Crowdsourcing

Massimo Poesio, Jon Chamberlain and Udo Kruschwitz

Abstract Most annotated corpora of wide use in computational linguistics were created using traditional annotation methods, but such methods may not be appropriate for smaller scale annotation and tend to be too expensive for very large scale annotation. This chapter covers crowdsourcing, the use of web collaboration for annotation. Both microtask crowdsourcing and games-with-a-purpose are discussed, as well as their use in computational linguistics.

1 Introduction

Most annotated corpora of wide use in Computational Linguistics (CL) were created using traditional annotation methods (this is the case, e.g., for most case studies in Part II of the Handbook) but such methods may not be appropriate for smaller scale annotation and tend to be too expensive for very large scale annotation. Outside CL, **crowdsourcing**[1]–outsourcing the creation of resources to large numbers of Internet users[2] –has become an established method for labelling data and for other resource

Massimo Poesio, Jon Chamberlain and Udo Kruschwitz
University of Essex, Language and Computation e-mail: {poesio,jchamb,udo}@essex.ac.uk

[1] The alternative term **human computation** is arguably more popular in other fields, but crowdsourcing is more popular in Computational Linguistics.

[2] A more formal and systematic definition of crowdsourcing has been provided in [19]:

> "Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken."

creation efforts [28]. In the last ten years, this methodology has also been adopted in Computational Linguistics as an alternative to traditional annotation methods, becoming the *de facto* standard for small-scale annotation projects. And as we will see in this Chapter, the methodology may also be the solution for projects whose objective is to create very large scale datasets. In this Chapter, we discuss the use of crowdsourcing for annotation in CL. (For a general introduction to crowdsourcing, we recommend the already mentioned book by Howe [28]; for more detailed information, and the applications of crowdsourcing in other fields, the *Handbook of Human Computation* [34].)

The structure of the Chapter is as follows. In Section 2 we discuss various types of crowdsourcing. In Section 3 we discuss the use of microtask crowdsourcing in computational linguistics. In Section 4 we discuss the use of games-with-a-purpose. Section 5 summarizes the lessons learned so far.

2 Approaches to collective resource creation

The different types of crowdsourcing can be distinguished on the basis of what **motivates** the participants to collaborate. At least three types of motivation can be distinguished: collaboration motivated by **shared intent**, by **financial incentives**, and by **enjoyment**. We briefly discuss each type in turn in this Section; for a more extensive discussion, see [13].[3]

2.1 Shared intent

One of the most potent motivations for large-scale collaboration on the Web is the desire to support a scientific enterprise, the creation of a shared resource, or in general an enterprise viewed as worthy.

Wikipedia

Wikipedia was perhaps the first project to show what can be really achieved through the willingness of Web users to collaborate in an enterprise to create a resource of general utility.[4] As of April 2015, English Wikipedia numbers 4,866,554 articles (i.e., 420,000 more than when the first version of this Chapter was written, in

[3] A number of alternative classification schemes for crowdsourcing have been proposed –see, e.g., [43, 50]. We return to the Wang *et al.* study below.

[4] The creation of the Oxford English Dictionary in the nineteenth century, which involved the collaboration of thousands of volunteers proposing candidate words and senses, is perhaps the best known example of the use of this approach in the pre-Web era.

February 2014), written by over 20 million collaborators and 1,400 editors[5]. By contrast, the edition of *Encyclopedia Britannica* of 2007 had 700 'Macro' articles and 70,000 'micro' articles, created by around 4,000 experts coordinated by 100 editors. Wikipedia is also fully multilingual: there are versions of Wikipedia in 288 languages, 8 of which number more than one million articles (Dutch, French, German, Italian, Polish, Russian, Spanish, and Swedish), and 43 more than 50,000. This extraordinary wealth of information makes wikipedia.org one of the top 10 most popular sites on the Web, and information extracted from Wikipedia itself or one of the many databases derived from Wikipedia (such as dbpedia or Yago) is used in an extraordinary number of projects in Computational Linguistics. Wikipedia also illustrates the effectiveness of 'bottom-up' or 'self-organizing' editorial control, where the reviewers are themselves volunteers who are considered by the Wikipedia community to be competent (i.e., by having an approval rate of over 75%).

Citizen science

Another powerful illustration of the potential of crowdsourcing is the success of projects like *Foldit*,[6] *Galaxy Zoo*[7] or *Phylo*[8] that have made genuine contributions to research in biology, astronomy, and other fields by recruiting thousands of web collaborators to help with time-consuming tasks such as galaxy classification. (The three projects mentioned are also examples of **games with a purpose**, see below.)

