EXPLORING THE "CROWD" AS ENABLER OF BETTER INFORMATION QUALITY

(Practice-oriented paper)

Pascal Wichmann
Karlsruhe Institute of Technology
pascal.wichmann@gmail.com

Alexander Borek
University of Cambridge
ab865@cam.ac.uk

Robert Kern
Karlsruhe Institute of Technology
robert.kern@ksri.uni-karlsruhe.de

Philip Woodall
University of Cambridge
phil.woodall@eng.cam.ac.uk

Ajith Kumar Parlikad
University of Cambridge
ajith.parlikad@eng.cam.ac.uk

Gerhard Satzger
Karlsruhe Institute of Technology
gerhard.satzger@kit.edu

Abstract: In any organization information is one of the key resources and its quality needs to be managed to achieve and sustain organizational success. So far, due to high costs of maintaining staff, literature has been focusing on information quality improvements that are driven by technological solutions. However, these solutions have their limits – both technologically and economically as many tasks require human intelligence. In analogy to the "cloud" in cloud computing, the "crowd" is available 24/7 whenever it is needed to provide human-based electronic services in a low cost and very flexible way in micro-task markets like Amazon Mechanical Turk. This paper explores the new opportunities for organizations to improve information quality by looking at the possible areas of application of crowdsourcing. Moreover, we have tested the use of crowdsourcing for information quality improvement in one of the identified application areas, namely natural language processing, in the context of medical record transcription in a health insurance company.

Key Words: Crowdsourcing, Information Quality Improvement, Micro Task Markets, Natural Language Processing.

INTRODUCTION

Despite the impressive progress in artificial intelligence (A.I.) research in the last decade, many tasks can still not be automated. Human intelligence remains superior to computer algorithms at least in certain domains, especially when it comes to visual perception, human common knowledge, natural language processing, and expertise on certain topics [1]. Even when automatic approaches are performing well, they tend to be highly specified and inflexible towards changes in the job description. The combination of both automated electronic processes and human intelligence, however, has led to remarkable successes when tasks are outsourced to a large network of people on the Web. The term "crowdsourcing" was coined by Jeff Howe as a blend of the two words "crowd" and "outsourcing" [13]. In a subsequent publication, Howe defines crowdsourcing as "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call" [12]. For example, millions of people transcribe the archives of the New York Times into machine-encoded text, mostly unaware of what they are actually doing. By entering the two scanned words shown in a *reCAPTCHA* interface they perform a task that is trivial to humans but one that computers cannot yet perform [2]. About 35-40 million words are being transcribed by internet users per day in this project[1]. Other applications are games like *Games With a Purpose*[2], *FoldIt*[3], *Phylo*[4] or EteRNA[5] which leverage human intelligence by letting human players perform tasks that are hard to automate, e.g. labelling of images or helping scientists in finding structures of proteins or configuration of DNA strings. They thereby provide metadata, make information hidden in images machine-accessible and solve complex problems.

The applications mentioned above solely rely on the intrinsic motivation of the users; other tasks like the ones traded on micro-task markets [18] like Amazon Mechanical Turk[6] (MTurk) are based on monetary incentives, for instance, by paying a certain amount of money per task performed. On MTurk, the service requester has the option of rejecting work if the task result does not comply with the task instructions. In this case, the worker will not be paid. Service requesters can also exclude certain workers from performing future work for them. The only general requirement for participating on MTurk is an Amazon account. A worker is presented with an overview of available task types and a short description of the work to be performed. Workers can preview a typical task before accepting to work on it. A service requester can restrict the participation with respect to the workers' location and task approval rate (the ratio of accepted to all submitted tasks as a measure of past performance). MTurk's business model is based on a 10% commission fee – at least, however, a fee of $0.005 per task. During the last few years numerous other micro-task market platforms for crowdsourcing have been sprouting, e.g. LiveWork[7], Samasource[8], CloudCrowd[9] or Clickworker[10].

