Machine-Crowd Annotation Workflow for Event Understanding Across Collections and Domains

Oana Inel[1,2(✉)]

[1] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
oana.inel@vu.nl
[2] IBM Center for Advanced Studies Benelux, Amsterdam, The Netherlands
oana.inel@nl.ibm.com

Abstract. People need context to process the massive information online. Context is often expressed by a specific event taking place. The multitude of data streams used to mention events provide an inconceivable amount of information redundancy and perspectives. This poses challenges to both humans, *i.e.*, to reduce the information overload and consume the meaningful information and machines, *i.e.*, to generate a concise overview of the events. For machines to generate such overviews, they need to be taught to understand events. The goal of this research project is to investigate whether combining machines output with crowd perspectives boosts the event understanding of state-of-the-art natural language processing tools and improve their event detection. To answer this question, we propose an end-to-end research methodology for: machine processing, defining experimental data and setup, gathering event semantics and results evaluation. We present preliminary results that indicate crowdsourcing as a reliable approach for *(1)* linking events and their related entities in cultural heritage collections and *(2)* identifying salient event features (*i.e.*, relevant mentions and sentiments) for online data. We provide an evaluation plan for the overall research methodology of crowdsourcing event semantics across modalities and domains.

Keywords: Crowdsourcing · Event extraction · Machine-human computation · Information extraction · Event semantics annotation

1 Introduction/Motivation

With the progress on the Web, significant amounts of information are made available online. The information ranges from different data types such as tweets, news, cultural heritage and news archives and across various distribution channels such as traditional or social media. This poses a lot of challenges for search engines and information retrieval systems as they need *(1)* to extract meaningful information from any modality (*i.e.*, text, image, video) and *(2)* to synthesize streams from various channels in order to provide succinct pieces of information

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 813–823, 2016.
DOI: 10.1007/978-3-319-34129-3_50

that answer the end user needs. Thus, there is a challenge to interpret the information gain of each data stream, identify the meaningful pieces of information and generate a concise and complete summary of all the information requested.

Events are by definition complex entities, essential for querying, perceiving and consuming the meaning of the information we are surrounded by. We need to understand what an event is, how to describe an event and to what extent an event is useful for searching on a given topic. Usually, events create context by introducing related entities such as participants involved, locations where the event takes place or the time period when the event takes place. For everyday events, the event space is represented in the different data streams and channels. Hence, besides relevance, we need to extend the event understanding with salience, novelty features, thus, minimized redundancy, multitude of perspectives and subjective semantics such as sentiments and sentiment intensities.

The natural language processing (NLP) community recognizes the importance of events [1,2]. While the accuracy of the NLP tools for extracting named entities (NE) is continuously improving, their performance in detecting events is still poor. The reasons are three-fold: *(1)* events are vague, *(2)* events carry multiple perspectives and *(3)* events have different granularity. The mainstream procedure for event annotation is by means of experts. However, even experts disagree a lot. To overcome this, people create strict annotation guidelines which instead, make the task rigid and hardly adaptable to other domains. This overgeneralization does not deal with the intrinsic ambiguity, the multitude of interpretations and perspectives of the language. Thus, many NLP tools suffer from lack of training and evaluation data [3], as well as understanding the ambiguity. Setting up annotation tasks is also time and cost consuming due to both the length of the process and the costs associated with the experts. The constant lack of training data is also a downfall for increasing the performance of tools to automatically assess event novelty [4] and event clustering (*e.g.*, Google News[1]).

Crowdsourcing has emerged as a reliable, time and cost efficient approach for gathering semantic annotations. Typical solutions for assessing the quality of crowdsourced data are based on the hypothesis [5] that there is only one right answer. However, this contradicts with the three angles of events, *i.e.*, vagueness, multiple perspectives and granularities and with the natural language ambiguity. Recent work [6,7] has shown that disagreement between workers is a signal for identifying low quality workers and provides better understanding of the data ambiguity. A major crowdsourcing bottleneck is that most practices are not systematic and sustainable, while state-of-the-art methods are only developed for a specific domain or input. Crowdsourcing became an efficient way of gathering ground truth data for active learning systems [8] as well. However, this did not change the assumption that there is only one correct answer. Thus, the variety of perspectives, interpretations and language ambiguity are still not considered.