Open Mind Commonsense

Open Mind Common Sense[9] [46] was perhaps the first demonstration that Web collaboration can be relied on to create resources for Artificial Intelligence, as well. More than 15,000 volunteers contributed over a million commonsense facts in the form of sentences, that were then compiled into a conceptual knowledge repository called ConceptNet [24]. The latest version of ConceptNet, ConceptNet5,[10] also includes knowledge from other collectively created knowledge resources such as DBPedia (created from Wikipedia) as well as from publically available resources such as WordNet, and, with other 10 million facts, is one of the largest sources of conceptual knowledge currently available. The Open Mind Common Sense project also led to the development of a 'quasi-game' for collecting commonsense knowledge, the system *LEARNER* [15].

---

[5] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[6] http://fold.it/portal

[7] http://www.galaxyzoo.org/

[8] http://phylo.cs.mcgill.ca/

[9] http://www.openmind.org

[10] http://conceptnet5.media.mit.edu

2.2 Financial incentives

The simplest way to incentivize collaborators is to pay them. Amazon Mechanical Turk[11] (AMT) pioneered the approach to resource creation called **microtask crowdsourcing**: outsourcing a piece of work to 'the crowd' using the Web as a way of reaching very large numbers of collaborators (called **workers** in this Chapter[12]) who get paid to complete small items of work called **human intelligence tasks** (HIT).

Advantages

The payment is typically fairly small, in the order of 1 to 20 US cents per HIT. AMT and CrowdFlower[13] demonstrated that crowdsourcing is very competitive with traditional resource creation methods from a financial perspective, because even very little payment is enough to attract large number of collaborators (many of which are students or otherwise unemployed, or live in countries in which the cost of living is lower). A further advantage is that workers work very fast–it is not uncommon for a HIT to be completed in minutes. These considerations resulted in crowdsourcing becoming a standard way of creating small- and medium- scale resources for computational linguistics, as discussed in the following Sections.

Issues

A number of questions have, however, been raised regarding this approach. One regards the quality of resources created this way. Crowdsourcing platforms provide a number of mechanisms for quality control. AMT provides three quality-control mechanisms: (i) each HIT can be completed by multiple workers, which makes it possible to identify noise; (ii) the requester can require that workers satisfy certain qualifications, such as a high acceptance rate for their previous HITs; and (iii) the requester can reject the work of workers. Crowdflower provides an extensive set of quality control mechanisms, as well [38]. For instance, gold standard data can be used can be used to block worker access to jobs if they cannot complete tasks whose answer is provided by the gold standard; or they can be mixed with previously unannotated data to get constant quality control. Yet doubts about the quality of the data thus created remain. Some studies showed that the quality of resources created this way is comparable to that of resources created in the traditional way, provided that multiple judgments are collected in a sufficient number [47, 9]. Other studies

[11] https://www.mturk.com/

[12] The term **turkers** is also often used on Amazon Mechanical Turk, but this term is often perceived as having a negative connotation.

[13] http://crowdflower.com

however found a substantial lower quality in comparison with resources created in the traditional way [6].

A second issue, raised e.g., by [22], concerns the wages paid to workers and more in general their rights. Other microtasking platforms, such as Samasource[14], guarantee workers a minimum payment level and basic rights. For additional information and discussion, see [22] as well as the relevant chapters of the *Handbook of Human Computation* such as [11] and the chapter in the same Handbook on legal issues[20] .

2.3 Enjoyment

Luis von Ahn from Carnegie Mellon University, Timothy Chklovsky from the Open Mind Common Sense group, and others argue that the desire to be entertained could be as powerful an incentive as financial reward. It is estimated that every year over 9 billion person-hours are spent by people playing games on the Web [1]. If even a fraction of this effort could be redirected towards resource creation via the development of Web games that achieve resource creation as a side effect of having people play entertaining games (von Ahn called such games **games-with-a-purpose** or GWAP) we would have enormous quantity of man-hours at our disposal.

