Crowdsourcing Research Opportunities: Lessons from Natural Language Processing

Marta Sabou
Dpt. of New Media Technology
MODUL University Vienna
Marta.Sabou@modul.ac.at

Kalina Bontcheva
Natural Language Processing Group
Dpt. of Computer Science
University of Sheffield
K.Bontcheva@dcs.shef.ac.uk

Arno Scharl
Dpt. of New Media Technology
MODUL University Vienna
Arno.Scharl@modul.ac.at

ABSTRACT

Although the field has led to promising early results, the use of crowdsourcing as an integral part of science projects is still regarded with skepticism by some, largely due to a lack of awareness of the opportunities and implications of utilizing these new techniques. We address this lack of awareness, firstly by highlighting the positive impacts that crowdsourcing has had on Natural Language Processing research. Secondly, we discuss the challenges of more complex methodologies, quality control, and the necessity to deal with ethical issues. We conclude with future trends and opportunities of crowdsourcing for science, including its potential for disseminating results, making science more accessible, and enriching educational programs.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Collaborative computing; I.2.1 [**Applications and Expert Systems**]: Games; I.2.6 [**Learning**]: Knowledge Acquisition; I.2.7 [**Natural Language Processing**]

General Terms

Measurement, Experimentation, Human Factors, Languages, Verification

Keywords

Crowdsourcing, Games with a Purpose, Resource Acquisition, Natural Language Processing

1. INTRODUCTION

The notion of citizen science, "a form of collaboration that involves the public in scientific research to address real-world problems" [50], has its roots in the early 19th century. The annual Christmas bird count organized by the National Audubon Society began in 1900 and to date leverages on the yearly contributions of 60,000 - 80,000 volunteers [9]. The emergence of the Internet and Web 2.0 technologies significantly lowered the cost of user participation and lead to citizen science projects that are entirely "virtual" [51]. Crowdsourcing techniques allow outsourcing a task to "an undefined, generally large group of people in the form of an open call" [21] and as such are the main techniques underpinning virtual citizen science projects.

Although crowdsourcing of scientific work is a natural continuation of citizen science projects and has lead to important discoveries already, many scientists still regard these type of approaches with a certain amount of skepticism [18], especially in those fields where relying on citizen scientists was not a common practice beforehand (e.g., molecular biology). Such skepticism is somewhat justified by the fact that lessons learned from crowdsourcing projects are often discussed within the boundaries of the specific scientific discipline and that crowdsourcing research opportunities for science have so far received little attention. Indeed, we are aware of several broad surveys, which overview human computation and crowdsourcing systems in general, employed both in research and commercial environments, and which aim to identify the critical issues for these systems and to organize those into meaningful taxonomies [11, 30, 40]. Other surveys are specific to a given domain and crowdsourcing approach, for example the use of games with a purpose for knowledge acquisition [46] or human computation [25]. Within the area of natural language processing (NLP), [7] provides an overview of 24 experiments in the area of NLP using crowdsourcing marketplaces, in particular Amazon Mechanical Turk (Mturk). [38] is broader in scope as it surveys both MTurk and game-based approaches and overviews over 30 diverse works. However, the limitation here is focusing on speech related works alone. Finally, [49] discuss the benefits and disadvantages of different crowdsourcing approaches (games, MTurk, volunteering) for NLP tasks. Savage discusses crowdsourcing approaches in various scientific disciplines, but focuses exclusively on game-based projects [42].

The contribution of this paper is in defining the transformative impact of crowdsourcing on NLP research, followed by a discussion of lessons learnt and outstanding research opportunities. We chose to focus on NLP, since it is a research field that naturally benefits from human language skills and where a rich repertoire of crowdsourcing approaches have already been implemented. Our observations are derived from

i-Know '12, Sep 05-07 2012, Graz, Austria
ACM 978-1-4503-1242-4/12/09.

our research into crowdsourcing for acquisition of linguistic resources [41, 44] a current crowdsourcing project in the climate change domain[1] [43], and the related body of knowledge in the field of NLP.

We start with considerations about how crowdsourcing is used to support science in general in Section 2. In Section 3 we describe the main impacts that crowdsourcing had on NLP research, before concluding with a discussion of future challenges for crowdsourcing in NLP and beyond.

2. CROWDSOURCING IN SCIENCE

The use of crowdsourcing techniques has impacted a broad range of science-related projects. It is therefore important to clarify the focus of this paper, to delimit the types of efforts we will cover and to differentiate them from others. Our aim is to focus primarily on *crowdsourcing efforts that have as their primary goal to gather data/resources from non-expert contributors in order to support scientific investigations.*

We will therefore not deal with crowdsourcing projects that do not have scientific data creation as their main goal, yet have created resources that are often used by researchers. A good example is Wikipedia, the crowdsourced online encyclopedia, which is frequently used by NLP researchers as an additional source of knowledge. We will also not address projects that rely on crowd participation to support auxiliary scientific processes such as bibliography management (e.g., Mendeley). Projects that recruit participants using ICT solutions such as social media but perform the data collection and experimentation in a closed-laboratory setting, are also out of our scope. Volunteer computing projects where participants contribute their computing resources rather than their own effort (e.g., SETI@home) will also not be discussed. Finally, we also did not consider distributed collaborative environments that support the curation and production of scientific data, but where contributors are only scientists and supporting professionals as opposed to non-expert contributors drawn from the crowd [14]. Note that we focus on a specific type of citizen science projects, namely those that are entirely mediated by ICT and categorized as "Virtual" projects in [51].