The current research defines events as *something that happened, is happening or will happen*, thus, using minimum restrictions. The primary purpose and focus is to gain event understanding by exploring event streams with regard to

[1] https://news.google.com.

(1) surface form, *i.e.*, event granularity; *(2)* space, *i.e.*, actors, location, time period; *(3)* relevance, *i.e.*, the most representative entities or phrases; *(4)* subjective perspectives, *i.e.*, the sentiment an event or entity triggers; *(5)* novelty and salience, *i.e.*, new or notable event features. Our experimental workflow builds on *(1)* machine-optimized stages where the data is pre-processed and pre-annotated with semantics and *(2)* crowd-driven gathering of event semantics ground truth. The novelty of the research comes from the event-centric approach of generating a ground truth of events, *i.e.* dealing with various concepts around events.

Our aim is to investigate how events are perceived and represented across data modalities (*e.g.*, text, image, video), sources (*e.g.*, news articles, tweets, video broadcasts) and languages. We want to analyze to what extent the crowd can help the NLP tools to understand events and improve their event detection. In summary, our aim is to improve event understanding and translate the data that we gather through experiments into a ground truth that can be used for training machines. Further, we aim to integrate machines and humans in a systematic way, *i.e.*, with focus on experimental methodologies and replicability and a sustainable way, *i.e.*, with focus on reusability of data, code and results.

2 State of the Art

In the literature, the process of detecting and extracting events presents high interest. The research focussed on this topic covers multiple perspectives, among others: event detection, related entities extraction, sentiments and novelty. Many NLP tools deal with extracting NE [1], but only a few with event detection such as, *e.g.*, OpenCalais[2], FRED[3]. One drawback of the supervised machine learning tools is the need of manually annotated data. For each new domain or data type new annotation guidelines need to be created. Preliminary research has been done on extracting earthquake-related events from tweets using distant supervision [9]. Their results show a performance of 88 % compared to a system that was trained on manually annotated microblogs. Using the sentence dependency tree provides F1-score of 53 % for event recognition on biomedical data [2].

For extracting event-related concepts such as people, locations, NER tasks are envisioned. In [10], the authors extract such concepts from video synopsis. State-of-the-art NE extractors [1] are developed using different algorithms and training data, making each targeted for specific NE recognition and classification tasks or more reliable on particular data [11]. [12] shows that evaluating agreement among extractors is effective: entities missed by one extractor can be found by others. Annotation of texts with heterogeneous topics and formats benefits from integrating extractors [12,13]. Nevertheless, semantic annotation of texts with heterogeneous topics, like news articles or TV-news bulletins is challenging, due to difficulties in training a single extractor to perform well across domains.

Agreement among NER tools is well captured by majority vote systems [14]. However, this could cut off relevant information such as, information supported

[2] http://viewer.opencalais.com/.
[3] http://wit.istc.cnr.it/stlab-tools/fred.

by only one extractor and cases with more than one solution. The evaluation of different extractor results will show disagreement, thus, aggregation of different NER tool results does not always solve the problem. Capturing events and other keywords from videos is even more challenging. A mainstream approach to tackle this problem is through automatic enrichment of the metadata [15]. This can be done either through machine processing of textual documents related to the videos [10,16], or through crowdsourcing descriptive tags for the media itself [17] and the media description [18]. Each of those approaches achieves reasonable results, however, each of them processes the videos from only one perspective.

Extensive research is also performed on event summarization, novelty and sentiment analysis. In [19], the authors perform single document summarization for creating news highlights by combining news articles with microblogs. However, this method has a very restrictive set of tweets that are considered relevant, *i.e.*, only the tweets that are linked to the article. An extensive literature study of automated novelty detection in texts is presented in [20]. Crowdsourcing proved to be a useful method for gathering ground truth on temporal events ordering in [21]. Another dimension of events is given by the sentiment [22,23]. Crowdsourcing has been also used on annotating NE in tweets [24], but its value on event annotation has not been thoroughly tackled until now.