von Ahn demonstrated his point through the development of several GWAP. The best known of these games is the ESP Game.[15] In the ESP Game two randomly chosen players are shown the same image. Their goal is to guess how their partner will describe the image (hence the reference to extrasensory perception or ESP) and type that description under strict time constraints. If any of the strings typed by one player matches the strings typed by the other player, they score both points. From the players' perspective that is all that matters. The descriptions of the images players provide are very useful information to train content-based image retrieval tools [2]. von Ahn's intuition that the game would attract very large numbers of Web visitors proved correct. The game attracted 13,000 players between August and December 2003 and has attracted over 200,000 players since, who have produced over 50 million labels. The quality of the labels has also been shown to be as good as that produced through conventional image annotation methods. A crucial advantage of GWAP over crowdsourcing is that, once the game has been developed and made available, it can continue to generate annotations with very little maintenance and very little cost. Indeed, the game was so successful that a license to use it was bought by Google, which developed it into the Google Image Labeler which was online from 2006 to 2011. The story of the Google Image Labeller[16] illustrates many useful points about what is required to make a GWAP successful: from the need to provide incentives to players, to that of continuously revising the game's

[14] http://samasource.org

[15] von Ahn's games used to be available from www.gwap.com, but the site is now dormant. ESP is still occasionally available at http://www.espgame.org

[16] http://en.wikipedia.org/wiki/Google_Image_Labeler

methods for controlling malicious behavior to stay one step ahead of the malicious players. We discuss these requirements in Section 4.

Many other GWAP have been developed by von Ahn and other labs to collect data for multimedia tagging (*OntoTube*,[17] *Tag a Tune*[18]) and for acquiring commonsense knowledge (*Verbosity*,[19] *OntoGame*,[20] *Categorilla*[21], *Free Association*[22]). The GWAP concept was also adopted in citizen science projects, such as the already mentioned *Foldit* (a GWAP about protein folding developed at the University of Washington) and *Phylo*.

3 Microtask crowdsourcing in computational linguistics

The form of web collaboration most used to create resources in computational linguistics is microtask crowdsourcing through Amazon Mechanical Turk (AMT) or CrowdFlower,[23] largely as a result of two influential papers by Snow *et al.* [47] and Callison-Burch [9]. We discuss each paper in turn.

3.1 Crowdsourcing for annotation

[47] explored the use of Amazon Mechanical Turk as an alternative to traditional annotation methods. Snow and colleagues used AMT workers for five annotation tasks on texts for which independently produced expert annotations already existed: sentiment analysis, word similarity, recognizing textual entailment, event temporal ordering, and wordsense disambiguation. For each of these tasks, Snow *et al.* collected annotations from 10 AMT workers, and then compared the (average) interannotator agreement between a turker and the average of the other workers with the average IAA between an expert and the average of the other experts. They found that generally speaking agreement between experts measured this was was higher than agreement between workers, but also that by raising the number of workers the interannotator agreement between workers would raise; and that at most 10 crowdsourced annotations (and in some cases less) would be required to achieve the same agreement as between experts. Snow *et al.* also compared training a sentiment analy-

[17] OntoTube used to be online at http://ontogame.sti2.at/games

[18] Tagatune used to be available as http://www.gwap.com/gwap/gamesPreview/tagatune or from Facebook. The site now appears to be dormant.

[19] Verbosity used to be accessible at http://www.gwap.com/gwap/gamesPreview/verbosity

[20] http://ontogame.sti2.at/games

[21] http://ai.stanford.edu/ dvickrey/wordgame/

[22] http://ai.stanford.edu/ dvickrey/wordgame/

[23] As of May 2015 Amazon Mechanical Turk requires payment with a US-based credit card hence most researchers outside the USA use CrowdFlower that does not have such a restriction.

sis system on the crowdsourced annotations with training it on the gold annotations, finding that comparable results could be achieved.

These results had a substantial impact; crowdsourcing with AMT or other platforms has been widely adopted in the computational linguistics community, and has now become the standard method for producing small-scale annotations.

3.2 Crowdsourcing for translation and evaluation

[9] showed that microtask crowdsourcing can also be used to evaluate tasks such as Machine Translation where simple comparison against a gold standard is not appropriate.

