Domain-Independent Quality Measures for Crowd Truth Disagreement

Oana Inel[13], Lora Aroyo[1], Chris Welty[2], and Robert-Jan Sips[3]

[1] VU University Amsterdam
oana.inel@vu.nl, lora.aroyo@vu.nl
[2] IBM Watson Research Center, New York
cawelty@gmail.com
[3] CAS Benelux, IBM Netherlands
robert-jan.sips@nl.ibm.com

Abstract. Using crowdsourcing platforms such as CrowdFlower and Amazon Mechanical Turk for gathering human annotation data has become now a mainstream process. Such crowd involvement can reduce the time needed for solving an annotation task and with the large number of annotators can be a valuable source of annotation diversity. In order to harness this diversity across domains it is critical to establish a common ground for quality assessment of the results. In this paper we report our experiences for optimizing and adapting crowdsourcing micro-tasks across domains considering three aspects: (1) the micro-task template, (2) the quality measurements for the workers judgments and (3) the overall annotation workflow. We performed experiments in two domains, i.e. events extraction (MRP project) and medical relations extraction (Crowd-Watson project). The results confirm our main hypothesis that some aspects of the evaluation metrics can be defined in a domain-independent way for micro-tasks that assess the parameters to harness the diversity of annotations and the useful disagreement between workers. This paper focuses specifically on the parameters relevant for the 'event extraction' ground-truth data collection and demonstrates their reusability from the medical domain.

Keywords: Crowdsourcing, Ground-Truth, Event Extraction, Relation Extraction, NLP, Newspaper corpus

1 Introduction

At the basis for machine learning and information retrieval systems is the collection of ground truth data. Typically, creating such a gold standard dataset requires domain expert annotations to ensure high quality of the training and evaluation data. However, expert-annotation may result in limitedly annotated datasets, which do not capture the evolution of human expressions and the diversity in their interpretations. With its large pool of human workers, crowdsourcing became a mainstream source for higher volume and continuous collection of training and evaluation data (specifically for tasks that do not require domain expertise). Thus, the new challenge became to correctly and efficiently identifying low quality or spam contributions of the micro-workers. Research shows that micro-workers' behavior (e.g. either as intentional spam or low quality contributions) can influence the overall quality of the final results [1]. Typically, the

quality is measured under the assumption that there is only one right answer for each micro-task and that it can be measured through annotators agreement [2].

Recently, however, there is evidence to support the hypothesis that harnessing diversity and disagreement between workers can improve the ground truth data [3]. Thus, it is critical to identify how much of the crowdsourced data is part of spam, low quality or actual meaningful disagreement between workers. There is an extensive body of research on spam detection through, e.g. majority decision [4], the expectation maximization [5]. Additionally, the micro-task template can impact the ability of the workers to complete the task successfully [6]. However, most of the studies have been focussing on addressing these issues as individual processes and less as part of a complete end-to-end workflow [7]. In this paper, we show that an optimal annotation workflow, which supports (1) apriori filtering of input data to maximize suitability for the workers and for the training, (2) crafting the templates to ensure proper disagreement collection and (3) defining appropriate metrics for low quality and spam filtering can impact beneficially the quality of the ground truth data, which we call *Crowd Truth* [8].

We conducted experiments in two domains - starting with **medical relation extraction** in the context of Crowd-Watson project[4] and adapting the experiences to **event extraction** in the context of DARPA's Machine Reading program (MRP)[5]. We used the same workflow in both domains: (1) *pre-processing of input data and micro-task template design*, (2) *data collection through automatic sequencing of micro-task batches*, (3) *disagreement analytics through quality metrics on workers judgments* and (4) *post-processing of the results for spam filtering and micro-task template adaptation*. The novel contribution of this work is twofold - on the one hand demonstrating a crowd truth collection workflow optimized for multiple domains; and on the other hand providing reusable disagreement-harnessing micro-task templates with the corresponding spam detection disagreement metrics.