We perform the crowdsourcing experiments in the context of the CrowdTruth [25] approach and methodology [6,7]. Using this approach the crowd annotations are stored in a vectorized fashion on which we apply cosine measures to identify low quality workers, unclear and ambiguous input units and annotation labels.

3 Problem Statement and Contributions

Current research builds upon the limitations of existing approaches on gathering and detecting event semantics presented in Sect. 2. Furthermore, we define the main research question: *can event detection tools benefit and gain event understanding by employing hybrid machine-optimized crowd-driven event semantics?*. The research novelty is two-fold: *(1)* the approach of gathering a ground truth by studying how events are represented in different modalities; *(2)* presenting the results in a machine readable form for improving NLP tools performance. To answer the main research question, we focus on the following sub-questions:

1. *How can we take advantage of existing natural language processing tools in order to ease and optimize the process of understanding events?*
   Many NLP tools deal with extracting NE that usually define the event space, but, they all have different precision. We aim to identify key features of the data for which different tools perform better than the rest and thus, take an informed decision of which tool should be used for the case at hand.
2. *How can we employ an optimized, replicable across data types, hybrid machine-crowd workflow for event understanding?*
   We focus on developing a workflow that combines machine processing and crowd perspectives for understanding event-related features across data types.

3. *How can we provide reliable crowdsourced training data to automated tools?* The multitude of crowdsourcing experiments that we need to perform to understand events requires us to validate the existing CrowdTruth metrics. Moreover, we need to adapt or define new disagreement metrics in the context of the CrowdTruth methodology to provide reliable crowdsourced data.
4. *How can we improve existing event and event-related feature annotation tools with machine-optimized and crowd-generated data?* Throughout the entire research process we focus on event exploration and understanding. As a final goal, we aim to grasp the expertise and semantics gained and ingest them as training data in specialized, existing NLP tools.

4 Research Methodology and Approach

We propose to answer the research questions defined in Sect. 3 and tackle some of the limitations presented in Sect. 2 through hybrid machine-crowd workflows. Our research methodology is three-fold, as shown in Fig. 1: *(i)* ***data enrichment*** layer where the data is enriched by machines and humans throughout a series of tasks; *(ii)* ***analytics*** layer that evaluates the results of the data enrichment layer and generates feedback; *(iii)* ***feedback*** layer that creates a continuous feedback loop throughout the methodology.

4.1 Data Enrichment Layer

The first step of our methodology is to apply relevant *machine processing and annotation* for the data input type and language at hand, while determining the *suitable input data*. In general, events are characterized by many features and thus, their detection and interpretation need multiple iterations and various annotation goals. Thus, the next step is *defining the annotation task*. At this stage we decide over the features that we want to annotate, implement the *annotation template*, choose the *crowdsourcing platform* and setup the *crowd-specific settings*. For every crowdsourcing experiment we have a *pilot run* that helps us identify the proper setup of the aforementioned methodology steps.

A couple of aspects make our methodology an efficient approach for gathering event understanding. On the one hand, our crowdsourcing tasks are replicable across data types (*e.g.*, running sentiment analysis and novelty ranking for news and tweets). This reduces the time spent to setup a proper crowdsourcing experiment for a new data type and generalizes the approach. One the other hand, we perform a gradual workflow of crowdsourcing experiments. The output of one crowdsourcing task becomes the input for a following task (*i.e.* data units that are not relevant for our event understanding goal are immediately discarded).

4.2 Analytics Layer

Human annotation is a process of semantic interpretation, often described by the triangle of reference [6]. Thus, disagreement signals low quality that can

be measured at any corner of the triangle: input, workers or annotation labels. In the **Analytics** layer we deal with the assessment of the data produced in the **Data Enrichment** layer. In order to evaluate the results of the crowd we first use the *vector representation* to measure the quality of the annotations in the three corners of the triangle. Thus, each worker annotation is stored in a vectorized data structure which eases the overall evaluation of the results. Second, we apply CrowdTruth metrics that provide useful insights for: *(1)* spam or low-quality *crowd workers*; *(2)* ambiguous or unclear *input units*; *(3)* unclear *annotation labels*. We iterate this process based on the input provided by the **Feedback Layer**, until we have a set of clear, reliable and correct results.