Problems regarding poor information quality arise in all domains where knowledge workers have to deal with excessive amounts of information. Many problems are hard to resolve automatically by computers and require hands-on work by humans. Conducting the information quality improvement work manually is however (a) very hard to coordinate and (b) requires a very volatile amount of workers with varying

---

[1] http://www.newsweek.com/blogs/techtonic-shifts/2009/11/13/recaptcha-a-k-a-those-infernal-squiggly-words-almost-done-digitizing-the-new-york-times-archive.html
[2] http://www.gwap.com/gwap/
[3] http://fold.it/portal/
[4] http://phylo.cs.mcgill.ca/eng/index.html
[5] http://eterna.cmu.edu/content/EteRNA
[6] https://www.mturk.com.
[7] https://www.livework.com/
[8] http://www.samasource.org
[9] http://www.cloudcrowd.com
[10] http://www.clickworker.com/en/

skills as there are often masses of data. To give an illustrative example, imagine you want to label several million images. Even if you have many employees committed full time to this task, it requires a lot of time and also a lot of effort to coordinate these tasks. It is therefore easy to envision that crowdsourcing offers a huge potential to organisations by providing a highly flexible and massive amount of workers whenever they are needed. We will show in this paper that there have been many promising individual examples of different types of applications of crowdsourcing to improve information quality presented in the literature and practice.

The rest of the paper is structured as follows: First, the paper discusses the general requirements of tasks to be suitable for crowdsourcing in micro-task markets and we present six potential application areas for improving information quality using crowdsourcing. Moreover, the paper presents a case study, in which we have applied crowdsourcing in one of the identified application areas in the context of medical record transcription in a health insurance company. The paper concludes with a discussion of future research.

General Requirements For Tasks To Use Crowdsourcing In Micro-Task Markets

One central characteristic of crowdsourcing is the high scalability of the resources (their suitability to cope with changes in demand for labor). In order to facilitate high scalability the following general requirements are applicable to the crowdsourcing approach:

- **Low complexity:** Tasks should be of low complexity or at least “formalizable” to allow for a low degree of interaction between workers and the service requester. Otherwise the limited capacity of the service requester will turn out to be a bottleneck for scalability. Crowdsourcing also demands scalable quality management with regard to the quality of the completed tasks; similarly quality management can be automated or crowdsourced itself. If the service requester needs to check all the task results manually, the system can not be considered to be scalable. This again underpins the requirement of highly formalized tasks.

- **Large but varying number of similar tasks:** Due to a scenario specific implementation effort, the micro-task approach pays off only if there is a large number of similar tasks. This – in combination with formalized tasks – allows for systematic optimization and quality control. The number of tasks should vary over time in order to profit from the scalability of the approach.

- **Tasks can be mapped to an electronic (Web-) interface:** As micro-task markets are Web applications, they only work with tasks that can be mapped to an electronic interface.

- **Generally low level of expertise:** The level of expertise required to perform tasks can impose a limit on the achievable scalability and available of the workforce in micro-task markets.

- **Difficult to automate and changing environment:** Typical scenarios are those where automatic solutions fail. This might be the case because the task is (still) generally too complex or because the automatic solution cannot be adapted quickly/cheaply enough to cope with changing circumstances or task specifications. This also includes one-off situations where a large number of tasks have to be performed and developing a software solution is not economically reasonable. Furthermore, it also encompasses tasks where automatic solutions have to be complemented by either human review or rework.

POTENTIAL APPLICATION AREAS FOR IMPROVING INFORMATION QUALITY USING CROWDSOURCING

We have found six different application areas (1) for which we have found that there are IQ problems that could be solved with the use of crowdsourcing in interviews with managers from the industry and (2) for which we have found examples of successful applications of crowdsourcing in literature and practice.

Image processing (visual perception)

IQ Problems in the Industry

In a public transport company we have interviewed, pictures of physical assets in the transportation network have been stored in the database. Pictures of physical assets are a form of unstructured data and cannot be accessed in suitable time if they are not classified in the right way. The same IQ problem can happen with motion pictures. Similarly, a UK based water utility company we interviewed also uses pictures to record the location of their assets. In the aerial pictures, the different type of assets need to be marked as, for example, water pipes, reservoirs, pumping stations etc. A common situation is that the assets are not labeled in the images and the resuting information about the assets is incomplete. Moreover, customers of a European energy company have to fill out forms when they want to register with the supplier, and these forms are handwritten. Automatic solutions cannot correctly capture all the data and carrying out the job manually is a very expensive and time-consuming process. Crowdsourcing can provide potential solutions to these problems.