The rest of the paper is organized as follows. Section 2 places this work in the context of crowdsourcing, evaluation metrics and event extraction. Section 3 presents the Crowd-Watson workflow and shows its adaptation for the event extraction task. Section 4 presents the experimental setup and Section 5 discusses the results. Section 6 draws the conclusions and presents the future work.

2 Related Work

The amount of knowledge that crowdsourcing platforms like CrowdFlower[6] or Amazon Mechanical Turk[7] hold fostered a great advancement in human computation [9]. Although the existing paid platforms manage to ease the human computation, it has been argued that their utility as a general-purpose computation platform still needs improvement [10]. Since the development of crowdsourcing has become more intensive, much research has been done in combining human and machine capabilities in order to obtain an automation of the crowdsourced process. Some state-of-the-art crowdsourcing frameworks are CrowdLang [10]

[4] https://github.com/laroyo/watsonc
[5] http://www.darpa.mil/OurWork/I2O/Programs/MachineReading.aspx
[6] https://crowdflower.com/
[7] https://www.mturk.com/mturk/

and CrowdMap [11]. However, CrowdLang restricts the users to work with its own internal programming language and CrowdMap solves only ontology alignment. Thus, both frameworks can be hardly adapted to another domain.

A lot of research has been focused on indentifying crowdsourced spam. Although a commonly used algorithm for removing spam workers is the majority decision [4], according to [12] it is not an optimal approach as it assumes all the workers to be equally good. Alternatively, expectation maximization [13] estimates individual error rates of workers. First, it infers the correct answer for each unit and then compares each worker answer to the one inferred to be correct. However, [14] shows that some tasks can have multiple good answers, while most spam or low quality workers typically select multiple answers. For this type of problem, some disagreement metrics [15] have been developed, based on workers annotations (e.g. agreement on the same unit, agreement over all the units) and their behavior (e.g. repetitive answers, number of annotations).

Research on events detection and extraction from medical texts [16], [17] is primarily focussed on improving the machine performance for it. In [16] the authors create an event trigger dictionary based on the original GENIA event corpus [18] and further, they apply dependency graphs for parsing the input corpus and extracting the putative events. [17] uses the Stanford Lexical Parser[8] for producing dependency graphs of the input corpus, as well as extracting the putative events. However, instead of using a dictionary for medical events, they only use the relations given by the dependency graphs.

Although there has been an extensive event extraction research using machines, the advantages of using crowdsourcing in this domain were not fully harnessed. Our new approach (fostering disagreement between annotators) [3] asks the crowd to judge the putative events and to provide event role-fillers at different granularities. The concept of harnessing disagreement in Natural Language Processing is not yet considered a mainstream process. In [19] disagreement is used as a trigger for consensus-based annotation in which all disagreeing annotators are forced to discuss and arrive at a consensus. This approach achieves κ scores above .9, but it is not clear if the forced consensus achieves anything meaningful. It is also not clear if this is practical in a crowdsourcing environment.

3 Adapting Crowd-Watson for Event Extraction

This section presents the workflow initially developed within the Crowd-Watson project (Figure 1) for creating ground truth data for medical relation extraction, that was further adapted for creating ground truth for newspaper events extraction. The resulting ground truth we refer to as *Crowd Truth*. In this paper we focus on the event extraction process for event crowd truth collection and its adaptation from the medical domain. A key point here is illustrating the reusability and optimization features of the workflow across the two domains.

The framework is designed as an end-to-end process which provides feedback loops that generate analysis for each stage of the workflow in order to improve future results. The *Pre-Processing* 3.1 component handles the adaptation of the input data for making it solving-affordable in terms of micro-tasks.