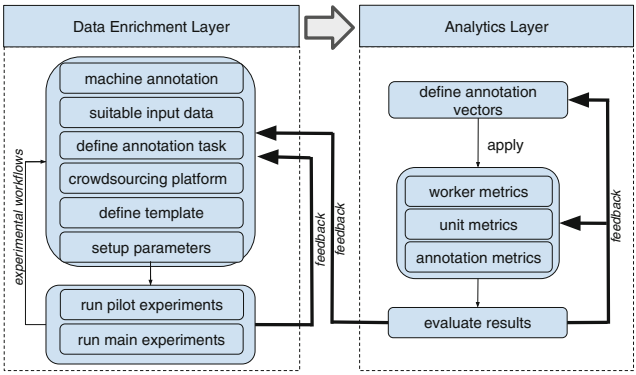

Fig. 1. Research methodology

4.3 Feedback Layer

The final part of our research methodology integrates various feedback loops that aim to improve the overall system. The feedback loops take advantage of the analysis generated in the **Analytics** layer and applies the necessary changes. Based on the current set of results, decisions are evaluated as follows. We assess the suitability of the data in the **Data Enrichment** layer and apply additional pre-processing or filtering steps. Similarly, based on the analysis of the *pilot experiments*, *i.e.*, the overall task clarity and crowd workers performance, we re-design the annotation task and template and tune the parameters. We analyze *(1)* the outcome of the *disagreement metrics* and improve the spam filtering (*i.e*, tune the worker metrics or perform additional filtering) and *(2)* the evaluation of the crowdsourcing outcomes (*i.e*, tune the worker and unit vector space).

5 Preliminary Results

In the initial stage of this PhD project we performed several experiments, across data streams and languages. We report on *(1)* event extraction from Dutch video

synopsis and *(2)* event space and sentiment analysis on English news articles and tweets. The crowdsourcing experiments were run on the CrowdFlower[4] platform and were analyzed using the CrowdTruth methodology and metrics [7,26].

The results of the first experiment are integrated in the DIVE[5] demonstrator [27], a linked-data digital cultural heritage collection browser for collections interpretation and navigation by means of events and related entities. The experiments were performed on a dataset of 300 video synopsis of news broadcasts from The Netherlands Institute for Sound and Vision. We use various NER tools to extract NE from video synopses, but their accuracy vary significantly. Empirical analysis of the results indicated that combining the output of several NER tools provides better results than using a single extractor (*e.g.*, a richer set of entities, types, surface forms). However, a thorough evaluation is necessary to answer the first sub-question. Section 6 provides a plan for this. Next, we designed crowdsourcing experiments for tasks were machines under-perform: *(i)* extracting events, *(ii)* linking events to their participating entities.

For the second set of experiments we identify relevant news snippets and tweets for the event of whaling. We performed crowdsourcing experiments to identify relevant texts, word phrases and sentiments that are expressed in news and tweets. The results of these experiments[6] present preliminary insights for identifying salient features across data sources. Further experiments will be performed for novelty assessment and determining the event features saturation, *i.e.*, compiling, over time, a complete set of side-events, participating actors, locations, time frames and subjective perspectives such as sentiment changes.

Overall, the experiments performed gave us an important head start for reasoning on the third research question. We gain understanding of various crowdsourcing tasks and we are able to analyze the crowd workers behavior in order to assess their work. We applied existing CrowdTruth disagreement metrics and defined new metrics and filters that identify with high precision, recall and accuracy the spam and low-quality workers. The metrics proved to support us in providing reliable crowdsourced event semantics annotations. However, we only performed manual evaluation of the data, which is not scalable over time. We plan to extend this evaluation with automated methods and ground truth data.