2.1 Types of Crowdsourcing Approaches

The use of crowdsourcing in science typically focuses on tasks with an Artificial Intelligence (AI) flavour. It includes a number of genres, which are typically classified along various dimensions, such as the motivation of human contributors (e.g., fun vs. altruism vs. payment), the way in which individual results are aggregated and how quality is managed. The three key crowdsourcing genres most widely adopted by the scientific community are [40]:

Mechanised labour (e.g., Amazon's Mechanical Turk) is a type of paid-for crowdsourcing, where volunteers choose to carry out small tasks and are paid a small amount of money in return (often referred to as micro-payments). Each such small task is called Human Intelligence Task (HIT).

Games with a purpose (GWAPs) [48] enable human contributors to carry out computation tasks as a side effect of playing online games. An example from the area of computational biology is the Phylo game[2] that disguises the problem of multiple sequence alignment as a puzzle like game thus "intentionally decoupling the scientific problem from the game itself" [23]. The challenges in using GWAPs in scientific context are in designing appealing games and attracting a critical mass of players.

Altruistic crowdsourcing refers to cases where a task is carried out by a large number of volunteer contributors. To reduce the incentive to cheat (e.g., for money or glory), altruistic crowdsourcing approaches leverage the intrinsic motivation of a community interested in a domain. The Galaxy Zoo[3] project, for example, seeks volunteers with a latent desire to help with scientific research for classifying Hubble Space Telescope galaxy images. The project has attracted more than 250K volunteers which provided over 150M galaxy classifications. The resounding success of this project, prompted the generalisation of the infrastructure created for Galaxy Zoo into a platform, named Zoonivers[4], where other similar, "citizen science" projects can be deployed. To date the platform offers a range of astronomy related projects and boasts a base of over 430K volunteers.

2.2 Typical Uses of Crowdsourcing

Crowdsourcing techniques can be used to support various stages of the scientific process. Firstly, crowdsourcing projects often produce data and other resources that are used as an input for designing or training algorithms or for clarifying a scientific hypothesis. Secondly, human problem-solving has been shown to be often more effective in certain problems than pure computation and that it can offer valuable contribution to scientific algorithms (i.e., by cutting down unpromising search trees in the solution space). Finally, some approaches employ human intelligence for evaluating the results produced by algorithms. We will discuss these different types of contributions to science and give relevant examples for each category.

Many scientific disciplines employ machine learning algorithms which, to be effective, rely on large, unbiased training data sets. For example, this is the case for NLP (Section 3.1) and also for the field of visual computing where image-processing algorithms support exploration within a wide range of scientific disciplines. There are various ways to collect such data sets. Firstly, participants might be asked to contribute data according to an established protocol. This is one of the classic methods used by citizen science projects. For example, the Great Sunflower Project[5] (which has attracted over 80,000 participants) asks participants to follow a given form-based protocol when reporting about their gardens and the observed activity of bees. Similarly, The Open Mind Common Sense project[6] collects general world knowledge from volunteers in multiple languages using as structured format and is a major source for the ConceptNet semantic network that can enable various text understanding tasks. Form-based data acquisition methods are easy to implement in crowdsourcing marketplaces such as MTurk and are routinely used to crowdsource data collection, in particular in the NLP field [8].

A second approach to collect data sets is to ask participants for completing structured recognition and classification tasks. Human visual skill, for example, is often used to recognize and label image content, as well as to classify

[1] http://www.ecoresearch.net/triple-c
[2] http://phylo.cs.mcgill.ca
[3] http://www.galaxyzoo.org
[4] http://www.zooniverse.org
[5] http://www.greatsunflower.org
[6] http://openmind.media.mit.edu

images. This kind of collection methods can also leverage other human capacities such as language skills (e.g., recognize the accents in spoken language [36]). Classification and labeling tasks can be implemented on crowdsourcing marketplaces, can be disguised as games or can be part of altruistic crowdsourcing projects such as Galaxy Zoo, where volunteers classify Hubble Space Telescope galaxy images.

Surveys are popular research instruments for clarifying a scientific hypothesis. It has been shown that, compared to the traditional approach of recruiting participants from a university's student base, crowdsourcing techniques lead to faster completion times, produce data that has a similar or even better quality and allow access to a participant base that is older, more ethnically diverse and has more work experience than the student population [3]. These are important factors for fields such as organizational research [3].

