Crowdsourcing Disagreement for Collecting Semantic Annotation

Anca Dumitrache(✉)

VU University Amsterdam, Amsterdam, The Netherlands
anca.dumitrache@vu.nl

Abstract. This paper proposes an approach to gathering semantic annotation, which rejects the notion that human interpretation can have a single ground truth, and is instead based on the observation that disagreement between annotators can signal ambiguity in the input text, as well as how the annotation task has been designed. The purpose of this research is to investigate whether disagreement-aware crowdsourcing is a scalable approach to gather semantic annotation across various tasks and domains. We propose a methodology for answering this question that involves, for each task and domain: defining the crowdsourcing setup, experimental data collection, and evaluating both the setup and the results. We present initial results for the task of medical relation extraction, and propose an evaluation plan for crowdsourcing semantic annotation for several tasks and domains.

Keywords: Crowdsourcing · Human computation · Ground truth · Natural language processing · Semantic annotation · Semantic ambiguity

1 Introduction/Motivation

As knowledge available on the Web expands, information extraction methods have become invaluable for facilitating data navigation and populating the Semantic Web. Gathering semantic data in the form of entities and relations between existing datasets, is central to information extraction systems (i.e. the task of machine learning for most analytics). Human-annotated gold standard, or ground truth, is used for training, testing, and evaluation of information extraction components. The traditional approach to gathering this data is to employ experts to perform annotation tasks.

However, such an annotation process can be both expensive, and time consuming [1], due to the costs associated with working with domain experts. Furthermore, experts might prove difficult to find for broad, open domains (e.g. sentiment analysis). This presents a challenge for extending information extraction methods, and concurrently Semantic Web systems into new domains. Human annotation is needed to solve this problem, but the process of gathering this data is not scalable at the level of the large datasets currently available on

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 701–710, 2015.
DOI: 10.1007/978-3-319-18818-8_43

the Web. Efficiently integrating human knowledge with automated procedures is necessary for tackling this issue [28].

IBM is facing this problem when adapting the question-answering system Watson [12] to new domains. To compete in the Jeopardy TV quiz show, Watson was trained on publicly available datasets, taxonomies, and ontologies. Adapting the system to the medical domain, however, requires large amounts of human-annotated data, as medical resources on the Web are not so readily available.

Furthermore, unlike the Jeopardy setup, where one correct answer exists for every question, the medical domain is more ambiguous – it is often the case that doctors disagree on the same diagnosis. The traditional approach to gathering annotation is based on restrictive annotation guidelines, and often results in over-generalized observations, as well as a loss of ambiguity inherent to language [1], thus becoming unsuitable for use in training information extraction systems.

Being cheaper and more scalable, crowdsourcing is a possible alternative to using dedicated annotators. Crowdsourcing also allows for collecting enough annotations per task in order to represent the diversity inherent in language. By employing a large crowd to collect semantic annotations, it becomes possible to observe inter-annotator disagreement. Previous research in crowdsourcing medical relation extraction [2,3] has shown that disagreement can be an informative, useful property, and its analysis can result in reduced time, lower cost, better scalability, and better quality human-annotated data.

This paper describes a PhD project within the context of CrowdTruth, a larger initiative investigating how disagreement-aware crowdsourcing can be used to collect annotations for text, videos, and images. Building on previous successful results [2,3], this paper aims to explore the question of how crowdsourcing can be employed as a tool for collecting semantic annotation. Specifically, we analyze the role of disagreement, and whether its analysis can be used to improve the quality of existing semantic ground truth. We propose a methodology to crowdsource a series of semantic annotation tasks (e.g. relation extraction, sentiment analysis), with the purpose of demonstrating that disagreement can be a task and domain-independent indicator of semantic ambiguity.

2 State of the Art

Crowdsourcing for collecting semantic annotation has been used before in a **variety of tasks and domains**: medical entity extraction [13,27], clustering and disambiguation [19], relation extraction [18], and ontology evaluation [20]. However, most of these approaches rely on the assumption that one universal gold standard must exist for every task. Disagreement between annotators is discarded by either restricting annotator guidelines, or picking one answer that reflects some consensus usually through using majority vote. The number of annotators per task is kept low, typically between one and three workers, also in the interest of eliminating disagreement.