Callison-Burch's first objective was to use AMT to evaluate translations. In this part of the work, he asked workers to judge the quality of machine translations produced by the systems participating in the German / English news translation task at the 2008 Workshop on Statistical Machine Translation (WMT08). The HIT exactly replicated the interface used for WMT08: the workers were shown a source sentence, a reference translation, and five translations produced by MT systems participating in the competition, and were asked to rank the system translations assigning scores from the best to the worst. 200 such HITs were produced, each shown to five different workers. The total cost was $9.75. Both the individual judgments and the combined ranked judgments were then compared with those of the experts. The comparison of individual workers with experts highlighted the great variety in quality between workers. This in turn suggested that workers's opinions should be assigned different weight depending on the reliability of the workers. Two types of weighing was tested: weighing a turker's contribution on the basis of how frequently he/she agrees with other workers; and weighing on the basis of agreement with experts on the first 10 assignments. Combined ranked judgments–unweighted, or weighted according to the two methods–were then compared with expert judgments. The results show that whereas experts agreed with each other 58% of the time, agreement between single workers and experts was 41% on average; agreement between experts and the unweighted combined ranking of 5 workers was 53%; and agreement between experts and weighted combined ranking of 5 workers was also 58%, i.e. identical to the agreement between experts. (For a discussion of Inter-Annotator Agreement, see Chapter I V f.)

Callison-Burch also tested using workers in a variety of more complex tasks, such as producing reference translations and scoring systems according to the official GALE scoring metric, HTER. For the first task, workers were asked to produce translations for 50 sentences in French, German, Spanish, Chinese and Urdu. The results showed that provided that filtering techniques were used to identify the translations that workers produced by cutting and pasting machine translations, these AMT-produced translations were of a quality almost as high as that of professionally produced translations.

3.3 Other uses of microtask crowdsourcing in computational linguistics

In the last five years microtask crowdsourcing has become the method of choice for creating small and medium scale resources for computational linguistics projects.

The methodology has been used, first of all, to create corpora for training and evaluation in tasks such as speech transcription [33]; part-of-speech tagging [31]; named entity recognition ([21]; see also Chapter 2 III d ii); ; wordsense disambiguation [39]; and deception detection [35]. Second, microtask crowdsourcing has been used for tasks that require more complex gold standards, such as machine translation and summarization [18]. Third, microtask crowdsourcing has been used to create resources for use in computational linguistics. For instance, [7] used AMT to create a wordsense dictionary, whereas [36] used it to create an emotional lexicon.

Indeed, crowdsourcing is now so popular in computational linguistics that whole workshops have been devoted to the topic–for instance, the workshops on *Collaboratively Created Language Resources* at ACL 2009, and on *Creating Speech and Language Data with Amazon's Mechanical Turk* at NAACL/HLT 2010– as well as a special issue of *Language Resources and Evaluation* in 2013 on Collaboratively Created Language Resources [23]. A number of conferences on crowdsourcing more in general also publish computational linguistics work or work relevant to annotation in computational linguistics–e.g., the annual Conference on Human Computation and Crowdsourcing (HCOMP) or the conference on Crowdsourcing and Data Mining (CSDM).

Crowdsourcing is also being applied for linguistic research beyond computational linguistics and for psycholinguistic research– for a discussion, see [37].

4 Games with a Purpose

A second type of crowdsourcing has also been used in computational linguistics to annotate corpora: the games-with-a-purpose (GWAP) approach pioneered by von Ahn [1]. This approach has not been used as widely as microtask crowdsourcing, for reasons discussed in Section 5.1, but a number of projects based GWAPs exist as this approach is perceived by many as holding more promise for the creation of truly large scale resources. We briefly survey in this Section the best known among these projects; one of the GWAPs summarized here, *Phrase Detectives*, is discussed in detail in a case study in Part 2 of this Handbook, in Chapter 2 IV a i.

4.1 Creating a Corpus for Translation: 1001 Paraphrases

1001 Paraphrases [16]–to our knowledge, the first GWAP whose aim was to create a corpus– was developed to collect training data for a machine translation system which needs to recognize paraphrase variants. In the game, players have to produce paraphrases of an expression shown at the top of the screen, like *this can help you*. If they guess one of the paraphrases already produced by another player, they get the number of points indicated on the window; otherwise the guess they produced is added to those already collected by the system, the number of points they can win is decreased, and they can try again. Chklovski reports collecting 20,944 contributions.

From a methodological point of view, the main point to note is that the task in this game is not annotation: as in the ESP game, players are required to enter text instead of choosing one interpretation. So the method could not be directly used for annotation, but could be tried for other translation-related applications, or possibly other tasks such as summarization or Natural Language Generation. However, many of the ideas developed by Chklovsky in *1001 Paraphrases* and the earlier *LEARNER* system (not really a game) are extremely useful, in particular the idea of **validation**– asking some of the collaborators to check the quality of what other collaborators have done. As we will see discussing quality control below, validation is one of the most powerful techniques for this purpose. It is difficult however to assess how successful the game was as the paper mentioned only reports a small-scale pilot study.