Crowdsourcing Applications

Typical tasks in image processing are the textual description or categorization of images: Deng et al. [7] introduce ImageNet, an annotated database of 3.2 million images that are hierarchically organized based on the WordNet ontology. The automatically gathered data is verified by human workers on MTurk. Hockenmaier et al. [25] let the crowd describe a given image in one complete but simple sentence. The transcription of single handwritten words [17] or handwritten sentences [20] has also been successfully explored. Automatic OCR software can be complemented by human-based optical character recognition as already successfully performed by ReCAPTCHA. In a geometric reasoning experiment by Corney et al. [6] workers have to pack two-dimensional parts as tightly as possible on a given canvas or select the image that is most representative of a three-dimensional model given images from different perspectives. More complex tasks are those dealing with video or three-dimensional visual representations: Yang et al. [27] let the crowd identify and locate objects in images. A more complex task they suggest is the identification of an activity performed by a person in a video snippet. The game FoldIt[11] asks users to fold protein structures in a three-dimensional space. Corney et al. [6] also rely on the human three-dimensional comprehension when they ask the crowd to select that image that is the most representative of a 3D model given a series of images of the same 3D model from various perspectives. The search algorithms for 3D object retrieval in CAD systems might be enhanced based on a human evaluation [15]. Especially where automatic approaches of image processing are already on the verge of replacing human work, these algorithms can often be trained by human feedback. The German company Clickworker cooperates with Honda to let the crowd train their image recognition systems[12].

Natural language processing (NLP)

IQ Problems in the Industry

In several production companies we found the following IQ problem pattern. Machines are inspected for their condition to determine when is the optimal time to maintain the machines. Problems with the machines identified by the inspections are manually recorded usually in a free text form in the system. We

[11] http://fold.it/portal/
[12] http://www.heise.de/tr/artikel/Eine-Nase-fuer-die-Nische-1027772.html?artikelseite=2

have observed many IQ problems that result from this completely unstructured information, such as database queries failing to retrieve all ‘overheating problems’ because the problem is not always described as ‘overheating’ (it is sometimes described as ‘temperature problems’ etc.). Furthermore, marketing departments in production companies have to make market analysis using the Internet, which is a very time-consuming process and leads to IQ problems related to the accessibility dimension. Extracting information from a website of potential clients, for example, can not usually be done with automated solutions as it requires natural language processing. Again, crowdsourcing could provide efficient solutions to this type of IQ problems.

Crowdsourcing Applications

Typical examples for crowdsourcing applications in the area of natural language processing are language translation or the evaluation of machine translation quality [5]. In the area of language translation specialized providers like myGengo[13] appear to dominate generic micro-task markets. Other applications have been word sense disambiguation and textual entailment[14] [24]. Little et al. [20] let workers on MTurk turn an outline of a document into text or let them change the tense of a given text. Similarly to this, a crowdsourcing task could require workers to translate a free-text description of a problem into one succinct description from a pre-specified list of problems; this could be used to address the ‘overheating problem’ mentioned above. Another practical application is spam protection: Systems like Vipul’s razor[15] use a collaborative filtering for spam prevention. Further examples are the creation of question-answer sentence pairs [16], the ranking of computer generated questions about provided texts [11] or rating Wikipedia articles [18].

Common sense knowledge

IQ Problems in the Industry

Many IQ problems in the industry could be significantly improved with the help of “intelligent” systems. Ontologies can provide a solution by feeding computers with explicit structured knowledge about a domain. This allows computers to make much more sophisticated reasoning. Semantic search, for example in the case of an information communications technology company, has solved many accessibility problems. Building ontologies is, however, a very time consuming process and can be very expensive. Crowdsourcing might offer a much cheaper way to create ontologies by feeding computers with common sense knowledge.