[8] http://nlp.stanford.edu/software/lex-parser.shtml

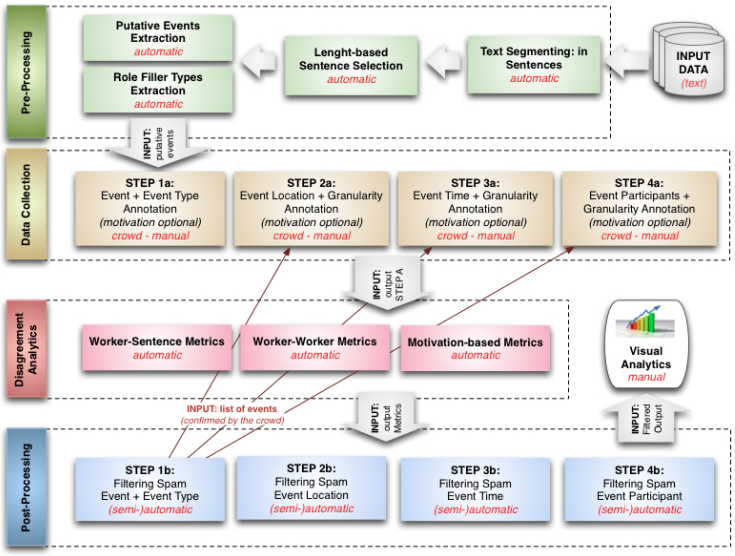

Fig. 1: Crowd-Watson Framework: Event Extraction Workflow Design

The *Data Collection* 3.2 uses CrowdFlower sequences of jobs for collecting judgments, while the *Disagreement Analysis* 3.3 component automatically handles the contributors evaluation. The *Post-Processing* 3.4 component automatically filters out the workers identified as spammers. Further, the process of collecting disagreement-based judgments can continue by reiterating each mentioned step.

3.1 Pre-Processing for Event Extraction

As, typically, the initial textual data collected from large sources, e.g. Wikipedia, newspapers first needs to be processed into smaller chunks suitable for micro-tasks (paragraphs, sentences). Further, to optimize its applicability for training, sentences that are not useful for training, e.g. too long or too short or contain specific words that increase the ambiguity need to be filtered out. The *Input Data Filtering* component clusters the input sentences based on their syntactic criteria, e.g. presence of semicolons, comma-separated lists, parentheses, etc. Each of those clusters can be either ignored or used for a specific micro-tasks. For example, sentences with putative events identified in them can be given to the crowd to confirm whether they refer to an event or not. Majority of those filters we directly reused from our medical relation extraction use case.

Input Data: For the experiments described in Section 4 we used articles from The New York Times. After their content was split into sentences (50 initial sentences), we removed the short sentences (less than 15 words). Compared with the task of medical relation extraction where the long sentences are typically difficult for the crowd, in the task of event extraction the longer the sentence the higher the chance that it will contain useful context for the event and the role fillers. This left us with 37 sentences to run the experiments with.

Putative Event Extraction: The first step in extracting events is to determine the *putative events* (verbs and nominalized verbs), i.e. word phrase that could

possibly indicate an event. This component first exploits the *context-free phrase structure grammar* representation from the Stanford Parser to extract all the verbs and the nouns. Further, it follows the *typed dependencies parses* (also from the Stanford Parser) to extract word phrases that being in relation with certain verbs might trigger events. In addition to the Stanford Parser we also used NomLex[9], a dictionary of nominalizations. Thus, we extracted 205 putative events from the 37 sentences of the input data. For the crowdsourced experiments we selected only 70 putative events. Table 1 presents the putative events dataset.

Table 1: Putative Events Overview

Category	Putative Events	
	Article 1	Article 2
VB, VBD, VBG, VBN, VBZ, VBP[10]	61	57
Phrasal Verb	3	2
Verb + Direct Object	21	18
Predicate + Infinitive Verb	9	9
Adjectival Complement	2	1
Nominalized Verb	10	11
Nominalized Verb + Preposition "of"	2	0
Total: 205		

Micro-Task Template Settings In order to collect maximum diversity of answers from workers, and thus explore the disagreement space, we focus on the design of specific micro-task templates. Here again, the initial template settings were adapted from the medical relation extraction templates [8] and [14]. For the event extraction template we use a sentence with one putative event capitalized. Each template is based on conditional statements ("if clause"), which lead the worker through the template parts (see Figure 2).
Event annotation: Judge whether the capitalized word phrase refers to an event and motivate the answer. If the answer is yes, choose the type of the event.
Event role fillers: Judge whether the selected word phrase refers to an event. If the answer is yes, highlight the words referring to the attribute and choose its type(s). For participants template there is a follow-up question to choose a second participant. By allowing the worker to highlight words directly in the text instead of retyping them (in the relation extraction task much of the low contributions came from this aspect) we aim to improve the annotations collected.