Some crowdsourcing projects aim not only to provide training data to algorithms, but also to harness human problem-solving abilities to help design and run computational algorithms. This is particularly true in the case of those algorithms that are computationally prohibitive, as they need to sieve through a solution space that grows exponentially with the size of the input. Human visual pattern recognition skills can help guide the algorithm away from unpromising regions of the solution space. For example, *Foldit*[7] [10] is a game in which players fold proteins and where researchers have used the best human tactics to further develop Rosetta, their protein folding algorithm [18]. Other projects that make use of human visual problem-solving skills are *EteRNA*[8] for detecting amino acid sequences that could fit best a given protein shape and Phylo, for aligning DNA sequences to study genetic diseases by asking volunteers to refine preliminary work performed by computers. From 2010, 35K people played Phylo and improved 70% of the given alignments [23]. We describe additional strategies for supporting and evaluating NLP algorithms in Section 3.5.

3. IMPACT OF CROWDSOURCING ON NLP RESEARCH

In the past five years, the use of crowdsourcing in the NLP field has intensified. For example, a search in the ACL Anthology for the term "Mechanical Turk" returned 128 results in November 2010 [16] while less than two years later, in June 2012, the same search returned 361 results, i.e., denoting that the volume of work using mechanised labour alone has almost tripled. In more detail, Figure 1 extends a subset of the bibliographic analysis in [16] by depicting the number of papers using MTurk that were published at the main NLP conferences in the last 5 years. Note that LREC and COLING were not held in 2011 as they are biennial conferences, and that at the time of writing (June 2012) we could only access the full papers presented at LREC, as the other conferences are held later in the year. The graph shows a significant increase in the number of papers that use mechanised labour, as all conferences publish 3 (LREC) to 7 (ACL) times more papers in this area as in previous years. Next, we discuss in what ways crowdsourcing has transformed the NLP research field.

[7] http://www.fold.it
[8] http://eterna.cmu.edu

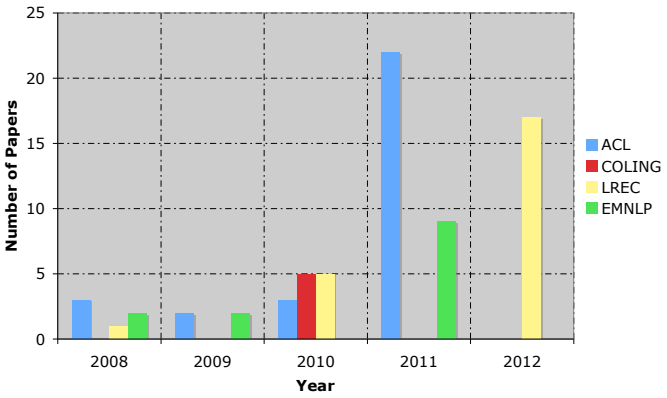

Figure 1: Number of papers that use MTurk in the major NLP conferences. 2008-2010 values from [16].

3.1 Affordable, Large-Scale Resource Acquisition

Linguistic resources such as (annotated) corpora or lexicons are core to the development of the NLP field. Indeed, over the past ten years, NLP research has been driven forward by a growing volume of annotated corpora, produced by evaluation initiatives such as the Document Understanding Conferences (DUC[9]), the Text Analysis Conferences (TAC[10]), SemEval and Senseval[11]. These corpora have been essential for training and domain adaptation of NLP algorithms and their quantitative evaluation, as well as in enabling algorithm comparison and repeatable experimentation. Yet, traditional expert-driven and tightly controlled corpus creation methodologies tend to be very expensive to implement, both in terms of time required to produce high-quality corpora of significant size, and in price per word annotated. The latter can vary between $0.36 and $1.0 [39, 52], which is unaffordable for corpora consisting of millions of words. Poesio and colleagues conclude that some NLP tasks such as training parsers are actually limited by the 1M word limit and they would ideally benefit from 100-million-word corpora (100M), which are simply out of the reach of the traditional methodologies which rely on expert scientists [39]. Instead, these are better solved either through active annotation (where manual annotation addresses only the needs of the system being trained rather than annotating the entire corpus) or Web collaboration approaches such as crowdsourcing. Therefore, cost reductions in resource acquisition are among the core motivations behind crowdsourcing in the NLP research field.

Commercial crowdsourcing marketplaces have been reported to be 33% less expensive than in-house employees on tasks such as tagging and classification [19]. Consequently, NLP researchers have started experimenting with MTurk and game-based approaches as less expensive alternatives for the creation of language resources through distributed human effort. Callison-Burch and Drezde have organized a workshop [8] where participants were offered a $100 voucher to crowdsource some NLP related task on MTurk. The reports from the workshop provide a first proof of the variety

[9] http://duc.nist.gov
[10] http://www.nist.gov/tac
[11] http://www.senseval.org

of resources that can be created with such a small sum [8]. In [39], the authors take a closer look at the cost reductions enabled by crowdsourcing, but unlike [8] which report on small-scale resources, they investigate the case of large resources, on the scale of 1M tokens. They estimate that, compared to the cost of expert-based annotation (estimated as $1.000.000), the cost of 1M annotated tokens could be indeed reduced to less than 50% by using MTurk (i.e., $380,000 - $430,000) and to around 20% of the expert-based approach's price (i.e., $217,927) when using a game based approach, such as their own PhraseDetectives game [39].

3.2 Diversification of Research

A direct consequence of the cost reductions enabled by crowdsourcing is that it is relatively cheap to produce resources needed for solving a broad range of NLP tasks. Therefore the field is not confined anymore to solving core tasks for which the manual creation of large resources has been funded but can also focus on specialized, niche tasks.