There exists some research on the impact of ambiguity on crowdsourcing annotation. In assessing the OAEI benchmark, [9] found that disagreement

between annotators (both crowd and expert) is an indicator for inherent uncertainty in the domain knowledge, and that current benchmarks in ontology alignment and evaluation are not designed to model this uncertainty. Reference [22] found similar results for the task of crowdsourced POS tagging – most inter-annotator disagreement was indicative of debatable cases in linguistic theory, rather than faulty annotation. In our approach, we use the crowd to collect semantic annotation, and harness inter-annotator disagreement as an inherent feature of semantic interpretation.

There is extensive literature on how to **measure crowdsourcing results**. Of particular interest are ways of identifying spam workers [8,15,16], and analyzing workers' performance for quality control and optimization of the crowdsourcing processes [23]. However, these approaches rely on a series of faulty assumptions about ground truth quality [5]: (1) that there exists a single, universally constant truth, (2) that this truth can be found through agreement between annotators, (3) that high agreement means high quality, and (4) that disagreement needs to be eliminated. Consequently, most crowdsourcing metrics attempt to measure the quality of the workers, without accounting for the ambiguity in the input text, and the clarity of the annotation task. Human annotation is a process of semantic interpretation, often described using the triangle of reference [17] linking: sign (input text), interpreter (worker), referent (annotation). Ambiguity for one aspect will propagate in the triangle – an unclear sentence will cause more disagreement between workers. Therefore, we design metrics to harness disagreement for each of the three aspects of the triangle, measuring the quality of the worker, as well as the ambiguity of the text and the task.

Evaluating crowd performance with existing benchmarks has been performed for a variety of tasks and domains. Reference [27] show that the crowd can perform just as well as experts in medical entity extraction. Reference [24] prove the crowd can match the experts for another five annotation tasks: affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation. Our approach mainly targets a crowd of lay workers, and we evaluate the results through comparison with existing gold standards, annotated by experts or automatically collected.

More knowledge-intensive tasks have proved difficult for the crowd to solve. Reference [21] show that tagging flowers with their botanical names could not be performed by a crowd of lay people. In these cases, nichesourcing [10], or employing crowds of experts to perform the annotations, can combine the advantages of using a crowd with the domain knowledge of experts. We define a generalizable methodology for crowdsourcing that we can then use to run nichesourcing experiments.

3 Problem Statement

Based on the issues we identified in Sect. 2, we define our main research question: *is disagreement-aware crowdsourcing a scalable approach to gather semantic annotation across various tasks and domains?* This can be broken down into several sub-questions:

1. How to measure disagreement in a crowdsourcing setup for semantic annotation? The triangle of reference [17] indicates disagreement holds different meanings for each of its three concepts, and therefore we need to define separate metrics that capture disagreement and use it to measure:
   (a) the ambiguity of an input *text unit*;
   (b) the clarity of an *annotation*;
   (c) the quality of a *worker*.
2. How is disagreement present in crowdsourcing across different domains and tasks related to semantic annotation? We investigate this by applying our disagreement metrics to crowdsourcing data and performing a comparative analysis across:
   (a) tasks: *text annotation* (entity extraction, relation extraction, relation direction), *alignment* (passage alignment, ontology alignment).
   (b) domains: requiring *expertise* (medical domain), requiring *no expertise* (open domain, sentiment analysis).
3. How does crowdsourcing compare to existing gold standard baselines? Factors to consider here are:
   (a) *implementation costs:* Is crowdsourcing cheaper, less time-consuming, more scalable than the usual gold standard approaches?
   (b) *quality of data:* Is crowdsourcing data more reliable than traditional ground truth data?

4 Research Methodology and Approach

To answer the research questions defined in Sect. 3, we aim to perform a series of crowdsourcing experiments across several types of annotation tasks, in a variety of domains. To conduct each of these experiments, we define a methodology consisting of three steps (Fig. 1).