Role Fillers Taxonomies: Providing role filler selection ranges is demanding to form the annotator disagreement space. Thus, we align events, their types and role fillers to a set of simplified existing ontologies (to increase workers efficiency). For the *event type taxonomy* we used the semantic parser Semafor (Semantic Analysis of Frame Representations) [20] which identifies the frames of FrameNet 1.5[11] evoked by each English word. We set up Semafor with "strict" automatic target identification model, graph-based semi-supervised learning and AD^3 decoding type. The taxonomy includes a total of 12 top frames and grouped frames with similar semantics. The taxonomy for *event location* is based on GeoNames[12]. From each main GeoNames category we chose the most common-sense entities and various commonly used subclasses. Annotators often disagree

[9] http://nlp.cs.nyu.edu/nomlex/
[10] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[11] https://framenet.icsi.berkeley.edu/fndrupal/
[12] http://www.geonames.org/ontology/documentation.html

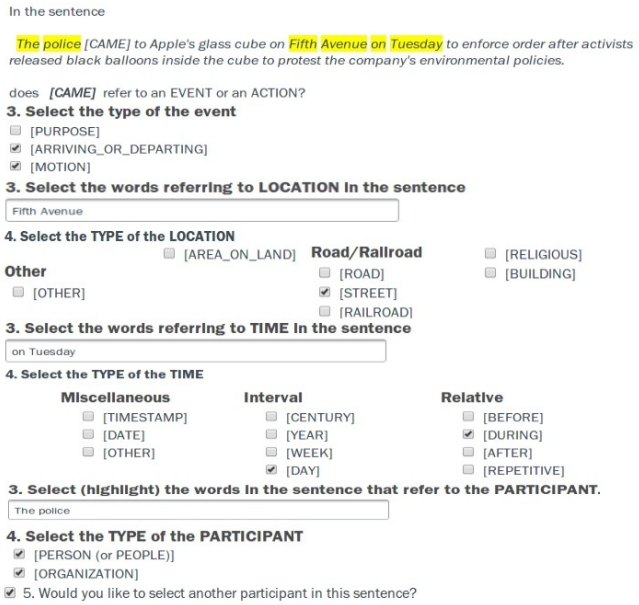

Fig. 2: Event Extraction Template Design

which is the relevant level of granularity for temporal expressions. However, when gathering gold standard data for events we are interested in collecting all possible temporal expression. Thus, we combined four relative classes from Allen's time theory [21] with two time points and five time intervals from from KSL time ontology [22]. According to [23] the proper nouns strongly relate to participants in events. Thus, our *participants taxonomy* considers 5 classes that can be mostly represented by proper nouns. To foster diversity and disagreement, we added to each taxonomy the value "Other". Table 2 presents each taxonomy.

Table 2: Event Role Fillers Taxonomies

Role Filler	Taxonomy
Event Type	Purpose, Arriving or Departing, Motion, Communication, Usage, Judgment, Leadership, Success or Failure, Sending or Receiving, Action, Attack, Political, Other.
Location Type	Geographical - Continent, Country, Region, City, State, Area on Land - Valley, Island, Mountain, Beach, Forest, Park, Area on Water - Ocean, River, Lake, Sea, Road/Railroad - Road, Street, Railroad, Tunnel, Building - Educational, Government, Residence, Commercial, Industrial, Military, Religious, Other
Time Type	Before, During, After, Repetitive, Timestamp, Date, Century, Year, Week, Day, Part of Day, Other
Participants Type	Person, Organization, Geographical Region, Nation, Object, Other

Target Crowd Settings component applies context-specific restrictions on contributors, i.e. origin country, native language. After these basic conditions are applied, the *Crowdsourcing Workflow Definition* element sets the actual flow of the micro-task, i.e. judgments per unit and per worker, channels, payment.