Crowdsourcing Applications

Eckert et al. [8] pose the question whether the wisdom of crowds can be used to create high quality concept hierarchies (ontologies) even in relatively challenging, abstract domains like philosophy. They show that bypassing the large amounts of time and money required for developing and maintaining formal ontologies via crowdsourcing is possible. Amazon uses its own platform MTurk to perform a number of data improvement and validation processes on its own product catalogue like product classification. Gordon et al. [9] let the crowd evaluate common sense knowledge (so called factoids) that has been extracted automatically from news and Wikipedia articles. The obtained ratings are used to improve automatic knowledge extraction systems. Another possible application is ontology matching where concepts from one hierarchy have to be mapped on the concepts of another. As stated by Eckert et al. [8] automatic ontology matching still suffers from severe problems with respect to the quality of matching results.

Information retrieval, search and relevance assessment

[13] http://mygengo.com/

[14] Textual entailment refers to deciding if the meaning of a hypothesis can be inferred from the meaning of a given text.

[15] http://razor.sourceforge.net/

IQ Problems in the Industry

Incorrect customer address data is a commonly encountered IQ problem in the industry. Verifying contact information is often not easy to automate and very costly. But also other types of accessibility problems are common when it comes to retrieving and assessing the relevance of information from the internet, for example, about competitors and trends in the market. Crowdsourcing can solve many IQ problems in this application area.

Crowdsourcing Applications

Examples of relevance assessment are provided by Grady and Lease [10] who ask users to make binary judgments (relevant/non-relevant) for a given query-document pair. Relevance evaluation is also explored by Alonso et al. [3]. Amazon mentions verification and search for addresses as possible applications for its platform MTurk. Letting the crowd perform simple research or information retrieval tasks on the Web is a further application. This way databases could be quickly and cheaply be filled with address data, current offers by competing companies, or publicly available research papers for any specified topic. When inconsistent entries have to be consolidated in a data de-duplication process, the crowd could be asked to identify which of the entries is the outdated one. Contact information could be easily verified via the Web. Given a set of documents and a question, the work of identifying the relevant documents or highlighting relevant paragraphs could be crowdsourced on a micro-task platform. Another plausible application could be to let the crowd answer free-text questions (in the spirit of quora.com) – which has been done by Clickworker.

Sentiment and opinion

IQ Problems in the Industry

Companies are usually interested in knowing what customers think about their products. Sentiment and opinion interpretation from free text in the Web is, however, very difficult to automate and nearly impossible to do manually because of the masses of information that are available on the internet, which is a typical IQ problem.

Crowdsourcing Applications

Hsueh et al. [14] consider the problem of classifying sentiment in political blog snippets using MTurk. Polarity scores are obtained that indicate whether the sentiment is positive or negative. Polarity scores can also be obtained for customer comments expressing opinions about a given topic [23]. Other applications are the capturing of the amount of action indicated by a sentence [21] or the ranking of t-shirt designs according to personal taste [20]. Barr and Cabrera [4] mention Web site reviews and marketing surveys are possible applications. Other possible applications that are based on user opinion are (beta-) testing or writing reviews.

Audio processing

IQ Problems in the Industry

Train operators record audio tracks of the squealing noise that wheels make when contacting with the track. They use this to determine when to replace track or wheels. However, this is unstructured information, which requires a lot of manpower to listen to and determine when the squealing occurs. Information is therefore not available in the format needed.

Crowdsourcing Applications

A typical application of human audio processing is the transcription of voice recordings [22]. The company CastingWords[16] is already specialized on the transcription of audio recordings based on the platform MTurk. Another application that is particularly hard to automate but nearly trivial for humans is

[16] http://castingwords.com/

the classification of a speaker's accent. In the experiment conducted by Kunath and Weinberger [19], the crowd was asked to classify the accent of non-native English speakers as Arabic, Mandarin or Russian. Another possible application may be the verification of sound where the crowd may be asked to record the time at which squealing occurs in a recording thereby aiding the maintainers of train track and train wheels as described in the example above.