The initial step in our approach consists of **defining the crowdsourcing setup**. We first *identify suitable data*, in the form of raw text together with available annotation, to perform our experiments. Suitable in this context refers

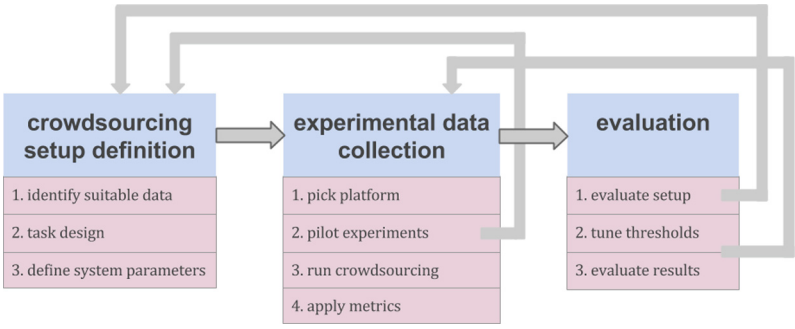

Fig. 1. Research methodology steps.

to data that (a) has some degree of complexity that makes it difficult to automatically extract annotation and (b) contains subjective opinion that could be interpreted in different ways by the crowd. Based on these features, we identified several candidate datasets:

- *Wikipedia medical sentences*: relations and entities can be automatically collected with distant supervision [26] and the UMLS vocabulary of medical terms [7], but the data contains noise and requires human input for correction;
- *Wikipedia open domain sentences*: relations and entities can be automatically collected with distant supervision [6] using DBpedia entities, also produces noisy data;
- *Twitter statuses*: contain a variety of subjective opinions on current events, that can be retrieved based on hashtags.

Next, we perform the *task design*, where we break down the annotation task into a workflow of independent micro-tasks that can be performed by the workers. The task design needs to structure the crowd annotations in a way that can be quantified by the disagreement metrics. When possible, the crowd will be asked to pick an answer from a given vocabulary (i.e. common medical relations, open domain relations, sentiments). For entity extraction, however, the goal is to crowdsource the vocabulary itself, so the workers are allowed to pick any combination of words in the input text. Entity clustering is then employed to reduce the noise in the answer set. To ensure that the data can be measured using our metrics, we aggregate the crowd answers into answer vectors per task.

We then *define the system parameters*. The metrics for measuring disagreement are defined at this step. The first to define is the sentence-annotation score – making use of the answer vectors introduced in the task design, it computes the likelihood that a given annotation exists in one input sentence. Based on it, we define metrics to harness disagreement at the level of the worker (to differentiate spammers from quality workers that nevertheless diverge from the majority), unit (to find unclear input data), and annotation (to find ambiguity in the annotation design). Also part of the system parameters are the settings for running the task: (1) how much *time* will the workers have to solve it, (2) how much the *payment* will be, and (3) the *number of workers* that will be solving one task.

The second step in our methodology is the **experimental data collection**. Here we *decide on the crowdsourcing platform* to run our experiments. Two established options exist: CrowdFlower[1] gives access to a larger workforce with less possibility to interact with the workers, whereas Amazon Mechanical Turk[2] has a smaller workforce with whom the task creators can communicate directly, thus creating a community of returning workers. To overcome these weaknesses and have an environment tailored to our crowdsourcing tasks, we also work at creating our own platform, through which we will be able to target

[1] http://CrowdFlower.com.
[2] http://MTurk.com.

a specific crowd that possesses domain expertise, for the purpose of nichesourcing experiments.

Next we perform several *pilot experiments* on a small sample of the data, for the purpose of tuning the crowdsourcing setup defined in the previous step of the methodology. According to these results, we adjust the task design, as well as the system parameters (e.g. worker pay, thresholds for detecting spam etc.). This enables us to *run the crowdsourcing tasks* on the entirety of the data. Finally, we *apply the metrics* to the data we collected, an use them to identify and remove spam workers, ambiguous input and unclear annotations.