A first dimension of diversification consists of the languages for which resources can be produced. One advantage of MTurk is that it allows "access to foreign markets with native speakers of many rare languages" [52]. This feature is particularly useful for those that work on less-resourced languages such as Arabic [13], Urdu [52] or others [2, 6, 22]. Irvine and Klementiev, for example, have shown that it is possible to create lexicons between English and 37 out of the 42 low resource languages that they experimented with [22]. There are also many researchers who deliberately create their crowdsourcing applications in a way that they can be easily re-used across languages. Examples include [22, 29] for MTurk and [39, 41] for GWAPs.

A second dimension of diversification is the possibility to study a variety of resource types besides news-wire text. Some researchers used crowdsourcing to create corpora of novel or special types of texts for which no resources are yet available, such as emails [28], Twitter feeds [15] or augmented and alternative communication texts [47]. Going beyond text, a recent survey shows an increasing use of paid-for crowdsourcing for speech corpora, in particular eliciting speech transcriptions [26], speech-accent ratings [36], and assessment of dialog systems [38] thus lowering the traditionally high acquisition barrier for speech based resources.

Low corpus creation costs also enable studying niche language phenomena. Munro and colleagues investigate the use of crowdsourcing for supporting linguists (and psycholinguists) to experimentally investigate language processing and linguistic theory [32]. They report on seven MTurk experiments, many of them reproducing "classic" experiments that were earlier performed only in lab conditions, but at the fraction of the lab's price. These experiments include (i) evaluating semantic transparency of verbs, (ii) investigating statistical word segmentation in audio files, (iii) contextual predictability studies also known as Cloze task, (iv) judgment studies of fine-grained probabilistic grammatical knowledge and (v) confirming corpus trends. A major differentiating feature of this class of experiments is that they aim to support the study of gradient phenomena where there is no right answer, and where the interest is mainly in the distribution of answers over participants rather than specific data points. These tasks also typically lack any gold standard to measure against, thus making quality control more cumbersome. Thanks to employing a set of pre-task quality assurance measures, Munro and colleagues find that the quality of crowdsourced results is often comparable to those obtained under controlled experiments (or sometimes even higher). They conclude that crowdsourcing allows investigating a broader range of linguistic phenomena, and willenable a more "expanded and dynamic NLP repertoire".

Crowdsourcing has also played a role in advancing research on tasks that are inherently subjective and for which no or only small-scale resources are available, since these resources are hard to create semi-automatically. These tasks include sentiment detection [41], translation [52], word sense disambiguation [37], anaphora resolution [39], question answering [31], textual entailment [33] and summarization [13].

3.3 Contributor Selection and Training

Crowdsourcing also changed significantly some scientific practices and methodologies in NLP. In particular, corpus acquisition methodologies were changed to accommodate new strategies for contributor selection (described here) and result aggregation and quality control (Section 3.4).

Traditional expert-based corpus acquisition projects typically hire and train expert contributors and managers in preparation for the resource creation phase [20]. The goal is to ensure that the contributors have the right expertise and understand the task at hand. Beyond the area of NLP, in many social science experiments understanding and controlling the demographics of the survey-population is of key importance. Yet this is difficult to achieve within virtual environments where, due to privacy issues, relatively little is known about the user. Therefore, crowdsourcing projects must invest significantly more thought and work into contributor selection and training than traditional approaches. Current techniques for this include screening, training and profiling contributors, as discussed next.

One of the experimentally grounded measures for *a priory* making crowdsourcing tasks more resistant against cheaters relates to controlling the *composition of worker crowd* [12]. Indeed, projects that run on crowdsourcing marketplaces routinely screen (filter) the workers prior to the task based on their previous performance (measured as the acceptance rate of their work for previous tasks), geographic origin, and initial training. GWAPs and altruistic crowdsourcing projects do not usually have this opportunity since most often their user community is not known a-priory, with the exception of Facebook-hosted games.

Training mechanisms ensure that the selected contributors understand the task at hand and acquire a basic skill for performing it. For example, the PhraseDetectives game includes a user training stage in which players' answers are contrasted with a Gold Standard and they are offered feedback on their answers to help them train for the task [39]. A user rating is derived which is used to determine whether the player needs more training. A training level is also used by the game supporting the GIVE challenge. In the area of mechanised labor, besides providing concise but precise instructions, many projects embed positive (and/or negative) gold standard examples within their tasks to determine the quality of data provided by each worker. While in the case of MTurk this data can only be used after the completion of the task (to exclude low-performing worker's data), CrowdFlower (crowdflower.com) the main competitor to MTurk, offers immediate feedback to workers when they complete a "gold"-unit, thus continuously training them for the task.