6 Evaluation Plan

Our modular and multi-layer research methodology with combined machine and human generated data stages, needs various types of evaluation. To answer the first sub-question from Sect. 3 we need to perform extensive automated annotations by means of NLP tools. Benchmarking such tools helps us to gain enough empirical evidence for choosing the most accurate tool based on the input (*e.g.* text dimension, language, domain) and the task (*e.g.* identification of locations,

[4] http://www.crowdflower.com/.
[5] Available at http://diveplus.beeldengeluid.nl/.
[6] http://data.crowdtruth.org/salience-news-tweets/.

actors, times) at hand. The "Open Knowledge Extraction" challenge at ESWC[7] and frameworks such as GERBIL [28] are good systems to validate our approach.

In the context of sub-question 3, we will perform various crowdsourcing tasks (*e.g.* event detection, sentiment analysis, novelty ranking). For each task we need to validate the correctness of the crowdsourced data. Simply applying the CrowdTruth disagreement metrics may not always correctly identifying the low-quality workers. The main component of the crowdsourcing task is the *task template*. Depending on its fields (*e.g.* multiple or single choice, free input text), low-quality workers can exhibit particular behaviors, such as, always choosing the same answer, taking the shortest path to solve the task. In order to guarantee reliable data, additional behavioral and disagreement metrics are applied. Further, curated annotations can be compared with existing ground truth datasets.

To evaluate our hybrid machine-crowd workflow we start by instantiating it with a specific input type and use case. This gives us insights to answer the inter-related sub-questions 1 and 3 and finally sub-question 2. Next, we identify similarities between input types and tasks, as well as their independent features, which helps us investigating to what degree our methodology is data agnostic. The semantic annotations generated throughout the various annotations tasks will be used to answer sub-question 4 and ultimately our main research question. We investigate here whether the event-specific generated data (*i.e.*, events, concepts, sentiments, salient features) can improve understanding of events and whether this ground truth can be used in improving machines performance.

Overall, the task of understanding events is tackled and investigated by many venues. TREC[8] is focusing on the Temporal Summarization Track, a track aiming to develop a system that is able to provide concise and non-redundant information with regard to a given event. The structure of this challenge is very close to our novelty approach: identifying relevant and novel texts that describe an event. In the past years, main conferences focused on sentiment analysis as well. For example, our crowd-annotated corpus could be evaluated in semantic web challenges, such as, "Concept-Level Sentiment Analysis" during ESWC.

7 Conclusions

The primary focus of this research proposal is to gain event understanding through employing automated tools and collecting diverse crowd semantic interpretations on different data modalities, sources and event-related tasks. Our main hypothesis is that existing tools for detecting events and event-related features can acquire event semantics and understanding by employing such additional training data or by validating their current results with crowd-empowered semantics. In order to answer our main research question, we defined a set of 4 related research questions to investigate: *(1)* how can we harness the semantics of existing tools; *(2)* how can we create data agnostic annotation workflows;

[7] http://eswc-conferences.org.
[8] http://trec.nist.gov.

(3) how can we guarantee reliable crowdsourced semantics; and *(4)* how can we improve existing tools with our machine-optimized and crowd-generated data.

The research methodology consists of a continuous experimental loop of interconnected components for providing the event space, context, semantics and perspectives. The methodology allows for: *(1)* running combined crowd and machine workflows across different data types; *(2)* analysis of the crowdsourced data; *(3)* extensive feedback layer for improving existing or future results. We presented preliminary results of in-use crowd-generated linked data and an example of an optimized and generic annotation workflow for identifying salient event features. Our evaluation plan is two-fold: *(1)* validating our methodology by testing it with new data types and *(2)* assessing our methodology by ingesting the semantics collected for training and testing existing tools.

Acknowledgements. We thank Lora Aroyo for helping in this research proposal, Robert-Jan Sips, Victor de Boer, Tommaso Caselli for assistance in performing the experiments.