3.2 Data Collection for Event Extraction

The Crowd-Watson[13] workflow framework [24] is targeted towards a crowd of lay workers and is developed as a micro-task platform on top of the crowdsourcing platform CrowdFlower. Additionally, Crowd-Watson supports also a gaming

[13] http://crowd-watson.nl

crowdsourcing platform[14], which targets nichesourcing with medical experts [25]. Crowd-Watson is specifically designed to stimulate the capture of disagreement and diversity in the annotator contributions. Figure 1 shows the specification of the components for event extraction.

3.3 Events Disagreement Analytics

This component assesses the quality of the workers contributions by analyzing the disagreement with *Worker Metrics* - worker agreement on a specific unit or across all the units that (s)he solved, and *Content Metrics* - the overall quality of the training data. This provides additional characteristics of the crowd truth, e.g. sentence clarity, similarity and ambiguity measures. In the event extraction task, the sentence vector is defined for each event property as the content of the aforementioned taxonomies and the "Not Applicable" value. This value is automatically added when: (1) the word phrase selected does not refer to an event, (2) there is no event property mentioned in the text snippet.

To avoid penalizing workers for contributing on difficult or ambiguous sentences, we filter sentences based on their clarity score [14]. Only then we apply the content-based worker metrics. The *worker-sentence agreement* measures the agreement between a worker annotation and the annotations of the rest of the workers for the same sentence (i.e. the averaged cosines between the worker sentence vector and the aggregated sentence vector, subtracting the worker's vector). The *worker-worker agreement* measures the agreement of the individual judgments of one worker with the judgments of all the other workers (i.e. the aggregated pairwise confusion matrix between the worker and the rest of the workers weighted by the number of sentences in common). The *number of annotations per sentence* is the average number of different types used by a worker for annotating a set of sentences.

3.4 Post-Processing for Event Extraction

The resulting analysis from these metrics provides input to *Post-Processing* to filter spam workers and spam-prone worker channels. The *Worker Spam Filtering* controls this flow. The list of spam micro-workers is sent to the *Data Collection* component to ban them from contributing to future tasks. Some statistics are also performed at the level of channels through *Crowdsourcing Channel Filtering*. Feedback is also sent to the *Pre-Processing* for improving the selection of input data, the optimization of micro-task settings and the workflow.

Finally, the *Visual-Analytics* component provides interactive visualization of (1) the workers behavior, (2) the sentence clarity and similarity. It provides a clear way to observe the dynamics in workers disagreement, completion time and the distribution of filters for spam contributions. The same visualization is used both for the relation extraction and for the event extraction tasks (Figure 3).

4 Experimental Setup

We adapted the Crowd-Watson medical relation extraction template for event extraction by constraining the workers to follow stricter rules, so that we can compare: (1) how does the new template influence the quality of the results;

[14] http://crowd-watson.nl/dr-detective-game/

and (2) and how does it effect the behavioral filters for spam and low quality contributions. We performed one preliminary experiment Exp0 3 to assess the applicability of the relation extraction template for the purposes of event extraction and established this as the baseline. We conducted four experiments

Table 3: Experiments Overview

	# Sents	#Judgts per Sent	Channels	Max # Sents per Worker	#Judgts for Batch	# of Workers for batch	# Unique Workers
Exp0 Event+Type	35	15	crowdguru, zoombucks, vivatic, amt, prodege	10	525	66	66
Exp1 Event+Type	70	15	— " —	10	1050	147	141
Exp2 Event Location	70	15	— " —	10	1050	143	132
Exp3 Event Time	70	15	— " —	10	1050	146	140
Exp4 Event Participants	70	15	— " —	10	1050	141	137
Totals	*70*	*15*	— " —	*10*	*4200*	*643*	*436*
Totals (no singletons)	*70*	*15*	— " —	*10*	*4143*	*580*	*428*
Totals (no singletons, doubletons)	*70*	*15*	— " —	*10*	*4102*	*539*	*421*

(each two batches of 35 sentences for each event property, i.e 70 sentences), see Table 3. All experiments had the same settings, i.e. 15 judgments per sentence, 10 sentences allowed per worker, AMT, Vivatic, Prodege, Zoombucks and Crowdguru channels. This setting of two sequential runnings of small batches (of 35 sentences) allowed to:

- get enough judgments per sentence given there is no golden standard;
- optimize the time to finish a large sentence set
- get a quick run of the entire workflow.