4.2 GWAPs for Anaphoric Reference: Phrase Detectives and PlayCoref

Phrase Detectives

Phrase Detectives,[24] discussed in more detail in Chapter 2 IV a i., is a single-player GWAP developed to collect data about English (and subsequently Italian) anaphoric coreference [42] The game architecture is articulated around a number of tasks and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. The game design is based on a detective theme, relating to the how the player must search through the text for a suitable annotation [14].

The players have to carry out two different tasks. Initially text is presented in Annotation Mode (called *Name the Culprit* in the game - see Figure 1). This is a straightforward annotation mode where the player makes an annotation decision about a highlighted **markable** (section of text). (The annotation scheme used in *Phrase Detectives* is a simplified version of the anaphoric annotation scheme used in the ARRAU corpus [41].) If different players enter different interpretations for a markable then each interpretation is presented to more players in Validation Mode

[24] http://www.phrasedetectives.com

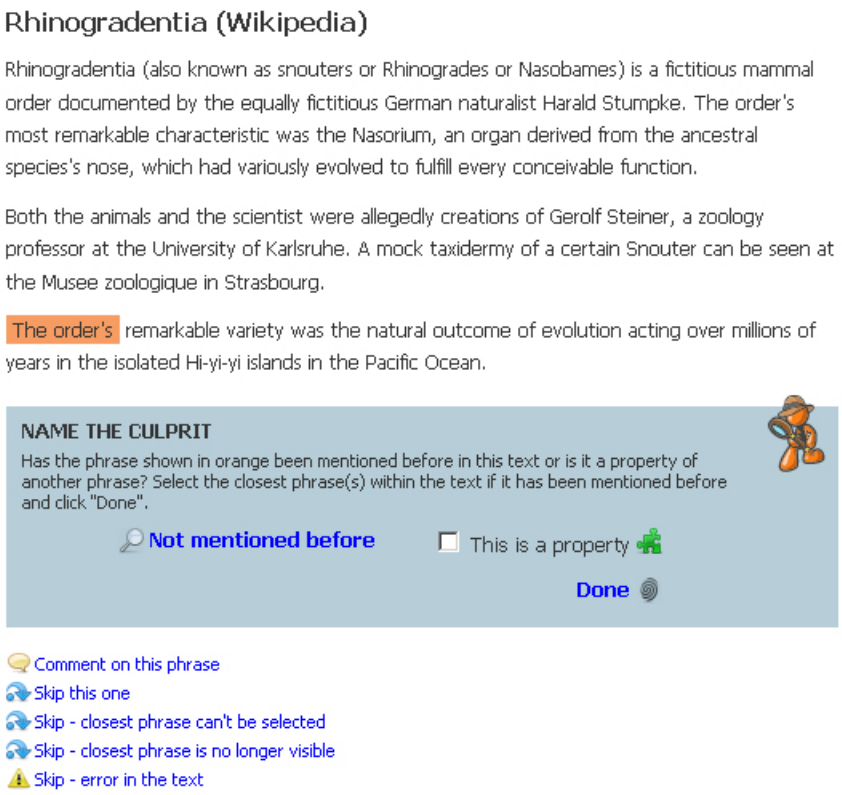

Fig. 1 Detail of a task presented in Annotation Mode.

(called *Detectives Conference* in the game). The players in Validation Mode have to agree or disagree with the interpretation.

Players are trained with texts from a gold standard. Players always receive a training text when they first start the game. Once the player has completed all of the training tasks they are given a rating (the percentage of correct decisions out of the total number of training tasks). If the rating is above a certain threshold (currently 50%) the player progresses on to annotating real documents, otherwise they are asked to do a training document again. The rating is recorded with every future annotation that the player makes as the rating is likely to change over time. The scoring system is designed to reward effort and motivate high quality decisions by awarding points for retrospective collaboration. A mixture of incentives, from the personal (scoring, levels) to the social (competing with other players) to the financial (small prizes) are employed.

The goal of the game was not just to annotate large amounts of text, but also to collect a large number of judgments about each linguistic expression. This led to the deployment of a variety of mechanisms for quality control which try to reduce the

amount of unusable data beyond those created by malicious users, from the level mechanism itself to validation to a number of tools for analysing the behavior of players.

A Facebook version of *Phrase Detectives*,[25] launched in February 2011, makes full use of socially motivating factors inherent in the Facebook platform. For instance, any of the player's friends who are playing the game form the player's team, which is visible in the left hand menu. Whenever a player's decision agrees with a team member they score additional points. The most interesting finding from this work is that although fewer players play it, the quality and quantity of their work is significantly superior to that of the players of the original game; more in general, knowing the identity of the player leads to much better quality [12].