Besides the few and generic worker details offered by crowdsourcing marketplaces, NLP projects additionally require insight into the workers' linguistic knowledge (e.g., the languages they speak and to what level). This information is often collected as part of the HITs - e.g., whether workers are native speakers, for how many years they speak a language [22, 52]. These details are part of a more detailed profile for each worker in order to better judge the quality of their work. In their pioneering work, Snow et al proposed a probabilistic model to correct annotator bias for categorical data, which allows modeling the reliability and bias of individual workers (as some embryonic profiling) and subsequently correcting them [45]. Ambati and colleagues replace an inter-annotator agreement based quality check with a technique where the reliable translators are identified and their decisions veto the cases when there is no agreement between turkers [2]. Zaidan and Callison-Burch use information about worker language abilities and location (collected through a questionnaire) as features in a model that selects the best translations for a given sentence [52]. Additionally, they compute worker competency (as another profile value) by comparing all crowdsourced translations against an available gold standard of professional translations. This feature plays a role when evaluating all their translations since reliable workers tend to provide good data, while cheaters have a consistent misbehavior.

3.4 Aggregation and Quality Control of Multiple, Noisy Annotations

Traditional, expert-driven methodologies for corpus creation have a rather straightforward process for aggregating the data provided by a small number of expert annotators and to create the final corpus [20]. They usually assess annotator performance over time, inter-annotator trends, and corpus characteristics (imbalance, sufficient size). In the crowdsourcing case, the challenge lies in aggregating the multiple, noisy contributor inputs to create a consistent corpus. Quality control plays an important role in this final stage as only contributions that surpass a quality threshold are selected for inclusion into the final resource.

Broadly speaking, there are two main approaches for aggregating contributions. Firstly, statistical processing is used to identify outliers and exclude them, both at the level of contributors and contributions. Majority vote is the simplest of such statistical measures. For example, Poesio and colleagues use voting to determine which markable, in an anaphora annotation task, is complete and require all players to agree [39]. A more complex technique is developed by [52] who train a model for selecting a translation that is likely to be the best from the total of 14 translations of each sentence. Their model takes into account features of the translated sentence (trying to discriminate between good and bad English sentences), worker-level features including their language abilities and location as well as features derived from the rankings of the translations. Only the best translations are selected for the final translation set.

A second set of approaches uses contributor-based, as opposed to statistical, measures to validate and select the best contributions of their peers, thus introducing a completely novel approach to result validation and aggregation than traditional methodologies. This is achieved by combining the basic data collection task with a verification task in a so-called "create-verify" workflow pattern. For example, Callison-Burch routinely includes a second, verification task following a data creation task, when creating multiple reference translations or collecting human-mediated translation edit rate corpora [6]. Negri and Mehdad organize the task of translating English hypothesis into Spanish into *translation-validation cycles* [34]. Translations judged correct with a confidence over 0.8 are added to the corpus, the rest are translated again, until they are approved by a majority vote.

3.5 Evaluation and Algorithmic Support

The performance of new algorithms on many NLP tasks (e.g. text summarization, natural language generation, machine translation) is typically best evaluated by human experts, following extensive guidelines and training. Researchers have recently turned to crowdsourcing in order to make such evaluations less costly to carry out. The main challenge here lies in defining the evaluation task, so that it can be crowdsourced with high quality results.

One innovative use of GWAPs is the GIVE challenge, which uses a treasure hunt 3D game to test and compare the output of Natural Language Generation (NLG) systems [24]. The tested systems must generate, in real time, natural language instructions that guide players within a virtual world. The game-based and lab-based experiments agree on the major differences between key features of the compared systems. Also, the crowdsourced evaluation finds considerably more differences and allows more fine-grained analysis of the data, since much more data is collected [24].

In contrast, the use of crowdsourcing marketplaces for evaluation has lead to mixed results. Gillick and Liu found that non-expert evaluation of summarization systems produces noisier results thus requiring more redundancy to achieve statistical significance and that MTurk workers cannot produce score rankings that agree with expert ranking [17]. They suggest crowdsourcing evaluation tasks to be designed differently from expert-based evaluation forms. This has indeed been shown to work well in crowdsourcing reading comprehension evaluation of machine translation, where a 4-phase workflow of tasks was used [6].

An emerging role of crowdsourcing is in supporting the automated algorithms by providing human input into hard-to-solve cases (Section 2.2). In the NLP field, research has focused on active learning, e.g. for sentiment classification [5] and named entity annotation [27]. These approaches leverage machine classifiers to predict which samples are the most informative (e.g., by measuring disagreement between multiple classifiers) to reduce the number of crowdsourced judgments. None of these approaches, however, makes use of crowdsourcing platforms (beyond the recruitment of contributors [27]), since the real-time interaction between the annotators and active learning requires a custom-built interface. On the one hand, the integration of human input within the algorithmic computations reduces significantly the amount of human input required, thus making crowdsourcing even more cost- and time-effective. On the other hand, it is far from straightforward and incurs additional interface design costs and time overheads.

3.6 Legal and Ethical Issues

The use of crowdsourcing in science raises three issues of legal and ethical nature, which have so far not received sufficient attention. The first one is how to acknowledge properly crowd contributions, i.e. having the "Crowd" as an

additional author. While no clear guidelines exist about this issue, some volunteer projects (e.g., FoldIt, Phylo) already include contributors in the authors' list [10, 23].