The final step in our methodology is to perform an **evaluation**. First, we *evaluate the setup*, and *tune the thresholds* of the system parameters, to determine whether the task design and system parameters performed well. Overwhelming agreement or disagreement between workers serve an indicator for bad selection of data, or faulty crowdsourcing setup, meaning that the corresponding steps in the methodology need to be repeated.

The quality of the data is measured by *evaluating the results* in comparison with existing baselines for semantic annotation. In knowledge-intensive domains, the baseline refers to gold standard expert annotation, whereas in cases where no domain knowledge is required, we can compare also to data that is automatically collected. Evaluation is done in two ways: (1) directly, by studying the overlap between crowd data and baseline and identifying the features of data units where baseline and crowd disagree, and (2) using machine learning, by training and testing a model for information extraction (i.e. relation extraction, named entity recognition, sentiment analysis) with both crowd and baseline data.

5 Evaluation Plan

We evaluate our methodology by **instantiating with a specific task and domain**, with the goal of outlining an answer for the second two research sub-questions: (1) defining disagreement-aware metrics, and (3) comparing to an existing baseline (Sect. 3). Based on the IBM Watson use case described in Sect. 1, we define a setup for the task of *medical relation extraction*. The purpose of this experiment is to evaluate crowdsourced semantic relation data in comparison with ground truth generated by medical experts, in the context of training a relation extraction classifier for IBM Watson. The results are detailed in Sect. 6.

The next step will be to **perform the same annotation task** (i.e. semantic relation extraction) **in other domains**, for the purpose of answering part of research sub-question (2) observing how our setup performs cross-domain, while also refining the answers for research sub-questions (1) and (3). As the previous experiment was based in a knowledge-intensive domain, we investigate next the open domain, using a different text corpus with the *DBpedia* vocabulary for identifying both terms and relations. For comparison purposes, the task design and workflow, as well as the metrics for harnessing disagreement are the same.

Consequently, we plan to **analyze other semantic annotation tasks** in the same domains. The purpose of these experiments is to analyze how semantic

disagreement is present cross-task, also as part of research sub-question (1). For the medical domain, we analyze crowdsourced *named entity recognition for medical terms*. For the open domain, we plan to perform sentiment analysis tasks over *Twitter* data. We are also planning to investigate semantic alignment tasks, such as passage alignment and ontology alignment. For knowledge-intensive tasks and domains, we plan to perform *nichesourcing* experiments involving crowds of experts. Specifically, we target the medical community through our own gamified crowdsourcing platform [11].

Finally, we will perform a **review across all experiments** that we investigated. The purpose is to identify the features of disagreement in crowdsourced annotation data that are independent of task and domain. We also investigate to what extent our methodology is generalizable for any combination of semantic annotation task and domain. The results of this analysis will then be used to answer our main research question on whether disagreement-aware crowdsourcing is a scalable approach to gather semantic annotation.

6 Preliminary Results

For the first set of experiments part of the evaluation plan (Sect. 5) – performing crowdsourced relation extraction over medical text – we designed a workflow of crowdsourcing tasks performing named entity correction, relation extraction, and relation direction (example templates in Fig. 2). We ran these tasks on both CrowdFlower and Amazon Mechanical Turk, collecting crowd judgments for around 2,000 medical sentences.

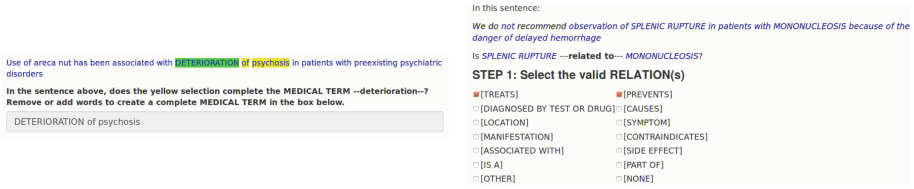

Fig. 2. Tasks on CrowdFlower (from left to right: entity correction, relation extraction)

To process these results and answer research sub-question (1) (Sect. 3), we modeled the crowd answers as vectors, and defined a set of metrics based on cosine similarity. The metrics are used to harness disagreement, and measure the quality of crowd workers, ambiguity of medical sentences, and the clarity of medical relations employed. Results on a series of pilot experiments have been published [4,25], using this setup on a restricted set of medical sentences, where the crowd results were manually evaluated against expert judgments. This thesis aims to build on these initial experiments by exploring how disagreement is present in large datasets, across a variety of tasks and domains, and analyze how it can be used to build a better ground truth for semantic annotation.