CASE STUDY: USING CROWDSOURCING FOR MEDICAL RECORD TRANSCRIPTION IN A HEALTH INSURANCE COMPANY

In this action research case study, we have investigated crowdsourcing to encode medical records based on the ICD [17], an international classification system for diseases and health related problems. The study was building on a project that IBM did together with a private health insurance company and a call center in Germany [26].

A customer of the insurance company has to pay the costs of medical examinations and treatments in advance. After handing in the bills to the health insurance the patient will be refunded, as illustrated in Figure 1.

Among other information, the bill contains a medical diagnosis. These medical diagnoses contain potentially useful information for the health insurance company. Not only can they be used for statistics on diseases or on effects of treatment but also for fraud detection with tremendous commercial relevance: The analysis of medical diagnosis might provide information on whether a certain therapy was justified by the diagnosis or indicate that a patient received drugs for (uncovered) family members. Furthermore, it allows for comprehending a patient's medical 5-year-history, checking on whether a therapy is covered by the patient's insurance-protection or if a medical doctor provided treatment that is out of his/her area of expertise. Long-term data may allow for identifying trends with respect to demographic influences on diseases or the occurrence of diseases over time.

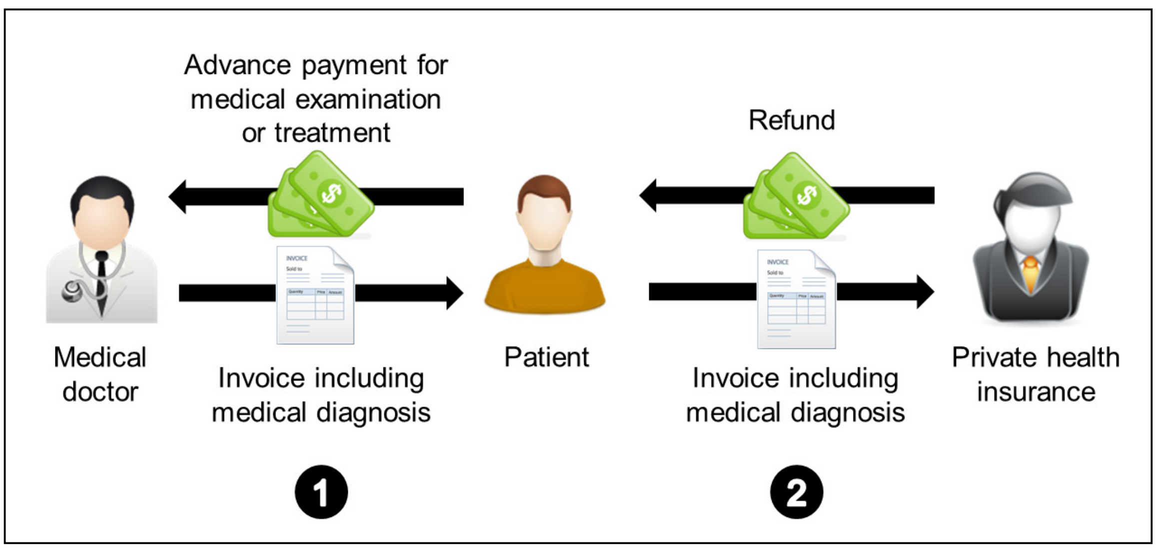

Figure 1. Case Scenario

[17] http://www.who.int/classifications/icd/en/

Despite their potential, medical diagnoses cannot be readily analyzed because they are written in natural language (derived from a free text field) and lack a standardized form, which is a major data quality problem for the insurance company. Furthermore, medical diagnoses tend to be composed of medical terms, domain-specific abbreviations and punctuation as well as other shortened expressions. Thus, data mining or other automated analysis techniques tend to fail.