References

1. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013)
2. McClosky, D., Surdeanu, M., Manning, C.D.: Event extraction as dependency parsing. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1626–1635 (2011)
3. Kim, S.M., Hovy, E.: Automatic detection of opinion bearing words and sentences. In: Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), pp. 61–66 (2005)
4. Soboroff, I., Harman, D.: Novelty detection: the TREC experience. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 105–112. ACL (2005)
5. 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 IR, pp. 557–566. ACM (2010)
6. Aroyo, L., Welty, C.: Truth is a lie: CrowdTruth and the seven myths of human annotation. AI Mag. **36**(1), 15–24 (2015)
7. Aroyo, L., Welty, C.: The three sides of CrowdTruth. J. Hum. Comput. **1**, 31–34 (2014)
8. Yan, Y., Fung, G.M., Rosales, R., Dy, J.G.: Active learning from crowds. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 1161–1168 (2011)
9. Intxaurrondo, A., Agirre, E., de Lacalle, O.L., Surdeanu, M.: Diamonds in the rough: event extraction from imperfect microblog data. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) (2015)
10. Li, Y., Rizzo, G., Redondo García, J.L., Troncy, R., Wald, M., Wills, G.: Enriching media fragments with named entities for video classification. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 469–476 (2013)

11. Rizzo, G., van Erp, M., Troncy, R.: Benchmarking the extraction and disambiguation of named entities on the semantic web. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, pp. 4593–4600 (2014)
12. Chen, L., Ortona, S., Orsi, G., Benedikt, M.: Aggregating semantic annotators. Proc. VLDB Endowment **6**(13), 1486–1497 (2013)
13. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) ISWC 2013, Part II. LNCS, vol. 8219, pp. 98–113. Springer, Heidelberg (2013)
14. Kozareva, Z., Ferrández, Ó., Montoyo, A., Muñoz, R., Suárez, A., Gómez, J.: Combining data-driven systems for improving named entity recognition. Data Knowl. Eng. **61**(3), 449–466 (2007)
15. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., et al.: Semantic annotation and search of cultural-heritage collections: the MultimediaN E-Culture demonstrator. Web Seman. Sci. Serv. Agents WWW **6**(4), 243–249 (2008)
16. Oomen, J., Belice Baltussen, L., Limonard, S., van Ees, A., Brinkerink, M., Aroyo, L., Vervaart, J., Asaf, K., Gligorov, R.: Emerging practices in the cultural heritage domain-social tagging of audiovisual heritage. In: Proceedings of the WebSci 2010: Extending the Frontiers of Society On-Line (2010)
17. Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G.J., Aroyo, L.: Crowdsourcing knowledge-intensive tasks in cultural heritage. In: Proceedings of the 2014 ACM Conference on Web Science, pp. 267–268. ACM (2014)
18. Maccatrozzo, V., Aroyo, L., Van Hage, W.R., et al.: Crowdsourced evaluation of semantic patterns for recommendation. In: UMAP Workshops (2013)
19. Wei, Z., Gao, W.: Utilizing microblogs for automatic news highlights extraction. In: COLING (2014)
20. Verheij, A., Kleijn, A., Frasincar, F., Hogenboom, F.: A comparison study for novelty control mechanisms applied to web news stories. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 431–436. IEEE (2012)
21. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in NLP, pp. 254–263 (2008)
22. Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M.: Building emotional dictionary for sentiment analysis of online news. World Wide Web **17**(4), 723–742 (2014)
23. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: Sentiment analysis in the news. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, pp. 2216–2220 (2010)
24. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in twitter data with crowdsourcing. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 80–88. ACL (2010)
25. Inel, O., Khamkham, K., Cristea, T., Dumitrache, A., Rutjes, A., van der Ploeg, J., Romaszko, L., Aroyo, L., Sips, R.-J.: CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In: Mika, P., et al. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 486–504. Springer, Heidelberg (2014)

26. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Measuring crowd truth: disagreement metrics combined with worker behavior filters. In: Proceedings of CrowdSem 2013 Workshop, ISWC (2013)
27. de Boer, V., Oomen, J., Inel, O., Aroyo, L., van Staveren, E., Helmich, W., de Beurs, D.: Dive into the event-based browsing of linked historical media. Web Semant. Sci. Serv. Agents WWW **35**(3), 152–158 (2015)
28. Usbeck, R., Röder, M., Ngonga Ngomo, A.C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al.: Gerbil: general entity annotator benchmarking framework. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1133–1143 (2015)