However, as the split in small batches could allow spam workers to perform every batch, this could have exponential negative effect on the quality of the annotations. Thus, we optimized the spam filtering by:

- limiting the number of judgments per worker in a batch;
- applying spam filtering after each batch and blocking them from future jobs.

5 Results and Discussion

In this section we analyze the entire experimental workflow. We observe the effect of the template design on the accuracy of the crowdsourced data, and we measure the accuracy of the worker metrics compared to the accuracy of the worker metrics together with the explanation-based metrics.

The preliminary experiment (Exp0) for identifying events and event types did not use a conditional micro-task template. Nine workers submitted only one or two judgments, which did not provide evaluation relevance, and were thus excluded from the analysis. The contributions of 66 remaining workers were analyzed further. The worker metrics identified 15 spam contributors, while the explanation-based filters, described in [14] identified another 10 spam contributors. However, upon a manual evaluation of the results 5 more workers were identified with an erratic behavior, selecting either "no event" and a type different than "Not Applicable", or "yes event" and "Not Applicable" type. Such contributions could be because of intentional spamming, or negligence or misunderstanding of the task. This result guided us to a more restrictive template to improve the job outcome. The new event extraction template (Figure 2) did not allow the workers to choose simultaneously: (1) "Not Applicable" and other

event property type, (2) "no event" and a type different than "Not Applicable", (3) "yes event" and "Not Applicable" type.

When adapting the taxonomies we tried to conceive different experimental cases that could give insights in the adaptability degree of the metrics:

- the list of event types, time types and participant types are similar to the number of relations provided in the medical relation extraction task;
- the taxonomy for location type is more diverse, with overlapping concepts;
- one event can have multiple participants, which can increase the number of annotation for a putative event; this is a relevant factor for evaluating the behavior the average number of annotations metric;

We performed a manual sampling-based evaluation of workers in order to determine the accuracy of the spam metrics. We examined all the workers marked as spam by the filters, as well as the ones ranked as best workers. Some workers in the gray area inbetween were also manually evaluated. Figure 4 shows the precision, recall and F-measure only for the worker metrics for each job type, i.e. event type, event location, event time and event participants. Although the worker metrics identified a high percentage of low-quality contributors, the accuracy and the precision of the metrics still need improvements. A reason for this behavior could strongly relate to event properties types distribution and similarity, which varies along the four classes of event characteristics.

For the *Event Type* task the *average number of annotations* metric was able to identify correctly a high amount of spammers. However, both the worker-sentence and worker-worker agreement had low values in terms of correctness. The distribution of event types among 35 sentences, i.e. putative events (see Figure 3) indicates high ambiguity of the event types. One reason for this could be our choice of event types, which might not be so appropriate for the workers. However, in the *Location Type* task we see a different picture. As expected, most workers chose multiple types for one identified location. For example, the highlighted location for one putative event was "Apple's glass cube on Fifth Avenue" and was classified as: [COMMERCIAL], [BUILDING], [ROAD], [STREET]. Although the worker did agree to a certain extent with the other workers solving the same sentence, he was wrongly classified as spammer based on the number of annotations. Thus, the accuracy of spam identification for location type is solely based on the worker-worker and worker-sentence agreements metrics. The F-measure for event type is equal to 0.89, while for event location is equal to 0.82. Even though the percentages of identified low-quality contributors were comparable for both experiments, the lower number of correct predictions for event location stands as a reason for a lower event location type F-measure. For the *Time Type* task the assessment of metrics behavior was the most challenging. More than half of the sentences used in the experiment did not contain any event time reference, and most of the workers chose "[NOT APPLICABLE]" as a type. This resulted in a high worker-worker and worker-sentence agreement scores. Thus, if a worker would disagree with other workers on just one sentence, (s)he would be identified as spam. Most of the spammers, however, were