A recent approach is to encode these medical diagnoses on the basis of the ICD[18] catalogue – the "International Statistical Classification of Diseases and Related Health Problems" by the World Health Organization (WHO). The most recent version is the 10th revision of the ICD (ICD-10[19]). The current German ICD derivative of the international standard is the ICD-10-GM. Diagnoses that have been ICD encoded can then be automatically analyzed. Figure 2 shows the activities that have to be carried out in order to encode each diagnosis: The diagnosis text has to be read and understood. This includes "mentally partitioning" the given diagnosis into subdiagnoses each requiring a separate ICD code, identifying signal words like "exclusion of", interpreting abbreviations in the given context, extraction of the parts of information that are relevant for the encoding.

In the scenario of the health insurance company, the medical diagnoses are already available as machine encoded text (albeit still being in natural language). The insurance company employs a software program to automatically assign ICD codes to the free text. The software program has not proven to be fully reliable: The exclusion of a disease has repeatedly been falsely identified as the presence of that disease. Only if the software is unable to extract any code at all (which happens in about 25% of the cases), a manual reworking is triggered. The manual rework is performed by specialized personnel – one of two teams with education as medical assistant – and (due to technical limitations) up to four diagnoses are identified for each free text. Every month about 10,000 invoices have to be manually reviewed and encoded. On average about two ICD codes are assigned to a diagnosis. There are some problems associated the way the encoding is done so far: As mentioned above, the software program is not yet able to identify all relevant codes and might still include erroneous codes. The medical personnel incur cost and the manual rework is tedious. Additionally, human workers tend to rely on a subset of well memorized codes which in turn might increase the risk that workers favor these over other less well-known codes.

[18] http://www.who.int/classifications/icd/en/
[19] http://apps.who.int/classifications/apps/icd/icd10online/