The second issue is contributor privacy and wellbeing. Paid-for marketplaces (e.g. MTurk) go some way towards addressing worker privacy, although these are far from sufficient and certainly fall short with respect to protecting workers from exploitation, e.g. having basic payment protection [16]. The use of mechanised labour (MTurk in particular) raises a number of worker right issues [16]: low wages (below $2 per hour), lack of worker rights, and legal implications of using MTurk for longer term projects.

The third issue is licensing and consent, i.e. making it clear to the human contributors that by carrying out these tasks they are contributing knowledge for scientific purposes and agree to a well-defined license for sharing and using their work. Typically open licenses, such as Creative Commons are used and tend to be quite prominently stated in volunteer-based projects/platforms [1]. In contrast, GWAPs tend to mostly emphasize the scientific purpose of the game, while many fail to state explicitly the distribution license for the crowdsourced data. In our view, this lack of explicit consent to licensing could potentially allow the exploitation of crowdsourced resources in a way, which their contributors could find objectionable (e.g. not share a new, GWAP-annotated corpus freely with the community). Similarly, almost one third of psychology reviews on MTurk post no informed consent information at all [3].

4. CONCLUSIONS AND FUTURE TRENDS

Research to date has shown that crowdsourcing is slowly revolutionising NLP research by significantly reducing the cost of acquiring linguistic resources, as well as by directly supporting algorithms and their evaluation. This has come at the expense of new, more complex resource creation methodologies, in particular for contributor management, result aggregation and quality control. Table 1 classifies the NLP approaches described in this paper in terms of the HC genre that they use, the research diversification type they enable, as well as the stage of the scientific process that they support. It is evident that, by large, mechanised labour is the most popular HC genre among NLP researchers, while altruistic crowdsourcing is the least frequently used approach. In terms of their motivation, most works aim to solve some task that is subjective and therefore amenable to human computation. Also, we conclude that acquiring input and training data sets is a major goal, as apposed to using HC for algorithmic support or evaluation. Nevertheless, so far NLP researchers have mostly crowdsourced small- to medium-sized projects. Scaling up is still a major challenges, as are the following challenges and future trends, which are valid for crowdsourcing in science as a whole.

Promoting Learning and Science. Learning and self-improvement through participation in crowdsourcing projects are a major, untapped opportunity and a powerful incentive mechanism. The newly released DuoLingo (duolingo.com) game is an important effort in this direction, i.e. contributors are trained in a new language, while helping with translation tasks. Another under-explored approach with huge potential is integrating scientific crowdsourcing projects with social networks. The added advantage here is that the increased visibility and ease of engagement, if harnessed successfully, could contribute to making STEM (science, technology, engineering, and mathematics) research more attractive and understandable, and hopefully motivating young people to take up science as their chosen career path.

Motivating Users. Typically the motivation mechanisms that underlie crowdsourcing projects in general are adopted unchanged in crowdsourcing projects with scientific purposes. This assumption, however, might be an oversimplification. For example, Nov and colleagues [35] discuss some fundamental differences between general crowdsourcing projects and those geared towards supporting scientific exploration such as the main beneficiary (e.g., the contributing crowd community vs. the scientists) and the visibility of each individual contribution (i.e., immediately published, clearly identifiable contribution associated with the contributor vs. small, unidentifiable part within a scientific work that is published several months after the contribution was made). Nov et al. argue that these differences impact the motivational factors of contributors. Even within crowdsourced science projects, Nov et al. find that the task granularity (i.e., whether only resources are contributed once or frequent and detailed contributions are required) also has an impact on motivation, as finer-grained tasks require substantially more effort [35].

Addressing Ethical and Legal Challenges. Similarly to paid-for marketplaces, volunteer- and game-based projects need to implement appropriate safeguards and warnings to ensure that no personal data is stored or transmitted and that prolonged, potentially health-damaging engagement, addiction, or unethical exploitation of users is prevented. In some cases, the unknown age of the volunteers/gamers could also be of concern (many teenagers and younger children are avid gamers). We also recommend that crowdsourcing projects adopt an open license, clearly stated and used as a motivating factor in recruiting contributors.

Preventing Bias. An overzealous contributor might introduce bias in the crowdsourced data by carrying out most of the work. Statistics from MTurk [16] and GWAPs [39] have shown that there are indeed a small number of people who carry out a large number of HC tasks (paid HITs or hours playing), which, if the aim is to have a more diverse set of linguistic choices, from different people, might bias the results. Similarly, "lazy" contributors might provide suboptimal results in an effort to cheat the system. The developers of the Stardust@Home citizen science project have found that "they would have to calibrate their volunteers just as they would any instrument" including assigning skill levels to players, monitoring skill levels and determining the contributor agreement required for judging a result relevant [18]. Emerging profiling techniques (Section 3.3) could be a solution to contributor management and bias prevention.