The aim is to develop a scalable, semi-automated method for collecting semantic annotation with crowdsourcing.

We have published work on CrowdTruth [14], our framework for crowdsourcing ground truth data. CrowdTruth connects with both Amazon Mechanical Turk and CrowdFlower for launching and monitoring tasks, and implements the disagreement metrics for a live analysis of the results from the crowd. We are also developing a gamified platform, Dr. Detective [11], targeting medical experts for nichesourcing tasks such as extracting entities related to a given diagnosis.

Currently, we are investigating the usefulness of crowdsourced data in training and evaluating a machine learning model, in order to answer research subquestion (3). We train a medical relation extraction classifier [26] using both crowd results and expert judgments in a cross-validation experiment, and compare the results of the evaluation for each dataset using accuracy and F1 score. In a paper that is currently in submission, we prove that, in training the model, crowdsourced data from the lay crowd, that has been weighted with disagreement scores, performs just as well as gold standard data from medical experts.

7 Conclusions

In this paper, we explored how crowdsourcing can be used to collect semantic annotation. We proposed an approach rejecting the notion that human interpretation can have a single ground truth, and is instead based on the observation that disagreement between annotators can signal ambiguity in the input text, as well as how the annotation task has been designed. In order to analyze this hypothesis, we defined three research questions: (1) how to measure disagreement in a crowdsourcing setup for semantic annotation, (2) how is disagreement present in crowdsourcing across different domains and tasks for semantic annotation, and (3) how does crowdsourcing compare to the existing gold standard.

We defined a three-step methodology for answering these questions: (1) defining the crowdsourcing setup, (2) experimental data collection, (3) evaluating both the setup and the results. As preliminary work, we presented the CrowdTruth platform for crowdsourcing tasks with disagreement metrics, and the results of an experiment comparing the crowd against experts for medical relation extraction. As part of our evaluation plan, we will extend our work by performing relation extraction experiments in the open domain, as well as implementing other annotation tasks (sentiment analysis, ontology alignment). To answer our research question, we will perform a comparative analysis of our results across tasks and domains, identifying the generalizable characteristics of disagreement in crowdsourced annotation.

Acknowledgments. We thank Lora Aroyo and Chris Welty for helping develop this research plan, Abraham Bernstein for help with editing the paper, Robert-Jan Sips, Anthony Levas and Chang Wang for assistance with performing the experimental work.