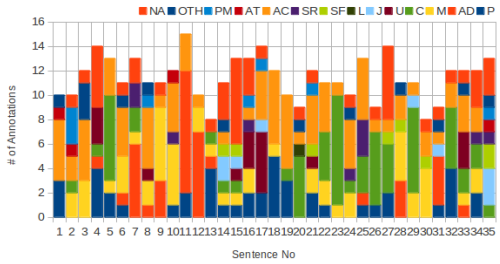

Fig. 3: Annotation Distribution - Event Types

captured by at least two worker metrics. The ones identified by less than two metrics justify the high recall and the low precision (Figure 4) (F=0.81).

One event could have multiple participants of different types. However, the highlighted participants mentions in the task were of the same type, which explains why the average number of annotations per sentence did not have an erratic behavior. The *event participants type* has the highest F-measure value (0.91). This value is a result of the high amount of spammers correctly identified as well as the high amount of spammers identified from the entire list of spammers. Thus, we can conclude that the participants taxonomy presented to the workers is concise and covers with high precision the possibilities of interpreting the participants of an event.

The high worker disagreement in the *event type* experiments gave the worker metrics an important boost of efficiency, by identifying a high amount of true spammers. However, the overall agreement was above mean expectations. Thus, the workers that did not highly agree with other workers were prone to be identified as spammers. For *event location*, however, the average number of annotations per putative event decreased the total precision of correctly identified spammers. As seen in Figure 4, the applied worker measurements had the most accurate behavior for the *event participants type* task.

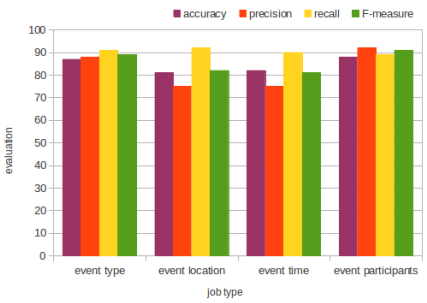

Fig. 4: Worker Metrics Evaluation

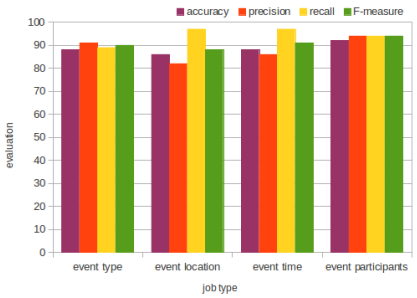

Fig. 5: Worker Metrics and Explanation-based Filters Evaluation

By looking at Figure 5 we can see that the explanations provided by the workers regarding their answers give an important boost of efficiency in identifying patterns that are associated with low-quality contributions and even spam contributions. In combination with the worker metrics, these explanation-based filters are able to increase the accuracy of detecting low-quality workers with at least 5%. This situation was possible because only a small number of workers

were identified as spammer by both explanation-based filters and disagreement-based metrics (worker metrics). Thus, for each batch, not more than 2 or 3 workers were identified by both quality measurements. Hence, it seems reasonable to further use the advantages brought up by those filters. This conclusion is also underlined by the usage of all the channels, situation that is usually associated with an increased percentage of spammed results. The results presented in Figure 5 make a good case to state that the usage of both worker metrics and explanation-based filters achieves high accuracy in terms of crowdsourced data. With the results mentioned in Figure 5 we can state that we succeeded to achieve high accuracy in identifying the spam workers, but we also showed how the metrics are suitable across domains.

6 Conclusions and Future Work

This paper presents the results of our experiments on estimating the reusability and domain-dependency of crowdsourcing workflow components, such as processing of input data, micro-task templates and result assessment metrics. We demonstrated how components defined in one domain (medical relation extraction) can be easily adapted to a completely different domain (event extraction from newspapers). Results from the experiments showed that some of the metrics for workers and content can be applied with high precision in those domains. For understanding to what extent the domains can be similarly treated, we conducted different research at each step of the crowdsourced process.