Increasing Repeatability. Although crowdsourcing approaches generally offer a more economic alternative to gathering scientific data than traditional approaches, a significant investment, both in terms of time and budget, is needed when setting up a crowdsourcing application from scratch. It is therefore vital for the development of this field to lower the access barrier to crowdsourcing methods and to increase repeatability by making it easier to reuse elements such as games, MTurk task definitions, licenses and consent forms. An encouraging step in that direction is Bossa[12] , an open-source framework for implementing "distributed thinking"

[12] `http://boinc.berkeley.edu/trac/wiki/BossaIntro`

Genre	Diversification Type				Stage of scientific process		
	Less-Resourced Languages	**New Resource Types**	**Study of Language Phenomena**	**Solve Subjective Tasks**	**Input& Training Data**	**Algorithm Support**	**Evaluation**
Mechanised Labour	[2, 6, 22] [52]	[15, 26, 28] [36, 38, 47]	[29, 32]	[13, 31, 33, 34] [37, 52, 45]	[1, 2, 22] [26, 28, 31] [33, 34, 36] [37, 45, 47] [52]	[27]	[6, 17]
GWAP	[41]			[39, 41]	[39, 41]		[24]
Altruistic Crowdsourcing	[1]			[5]	[1]	[5]	

Table 1: Correlation between crowdsourcing genres, the research diversification they enable and the stages of the scientific process that they support.

projects that can be used by anyone. Striking a balance between offering maximum flexibility to researchers wishing to use such platforms and making sure that only properly-committed scientists have access to the valuable resources pro-vided by volunteers remains a controversial issue [18].

Towards an emergent, hybrid-computing infrastructure. Crowdsourcing projects in general, and those geared towards scientific exploration in particular, facilitate a closer integration of human and machine computation than it was ever possible before. In the area of NLP, such symbiosis is reached through increasingly complex workflows where machines and humans take turns in solving different tasks within the same workflow. Similarly, active learning methods reduce the feedback cycles between algorithms and humans, with humans solving on-demand tasks identified by the algorithm. These are early signs of what experts in the field of human computation converge to foresee as the advent of a novel "architecture to compute on" [42] or an "emerging human-computer network constituting the global brain" [4].

ACKNOWLEDGMENTS

The work presented in this paper was developed within DIVINE (www.weblyzard.com/divine), a project funded by FIT-IT Semantic Systems of the Austrian Research Promotion Agency (www.ffg.at) and the Federal Ministry for Transport, Innovation and Technology (www.bmvit.gv.at). K. Bontcheva is supported by funding from the Engineering and Physical Sciences Research Council (grant EP/I004327/1).

5. REFERENCES

[1] T. Abekawa, M. Utiyama, E. Sumita, and K. Kageura. Community-based Construction of Draft and Final Translation Corpus through a Translation Hosting Site Minna no Hon'yaku (MNH). In *Proc. of LREC*, 2010.

[2] V. Ambati and S. Vogel. Can Crowds Build Parallel Corpora for Machine Translation Systems? In Callison-Burch and Dredze [8], pages 62–65.

[3] T.S. Behrend, D.J. Sharek, A.W. Meade, and E.N. Wiebe. The viability of crowdsourcing for survey research. *Behav. Res.*, 43(3):800–813, 2011.

[4] A. Bernstein, M. Klein, and T. W. Malone. Programming the Global Brain. *Commun. ACM*, 55(5):41–43, 2012.

[5] A. Brew, D. Greene, and P. Cunningham. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proc. of the European Conf. on Artificial Intelligence*, pages 145–150, 2010.

[6] C. Callison-Burch. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proc. of the Conf. on Empirical Methods in NLP*, pages 286–295, 2009.

[7] C. Callison-Burch and M. Dredze. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* [8], pages 1–12.

[8] C. Callison-Burch and M. Dredze, editors. *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[9] P. Cohn. Can Volunteers Do Real Research? *BioScience*, 58(3):192–197, 2010.

[10] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[11] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing Systems on the World-Wide Web. *Commun. ACM*, 54(4):86–96, April 2011.

[12] C. Eickhoff and A. de Vries. Increasing Cheat Robustness of Crowdsourcing Tasks. *Information Retrieval*, 15:1–17, 2012. 10.1007/s10791-011-9181-9.

[13] M. El-Haj, U. Kruschwitz, and C. Fox. Using Mechanical Turk to Create a Corpus of Arabic Summaries. In *Proc. of LREC*, 2010.

[14] T.A. Finholt. Collaboratories. *Annual Review of Info. Science and Technology*, 36:74 –104, 2002.

[15] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating Named Entities in Twitter Data with Crowdsourcing. In Callison-Burch and Dredze [8], pages 80–88.

[16] K. Fort, G. Adda, and K.B. Cohen. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37(2):413 –420, 2011.

[17] D. Gillick and Y. Liu. Non-Expert Evaluation of Summarization Systems is Risky. In Callison-Burch and Dredze [8], pages 148–151.

[18] E. Hand. Citizen science: People power. *Nature*, 466:685 – 687, 2010.

[19] L. Hoffmann. Crowd Control. *Commun. ACM*, 52(3):16 –17, 2009.

[20] E.H. Hovy, M. P. Marcus, M. Palmer, L. A. Ramshaw, and R. M. Weischedel. OntoNotes: The 90% Solution.

In *Proc. of HLT-NAACL*, 2006.

[21] J. Howe. Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business, 2009. http://crowdsourcing.typepad.com/.

[22] A. Irvine and A. Klementiev. Using Mechanical Turk to Annotate Lexicons for Less Commonly Used Languages. In Callison-Burch and Dredze [8], pages 108–113.