References

1. Aroyo, L., Welty, C.: Harnessing disagreement for event semantics. In: 11th International Semantic Web Conference Proceedings of the 2nd International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012), p. 31 (2012)
2. Aroyo, L., Welty, C.: Crowd truth: harnessing disagreement in crowdsourcing a relation extraction gold standard. In: WebSci2013. ACM (2013)
3. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. In: AAAI 2013 Fall Symposium on Semantics for Big Data (2013)
4. Aroyo, L., Welty, C.: The three sides of crowdtruth. J. Hum. Comput. **1**, 31–34 (2014)
5. Aroyo, L., Welty, C.: Truth Is a Lie: 7 Myths about Human Computation. AI Magazine (2014) (in press)
6. Augenstein, I.: Joint information extraction from the web using linked data. In: Mika, P. (ed.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 505–512. Springer, Heidelberg (2014)
7. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(suppl 1), D267–D270 (2004)
8. Bozzon, A., Brambilla, M., Ceri, S., Mauri, A.: Reactive crowdsourcing. In: Proceedings of the 22nd International Conference on World Wide Web WWW 2013, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 153–164. (2013). http://dl.acm.org/citation.cfm?id=2488388.2488403
9. Cheatham, M., Hitzler, P.: Conference v2.0: an uncertain version of the OAEI conference benchmark. In: Mika, P. (ed.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 33–48. Springer, Heidelberg (2014)
10. de Boer, V., Hildebrand, M., Aroyo, L., de Leenheer, P., Dijkshoorn, C., Tesfa, B., Schreiber, G.: Nichesourcing: harnessing the power of crowds of experts. In: ten Teije, A. (ed.) EKAW 2012. LNCS, vol. 7603, pp. 16–20. Springer, Heidelberg (2012)
11. Dumitrache, A., Aroyo, L., Welty, C., Sips, R.J., Levas, A.: "Dr. detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. In: 12th International Semantic Web Conference Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013) (2013)
12. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building watson: an overview of the deepqa project. AI Mag. **31**, 59–79 (2010)
13. 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 CSLDAMT 2010, pp. 80–88. Association for Computational Linguistics (2010)
14. Inel, O., et al.: CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In: Mika, P. (ed.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 486–504. Springer, Heidelberg (2014)
15. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. In: HCOMP 2010 Proceedings of the ACM SIGKDD Workshop on Human Computation. pp. 64–67. ACM, New York (2010). http://doi.acm.org/10.1145/1837885.1837906

16. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: CHI 2008 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 453–456. ACM, New York (2008). http://doi.acm.org/10.1145/1357054.1357127
17. Knowlton, J.Q.: On the definition of "picture". AV Commun. Rev. **14**(2), 157–183 (1966)
18. Kondreddi, S.K., Triantafillou, P., Weikum, G.: Combining information extraction and human computing for crowdsourced knowledge acquisition. In: 2014 IEEE 30th International Conference on Data Engineering (ICDE), pp. 988–999. IEEE (2014)
19. Lee, J., Cho, H., Park, J.W., Cha, Y.R., Hwang, S.W., Nie, Z., Wen, J.R.: Hybrid entity clustering using crowds and data. The VLDB J. **22**(5), 711–726 (2013). http://dx.doi.org/10.1007/s00778-013-0328-8
20. Noy, N.F., Mortensen, J., Musen, M.A., Alexander, P.R.: Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 262–271. ACM (2013)
21. Oosterman, J., Nottamkandath, A., Dijkshoorn, C., Bozzon, A., Houben, G.J., Aroyo, L.: Crowdsourcing knowledge-intensive tasks in cultural heritage. In: WebSci 2014 Proceedings of the 2014 ACM Conference on Web Science, pp. 267–268. ACM, New York (2014). http://doi.acm.org/10.1145/2615569.2615644
22. Plank, B., Hovy, D., Søgaard, A.: Linguistically debatable or just plain wrong? In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 507–511. Association for Computational Linguistics, Baltimore, June 2014. http://www.aclweb.org/anthology/P/P14/P14-2083
23. Singer, Y., Mittal, M.: Pricing mechanisms for crowdsourcing markets. In: WWW 2013 Proceedings of the 22nd International Conference on World Wide Web, pp. 1157–1166. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2013). http://dl.acm.org/citation.cfm?id=2488388.2488489
24. 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: EMNLP 2008 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics, Stroudsburg (2008). http://dl.acm.org/citation.cfm?id=1613715.1613751
25. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Measuring crowd truth: disagreement metrics combined with worker behavior filters. In: 12th International Semantic Web Conference on Proceedings of the 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem 2013) (2013)
26. Wang, C., Fan, J.: Medical relation extraction with manifold models. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 828–838. Association for Computational Linguistics (2014). http://aclweb.org/anthology/P14-1078
27. Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., Solti, I.: Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J. Med. Internet Res. **15**(4), e73 (2013)
28. Zhong, N., Ma, J.H., Huang, R.H., Liu, J.M., Yao, Y.Y., Zhang, Y.X., Chen, J.H.: Research challenges and perspectives on wisdom web of things (w2t). J. Supercomput. **64**(3), 862–882 (2013)