results in terms of reducing the efforts and costs for the creation of new Language Resources.
As for the issue of quality control, we have exploited the use of Gold Standard data and recruited motivated workers by means of a campaign on social platforms such as Facebook and Twitter. This latter element has proved important in avoiding the presence of spammers. As for the data collected, the combination of majority voting and reliability scores has proven useful for the identification both of reliable workers and correct data. However, the identification of the reliable crowd is still an open issue (see Ipeirotis et al., 2010) and better mechanisms of crowd selection should be integrated into existing (and new) crowdsourcing platforms. The solution we have adopted is partial though it proved to be efficient.
Finally, it is interesting to notice that average speakers are aware of the notion of event type, but as the results prove, they have problems to project the event type category on the actual context of occurrence.
In order to get better results in terms of quality and quantity, we are planning to further exploit the Gold Standard to identify the subset(s) of participants who is good at the sub-tasks of annotating each event type separately (i.e. state, activity, and transition respectively). This may even include workers whose reliability is below the threshold for the whole task (i.e. identify the three event types), but, on the contrary, is (almost) perfect on the sub-tasks. Moreover, we will extend this experiment with data from other languages such as English and Chinese to provide further support to our observations and, most importantly, to the reliability threshold. Finally, we aim at using the collected data for testing a classifier of event types in context. This will be the first step of a more complex task involving the identification of event internal structures (Im and Pustejovsky, 2009; 2010), which will contribute to the development of a new lexicon on events for complex NLP systems such as Question Answering and Recognizing Textual Entailment.

Acknowledgments

This research was supported in part by the Erasmus Mundus Action 2 program MULTI of the European Union, grant agreement number 2009-5259-5.

References

Akkaya, C., Alexander, C., Janyce, W., and Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Akkaya, C., Janyce, W., and Rada, M. (2009). Subjectivity word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.

Ambati, V. and Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Bertinetto, P. and Delfitto, D. (1996). Aspect vs. actionality. In Bertinetto, P., editor, *Il dominio tempo-aspettuale*, Torino. Rosebnerg Sellier.

Bertinetto, P. M. (1986). *Tempo, Aspetto e Azione nel verbo italiano. Il sistema dell'indicativo*. Accademia della Crusca., Firenze.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Dowty, D. (1979). *Word Meaning and Montague Grammar*. D. Reidel,, Dordrecht, W. Germany.

Im, S. and Pustejovsky, J. (2009). Annotating event implicatures for textual inference tasks. In *Proceedings of G.L. 2009*.

Im, S. and Pustejovsky, J. (2010). Annotating lexically entailed subevents for textual inference tasks. In *Proceedings of FLAIRS-2010*.

Ipeirotis, P., Provost, F., and Wang, J. (2010). Quality management on amazon mechanical turk. In *Proceedings of KDD-HCOMP '10*.

Jonathan, G., Benjamin, V. D., and Schubert, L. (2010). Evaluation of commonsense knowledge with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Klavans, J. and Chodorow, L. (1992). Degrees of stativity: The lexical representation of verb aspect. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*.

Korhonen, A., Krimolowsky, Y., and Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th International Language Resources and Evaluation (LREC'06)*.

Lakoff, G. (1970). *Irregularity in syntax*. Holt, Rinchart and Winston, New York.

Moens, M. (1987). *Tense, Aspect and Temporal Reference*. Centre for Cognitive Science, University of Edinburgh, Edinburgh, Scotland.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4):409–441.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press.

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Fiorentino, M. D., Ulivieri, M., and Rossi, S. (2003). A computational semantic lexicon of italian: Simple. *Linguistica Computazionale*, XVIII-XIX:821–864.

Rumshisky, A. (2011). Crowdsourcing word sense definition. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*.

Snow, R., Connor, B. O., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fastut is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*.

Vendler, Z. (1995). *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *ACM Conference on Human Factors in Computing Systems, CHI 2004*.

Wang, R. and Callison-Burch, C. (2010). Cheap facts and counter-facts. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Zarcone, A. and Lenci, A. (2006). Un modello stocastico della classificazione azionale. In G., F., Benatti, R., and Mosca, M., editors, *Linguistica e modelli tecnologici della ricerca. Atti del XL Congresso SLI - Vercelli, Settembre 2006*.

Zarcone, A. and Lenci, A. (2008). Computational models of event type classification in context. In *Proceedings of the 6th International Language Resources and Evaluation (LREC8)*.