Phrase Detectives is one of the most successful GWAPs for computational linguistics. Started in December 2008, it is still being played. As of April 2015, about 40,000 players have registered (i.e., 6,000 more than when the first draft of this Chapter was completed); of these, 4,000 passed the training phase–around 1,000 of which on *Facebook Phrase Detectives*. Over 2 million annotation judgments have been collected (280,000 more than in February 2014) and 444,000 validations (145,000 more); 546 documents have been completely annotated (up from 494) for a total of around 316,000 words, up from 229,453 (the complete corpus will be of 1.2 million words).

PlayCoref

Another GWAP for anaphoric annotation exists: *PlayCoref*, developed at Charles University in Prague [25]. PlayCoref is a two-player game in which players can interact with each other. A number of empirical evaluations have been carried out showing that players find the game very attractive but to our knowledge the game has not yet been put online to collect data on a large scale.

4.3 Sentiment Analysis

As already discussed regarding *Phrase Detectives*, GWAPs integrated into social networking sites such as *Sentiment Quiz*[26] on Facebook show that social interaction within a game environment does motivate players to participate [44]. The *Sentiment Quiz* asks players to select a level of sentiment (on a 5 point scale) associated with a word taken from a corpus of documents regarding the 2008 US Presidential election. The answer is compared to another player and points awarded for agreement.

[25] http://apps.facebook.com/phrasedetectives

[26] https://www.modul.ac.at/about/departments/new-media-technology/projects/sentiment-quiz/

4.4 Creating (and Annotating) a Corpus for Generation: GIVE

A family of GWAP have been used to collect data actually used in Computational Linguistics: the GIVE games,[27] developed in support of the the GIVE-2 challenge for generating instructions in Virtual Environments, initiated in the Natural Language Generation community [30]. GIVE-2, for instance, is a treasure-hunt game in a 3D world. When starting the game, the player sees a 3D game window, which displays instructions and allows the players to move around and manipulate objects. In the first room players learn how to interact with the system; then they get in an evaluation world where they perform the treasure hunt, following instructions generated by one of the systems participating in the challenge. The players can succeed, lose, or cancel the game; this outcome is used to compute the **task success** metric, one of the metrics used to evaluate the systems participating in the challenge.

GIVE-2 was extremely successful as a way to collect data for HLT, collecting over 1825 game sessions in three months, which played a key role in determining the results of the challenge. GIVE-2 is an extremely attractive game to play, which no doubt contributed in part to its success. Again, this methodology would not be appropriate to annotate pre-existing text; it may be possible however to learn about anaphora from the data produced this way.

4.5 GWAPs for Parsing: PhraTris

PhraTris [4] is a GWAP for syntactic annotation developed by Giuseppe Attardi's lab at the University of Pisa using a general-purpose GWAP development platform called GALOAP.[28] *PhraTris* is a very entertaining game and won the INSEMTIVES game challenge 2010 but has not yet been put online to collect data.

4.6 The Groningen Meaning Bank

At present, the most ambitious project using web collaboration to annotate data for Computational Linguistics is the *Groningen Meaning Bank* (GMB),[29] (discussed in more detail in Chapter 2 I b). The GMB effort has three key characteristics [5]. First, web collaboration is used to annotate *all* linguistic levels, from POS to syntax to semantics to discourse, including discourse relations. Second, the aim is to annotate 'deep' linguistic information, i.e., associating text with its full linguistic analysis at a given level, all the way to a full representation of the meaning of a discourse in Discourse Representation Theory [29]. Third (and a virtual corollary of the pre-

[27] http://www.give-challenge.org

[28] http://galoap.codeplex.com

[29] http://gmb.let.rug.nl/

vious point, given that manually constructing such full representations would be prohibitively time consuming), the use of a *human-aided machine annotation approach*, in which the full linguistic analyses are first produced by a POS tagger, or parser, or semantic interpreter, so that the task of the collaborator is to correct them.

Two main forms of crowdsourcing are used in the GMB: some of the work is carried out as shared intent, but a suite of GWAPs called *WordRobe*[30] is used for some annotation tasks, including POS tagging, named entity tagging, anaphora, and wordsense. The wordsense annotation GWAP, *Senses*, is discussed in [49].