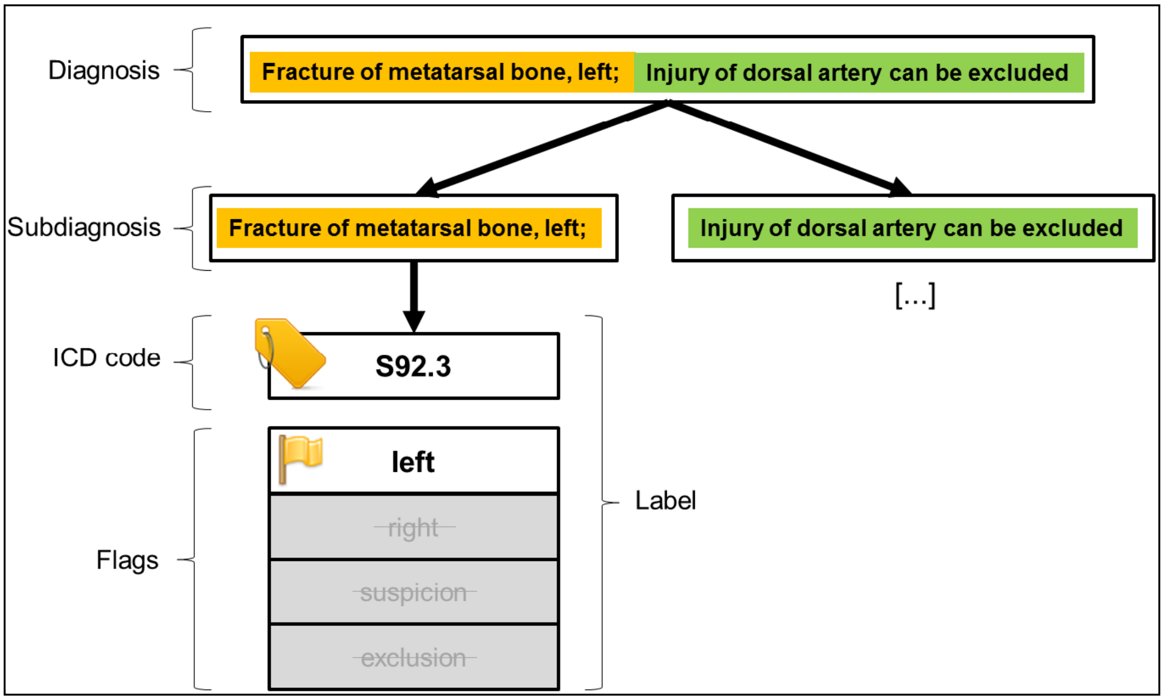

Figure 2. Medical Record Transcription

A conceivable alternative to the software and in-house encoding is, therefore, to outsource the work via a crowdsourcing platform, see Figure 3. We have used two different platforms in this case study: (1) Amazon's Mechanical Turk and (2) we cooperated with a call center and used call center agents in their idle time. In the second case, we have used a Java based crowdsourcing platform that we have developed as part of our project. In order to improve the quality of the encoding results, we allowed the system to issue the same encoding task to multiple workers and we developed an algorithm to automatically assess the quality of the delivered results based on the workers' previous quality as well as the congruence of results (in case tasks have been published redundantly). Moreover, workers were using the encoding tool Semfinder Online that supported their transcription activities. The overall results were very encouraging. The work has been done quickly and with high quality. To our surprise, we have experienced no misuse, for instance, workers did not attempt to cheat and earn credits by entering random work results. Workers had not more problems regarding understanding of the tasks they were asked to perform than conventional workers.

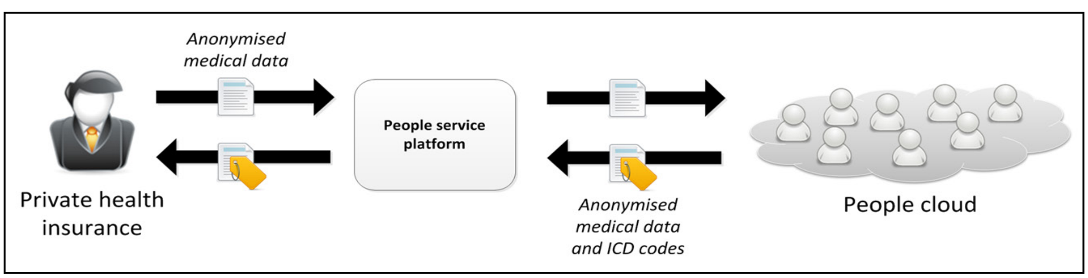

Figure 3. Using Crowdsourcing for Medical Record Transcription

Compared to the status quo the ICD encoding via crowdsourcing promises to bring along the following advantages from an insurance company's point of view.

1. An improved and structured data foundation enables an automatic analysis of the medical records.
2. In the case of fraud detection, suspicious medical records could then be individually investigated after being flagged by an automatic fraud detection software system. The capability of detecting fraud more easily can lead to a cost reduction for health insurance and patients.
3. ICD-encoded patients' health records would allow for enhanced risk profiles and more sophisticated analyses might enable more flexible tariff models as it is now possible to automatically detect medical services not being covered and requests for refund can be rejected. In case of fraud, the insurance company can withdraw from the contract.
4. Crowdsourcing the ICD-encoding work also allows employees to focus on core tasks and competencies.
5. Compared with a software system, a crowdsourcing platform promises to provide improved encoding quality – at least if it is combined with an appropriate automatic quality control.
6. Crowdsourcing is very useful in this kind of scenario, because the demand for labor fluctuates. Seasonal variations in the number of diagnoses can be counterbalanced via crowdsourcing.

CONCLUSION

This paper investigates potential applications of crowdsourcing for information quality improvement. We have done this by providing application examples from practice and the literature and by grouping these examples along six application areas. We have also demonstrated in an action research study in one of the application areas, i.e. natural language processing, that it is feasible to use crowdsourcing to solve IQ problems and that it can generate very encouraging results. The paper shows that crowdsourcing can be often a good alternative or complement to current data quality software tools used in the industry. Future research should explore additional application areas for crowdsourcing to improve IQ. Moreover, it is necessary to validate the usefulness of crowdsourcing in the examined application areas in further action research studies in the industry.

ACKNOWLEDGEMENTS

This research has been partly funded by EPSRC project "Information Quality in Asset Management", reference number EP/G038171/1.

REFERENCES

[1] Von Ahn, L., Blum, M., Hopper, N., and Langford, J., "CAPTCHA: Using hard AI problems for security," *Advances in Cryptology—EUROCRYPT 2003*, 2003, pp.646–646.