By directly reusing pre-processing filters from the medical relation extraction domain we showed that the input data can be optimized using syntactic text features. Thus, we can argue that the syntactic text features are mostly domain-independent. Although the template design was adapted for stimulating diversity in worker answers, the metrics were still able to capture the low quality contributions in both domains. With the final template design for event extraction, the workers were less prone to spam the results by mistake. We have showed that especially for domains were there is no golden data known in advance, the explanations can be successfully used to identify more spam or low-quality workers. When the explanation filters are combined with the disagreement worker metrics the accuracy of detecting those low-quality contributors reaches a value greater than 92%. To sum up, the adaptation of the disagreement analysis component from the medical relation extraction to the event extraction preserved its good outcomes, and thus, these disagreement metrics are domain-independent.

As part of this research, our future work should focus on solving ambiguity-related aspects. First, our analysis showed that there is still space for improving the event properties types. The event-type taxonomy shows a lot of ambiguity when looking at the workers annotations distribution. Further experiments should clarify whether a different classification of the putative events can achieve a better performance compared to the current experiments. Also, we need to conclude how the types that are overlapping influence the results. Furthermore, each word phrase highlighted from the sentences needs to be clustered in order to determine the most appropriate structure of the event role filler.

References

1. Vuurens, J., de Vries, A.P., Eickhoff, C.: How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In: Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIRâĂŹ11). (2011) 21–26
2. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on Multimedia information retrieval, ACM (2010) 557–566
3. Aroyo, L., Welty, C.: Harnessing disagreement for event semantics. Detection, Representation, and Exploitation of Events in the Semantic Web (2012) 31
4. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. In: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), IEEE (2011) 316–321
5. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: ACM SIGKDD workshop on human computation, ACM (2010) 64–67
6. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: SIGCHI conference on human factors in computing systems, ACM (2008)
7. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 conference on Computer supported cooperative work, ACM (2013) 1301–1318
8. Aroyo, L., Welty, C.: Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM (2013)
9. Quinn, A.J., Bederson, B.B.: Human computation: a survey and taxonomy of a growing field. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2011) 1403–1412
10. Minder, P., Bernstein, A.: Crowdlang-first steps towards programmable human computers for general computation. In: Human Computation. (2011)
11. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. In: The Semantic Web–ISWC 2012. Springer (2012) 525–541
12. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. The Journal of Machine Learning Research **99** (2010)
13. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the em algorithm. Applied Statistics (1979) 20–28
14. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Crowd truth metrics. Technical report, VU University Amsterdam (2013)
15. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. AAAI2013 Fall Symposium on Semantics for Big Data (in print) (2013)
16. Buyko, E., Faessler, E., Wermter, J., Hahn, U.: Event extraction from trimmed dependency graphs. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task, Association for Computational Linguistics (2009) 19–27
17. Kilicoglu, H., Bergler, S.: Syntactic dependency based heuristics for biological event extraction. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. (2009) 119–127
18. Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC bioinformatics **9**(1) (2008) 10
19. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: INTERSPEECH, Citeseer (2002)
20. Das, D., Schneider, N., Chen, D., Smith, N.A.: Semafor 1.0: A probabilistic frame-semantic parser. Language Technologies Institute, School of Computer Science, Carnegie Mellon University (2010)
21. Allen, J.F., Hayes, P.J.: A common-sense theory of time. Volume 85. (1985)
22. Zhou, Q., Fikes, R.: A reusable time ontology. Technical report, KSL-00-01, Stanford University (2000)
23. Hatzivassiloglou, V., Filatova, E.: Domain-independent detection, extraction, and labeling of atomic events, Proceedings of the RANLP Conference (2003)
24. Lin, H., Inel, O., Soberón, G., Aroyo, L., Welty, C., Overmeen, M., Sips, R.J.: Crowd watson: Crowdsourced text annotations. Technical report, VU University Amsterdam (2013)
25. Dumitrache, A., Aroyo, L., Welty, C., Sips, R.J., Levas, A.: Dr. detective: combining gamification techniques and crowdsourcing to create a gold standard for the medical domain. Technical report, VU University Amsterdam (2013)