5 Using crowdsourcing for annotation

Computational linguists have accumulated by now considerable experience in using crowdsourcing for annotation. In this Section we summarize some of the lessons learned through this experience.

5.1 Microtask crowdsourcing vs GWAP

The first question for a CL practitioner is whether to use microtask crowdsourcing or GWAPs. A very useful comparison between the two approaches can be found in [50]. Wang *et al.* identify five dimensions along which these approaches to crowdsourcing discussed in Section 2 can be compared–**motivation**, **annotation quality**, **setup effort**, **human participation**, and **task character**–and score nine uses of crowdsourcing along each dimension: two GWAPs (*Phrase Detectives* and ESP), six uses of microtask crowdsourcing (the five case studies by [47] and the use of in TREC Blog Assessment), and two approaches based on shared intent (Wikipedia and the Open Mind Initiative). Their conclusions are as follows:

GWAPs Pros: they have the lowest long-term costs so are potentially usable for bigger annotation projects. The cons are the costs to setup the game, and the slow pace at which the annotation procedes. Also, not all CL annotation tasks can be turned into a fun or at least moderately interesting game.

Microtask crowdsoourcing Pros: the setup cost is almost nil, and the task can be completed very quickly and spending very little. Cons: the costs for really big annotation projects end up being higher than with GWAPs. (See below.) The quality of the annotation may be low.

Shared interest Pros: the quality of annotation produced by people who are doing this as a labour of love can be quite high. Cons: altruism or scientific interest are not as powerful an incentive as financial considerations or entertainment.

[42] attempted to estimate the difference in cost between the different types of annotation more precisely. They distinguished between four types of annotation.

[30] http://www.wordrobe.org

The first type is **Traditional High Quality (THQ)** as in projects like OntoNotes [27] or SALSA [8], which involves the development of a very formal annotation scheme, dedicated annotation tools, and double or triple coding of each item under the supervision of an expert. The cost of such annotation was estimated by Poesio *et al.* at around $1 per corpus token (word). **Traditional, Medium Quality (TMQ)** annotation also involves the development of a formal coding scheme and training of annotators, but most items will be typically annotated only once, although around 10% of items will be double-annotated to spot misunderstandings and other problems. The cost of this annotation, including the salary of a supervisor, works at around $.4 per token. The costs for **crowdsourcing** depend on the amount paid per HIT and on the number of multiple judgments collected. In our experience, .05 US $ per HIT is the minimum required for non-trivial tasks, and for a task like anaphora, the cost is typically around .1 US $ per hit, i.e., .1 US $ per markable, which at the rate of 3 tokens per markable, works out at around .03 US $ per token. Many researchers only require five judgments per item, but in practice we find that 10 is more like the number needed; this results in a cost of 1 US $ per markable, i.e., .33 $ per token. Adding the salary of a supervisor, we end up with a cost of .38 - .43 $ per token / 1.2–1.3 US$ per markable, which is about the cost with TMQ. By contrast, the cost for a **GWAP** like *Phrase Detectives* was quite high at the beginning as the game had to be created–65,000 US $ for the first two years– but after that the only real cost has been the prizes, around £1,000 a year, as checking of annotations is done by the players themselves. The total cost so far has been around 100,000 US $ for around 316,000 completely annotated tokens. If the current rate of growth of 80,000 tokens per year (at a cost of $ 1,500 per year) remains the same, we can project a total cost of US $ 110,000 to annotate 1 million words, i.e., $ .11 per token. The real tradeoff regards time: one of big advantages of microtask crowdsourcing is speed, whereas even if the current rate of growth could be maintained, it would have taken about 13 years to annotate 1.2 M words with *Phrase Detectives*. The following table summarizes the costs for creating a corpus of 1 million words using the four methods.

Table 1 Comparison of costs in US$ using four different annotation methods.

Method	Cost/token	Cost/markable	Cost/million tokens
Traditional, High Quality	1	3	1,000,000
Medium, High Quality	.4	1.2	400,000
Amazon Mechanical Turk	.38-.43	1.2-1.3	380,000-430,000
Games With A Purpose	.11	.33	110,000

5.2 Quality control

Quite a few lessons have also been learned about how best to use crowdsourcing for annotation, many of which, in particular those regarding quality control, can already be found in [47]. The first lesson is the need for **redundancy** to achieve comparable quality to traditional annotation: at least 4 and in fact typically more workers are needed for each item. This finding is pretty robust and holds both for microtask crowdsourcing and when using GWAPs.