[23] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, E. Zarour, and Phylo players. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. *PLoS ONE*, 7(3):e31362, 2012.

[24] A. Koller, K. Striegnitz, A. Gargett, D. Byron, J. Cassell, R. Dale, J. Moore, and J. Oberlander. Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In *Proc. of the Int. Natural Language Generation Conf.*, pages 243–250, 2010.

[25] M. Krause and J. Smeddinck. Human Computation Games: a Survey. In *Proc. of 19th European Signal Processing Conference (EUSIPCO 2011)*, 2011.

[26] S. A. Kunath and S. H. Weinberger. The Wisdom of the Crowd's Ear: Speech Accent Rating and Annotation with Amazon Mechanical Turk. In Callison-Burch and Dredze [8], pages 168–171.

[27] F. Laws, C. Scheible, and H. Schütze. Active Learning with Amazon Mechanical Turk. In *Proc. of the Conf. on Empirical Methods in NLP*, pages 1546–1556, 2011.

[28] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. In Callison-Burch and Dredze [8], pages 71–79.

[29] N. Madnani, J. Boyd-Graber, and P. Resnik. Measuring Transitivity Using Untrained Annotators. In Callison-Burch and Dredze [8], pages 188–194.

[30] T. W. Malone, R. Laubacher, and C. Dellarocas. The Collective Intelligence Genome. *MIT Sloan*, 51(3):21–31, 2010.

[31] J. Mrozinski, E. Whittaker, and S. Furui. Collecting a Why-Question Corpus for Development and Evaluation of an Automatic QA-System. In *Proc.of ACL: HLT*, pages 443–451, 2008.

[32] R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. Crowdsourcing and Language Studies: The New Generation of Linguistic Data. In Callison-Burch and Dredze [8], pages 122–130.

[33] M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora. In *Proc. of the Conf. on Empirical Methods in NLP*, pages 670–679, 2011.

[34] M. Negri and Y. Mehdad. Creating a Bi-lingual Entailment Corpus through Translations with Mechanical Turk : 100 for a 10-day Rush. In Callison-Burch and Dredze [8], pages 212–216.

[35] O. Nov, O. Arazy, and D. Anderson. Crowdsourcing for Science: Understanding and Enhancing Science Sourcing Contribution. In *WS. on the Changing Dynamics of Scientific Collaborations*, 2010.

[36] S. Novotney and C. Callison-Burch. Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription. In *Proc. of HLT-NAACL*, pages 207–215, 2010.

[37] G. Parent and M. Eskenazi. Clustering Dictionary Definitions Using Amazon Mechanical Turk. In Callison-Burch and Dredze [8], pages 21–29.

[38] G. Parent and M. Eskenazi. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *Proc. of INTERSPEECH*, pages 3037–3040, 2011.

[39] M. Poesio, U. Kruschwitz, J. Chamberlain, L. Robaldo, and L. Ducceschi. Phrase Detectives: Utilizing Collective Intelligence for Internet-Scale Language Resource Creation. *Transactions on Interactive Intelligent Systems*, 2012. To Appear.

[40] A. J. Quinn and B. B. Bederson. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proc. of Human Factors in Computing Systems*, pages 1403–1412, 2011.

[41] W. Rafelsberger and A. Scharl. Games with a Purpose for Social Networking Platforms. In *Proc. of the Conf. on Hypertext and Hypermedia*, pages 193–198, 2009.

[42] N. Savage. Gaining Wisdom from Crowds. *Commun. ACM*, 55(3):13–15, 2012.

[43] A. Scharl, M. Föls, and M. Sabou. ClimateQuiz: A Game for Gathering Environmental Knowledge, 2012. Under Peer Review.

[44] A. Scharl, M. Sabou, S. Gindl, W. Rafelsberger, and A. Weichselbraun. Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources. In *Proc. of the LREC*, 2012.

[45] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and Fast—but is it Good?: Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proc. of EMNLP*, pages 254–263, 2008.

[46] S. Thaler, K. Siorpaes, C. Hofer, and E. Simperl. A Survey on Games for Knowledge Acquisition. Technical report, Semantic Techology Institute, Innsbruck, 2011.

[47] K. Vertanen and P. O. Kristensson. The Imagination of Crowds: Conversational AAC Language Modeling using Crowdsourcing and Large Data Sources. In *Proc. of the Conf. on Empirical Methods in NLP*, pages 700–711, 2011.

[48] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[49] A. Wang, C.D.V. Hoang, and M. Y. Kan. Perspectives on Crowdsourcing Annotations for Natural Language Processing. *Language Resources and Evaluation*, 2012.

[50] A. Wiggins. Crowdsourcing Science: Organizing Virtual Participation in Knowledge Production. In *Proc. of the Int. Conf. on Supporting Group Work*, pages 337–338, 2010.

[51] A. Wiggins and K. Crowston. From Conservation to Crowdsourcing: A Typology of Citizen Science. In *Proc. of the Hawaii Int. Conf. on System Sciences*, pages 1–10, 2011.

[52] O. F. Zaidan and C. Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proc. of ACL: HLT*, pages 1220–1229, 2011.