[2] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M., "recaptcha: Human-based character recognition via web security measures," *Science*, 321 (5895), 2008, p.1465.

[3] Alonso, O., Rose, D.E., and Stewart, B., "Crowdsourcing for relevance evaluation," ACM SIGIR Forum*42*, 2008, pp.9–15.

[4] Barr, J., and Cabrera, L.F., "AI gets a brain," *Queue*, 4 (4), 2006, pp.24–29.

[5] Callison-Burch, C., "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk," Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, 2009, pp.286–295.

[6] Corney, J.R., Torres-Sánchez, C., Jagadeesan, A.P., Yan, X.T., Regli, W.C., and Medellin, H., "Putting the crowd to work in a knowledge-based factory," *Advanced Engineering Informatics*, 2010.

[7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database" 2009.

[8] Eckert, K., Niepert, M., Niemann, C., Buckner, C., Allen, C., and Stuckenschmidt, H., "Crowdsourcing the

Assembly of Concept Hierarchies," Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL), Brisbane, Australia. ACM Press, 2010.

[9] Gordon, J., Van Durme, B., and Schubert, L.K., "Evaluation of commonsense knowledge with Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.159–162.

[10] Grady, C., and Lease, M., "Crowdsourcing document relevance assessment with Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.172–179.

[11] Heilman, M., and Smith, N.A., "Rating computer-generated questions with Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.35–40.

[12] Howe, J., "Crowdsourcing: A Definition," 2 June 2006, <http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html> (18 February 2010).

[13] Howe, J., "The rise of crowdsourcing," *Wired Magazine*, 14 (6), 2006, pp.1–4.

[14] Hsueh, P.Y., Melville, P., and Sindhwani, V., "Data quality from crowdsourcing: a study of annotation selection criteria," Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, 2009, pp.27–35.

[15] Jagadeesan, P., Wenzel, J., Corney, J.R., Yan, X.T., Sherlock, A., Torres-Sanchez, C., and Regli, W., "Validation of purdue engineering shape benchmark clusters by Crowdsourcing," Proceedings of the International Conference on Product Lifecycle Management, 2009.

[16] Kaisser, M., and Lowe, J.B., "Creating a research collection of question answer sentence pairs with amazons mechanical turk," Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC), 2008.

[17] Kern, R., Thies, H., and Satzger, G., "Statistical quality control for human-based electronic services," *Service-Oriented Computing*, 2010, pp.243–257.

[18] Kittur, A., Chi, E., and Suh, B., "Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements," *Proc. CHI 2008*.

[19] Kunath, S.A., and Weinberger, S.H., "The wisdom of the crowd's ear: Speech accent rating and annotation with Amazon Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.168–171.

[20] Little, G., Chilton, L.B., Goldman, M., and Miller, R.C., "Turkit: tools for iterative tasks on mechanical turk," Proceedings of the ACM SIGKDD workshop on human computation, 2009, pp.29–30.

[21] Madnani, N., Boyd-Graber, J., and Resnik, P., "Measuring transitivity using untrained annotators," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.188–194.

[22] Marge, M., Banerjee, S., and Rudnicky, A.I., "Using the Amazon Mechanical Turk for transcription of spoken language," Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp.5270–5273.

[23] Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M.R., and Banchs, R., "Opinion mining of spanish customer comments with non-expert annotations on Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.114–121.

[24] Parent, G., and Eskenazi, M., "Clustering dictionary definitions using Amazon Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.21–29.

[25] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J., "Collecting image annotations using Amazon's Mechanical Turk," Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010, pp.139–147.

[26] Satzger, G., Kern, R., Pasing, A., Pfau, G., Fritz, R., Alexander, S., and Roland, P., "People Services: Efficient work processes leveraging the power of the crowd," presented at 20th Annual Frontiers in Service Conference, 2011, Columbus, Ohio, USA.

[27] Yang, Y., Zhu, B.B., Guo, R., Yang, L., Li, S., and Yu, N., "A comprehensive human computation framework: with application to image labeling," Proceeding of the 16th ACM international conference on Multimedia, 2008, pp.479–488.