One of the most successful techniques for ensuring quality is **validation**–having other collaborators checking the quality of what previous collaborators have done. This is the principle that makes Wikipedia work and it has been shown to work both for microtask crowdsourcing (e.g., [9]) and for GWAPs (e.g., [42]).

5.3 Finding reliable annotators and reliable annotations

One of the more important lessons about crowdsourcing (in fact, about annotation in general) is that annotation quality varies a lot from collaborator to collaborator [47, 9] hence methods are needed to identify poor-quality collaborators and/or items. [47] developed a method to estimate workers' judgments; [9] developed techniques for weighing the annotators; more recently, Bayesian models originally developed to assess the quality of multiple judgments in diagnosis have become widely used to simultaneously assess the quality of workers and labels.

The first of such models we are aware of was proposed by David and Skene [17]. In this (generative) model, the probability that the actual label of item i is z_i, given the observed labels $\overline{y_i}$ produced by the annotators, is specified as follows:

$$p(z_i|y_i, \theta, \pi) \propto p(z_i|\pi) * p(y_i|z_i, \theta)$$

where π_k is **prevalence**, i.e., the probability that an item belongs to category k, whereas $\theta_{j,k,k}$ is **annotator response**, i.e., the probability that annotator j labels an item as k' when its actual category is k. The parameters of such a model can be estimated using EM, obtaining as a result both the probability of each label for item i and an assessment of the quality of annotator j. Carpenter and Passonneau used the Dawid and Skeene model to assess the quality of wordsense in the MASC corpus [39, 40] (the MASC corpus is discussed in Chapter 2 I c). More advanced Bayesian models have also been proposed. The models proposed in [48, 51] also include explicit models of the difficulty of items, and the model proposed by Carpenter [10] provides an explicit estimate of the probability distribution of workers. Raykar *et al.* [45] propose a model that simultaneously also trains a classifier from the crowdsourced data. More recently, a simplified version of the Dawid and Skene model, MACE, has been proposed by Hovy *et al.* [26]. Hovy *et al.* showed that MACE is very effective at estimating the actual labels of items, and requiring fewer parameters, it can be estimated very efficiently.

5.4 Other aspects of best practice

One of the more debated issues in crowdsourcing is whether paying workers more affects the quality or not. [32] and others found that increased payments simply increase noise. On the other end, [3] found evidence to the contrary.

6 Conclusions

Microtask crowdsourcing has become the most widely used form of annotation to annotate small-to-medium sized resources, particularly when quality only needs to be adequate (i.e., the kind of resources that are often used in computational linguistics to work on problems when no resources exist–see, e.g., [35]). However, for resources that have to be used over and over, traditional annotation methods are still used. The big question is whether microtask crowdsourcing will take over for more substantial annotation projects, as well.

By contrast, creating a GWAPs only really makes sense to annotate very large datasets. So far however no computational linguistics GWAP has replicated the success of *Foldit* and similar games–the challenge here is which annotation tasks in CL lend themselves to the development of such games.

Acknowledgements This work was in part supported by the SENSEI project.[31] The development of *Phrase Detectives* was in part supported by EPSRC. Jon Chamberlain is currently supported by EPSRC.

[31] http://www.sensei-conversation.eu/

References

1. von Ahn, L.: Games with a purpose. Computer **39**(6), 92–94 (2006)
2. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the conference on Human factors in computing systems, pp. 319–326. ACM (2004)
3. Aker, A., El-Haj, M., Albakour, M., Kruschwitz, U.: Assessing crowdsourcing quality through objective tasks. In: Proc. of LREC (2012)
4. Attardi, G., the Galoap Team: Phratris. Demo presented at INSEMTIVES 2010 (2010)
5. Basile, V., Bos, J., Evang, K., Venhuizen, N.: Developing a large semantically annotated corpus. In: Proc. of LREC, pp. 3196–3200. Istanbul, Turkey (2012)
6. Bhardwaj, V., Passonneau, R., Salleb-Alouissi, A., Ide, N.: Anveshan: a tool for analysis of multiple annotators' labelling behavior. In: Proc. of the 4th LAW (2010)
7. Biermann, C.: Creating a system for lexical substitutions from scratch using crowdsourcing. Language Resources and Evaluation **47**(1), 97–122 (2013)
8. Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., Pinkal, M.: Framenet for the semantic analysis of German: Annotation, representation and automation. In: H.C. Boas (ed.) Multilingual FrameNets in Computational Lexicography: Methods and Applications. Mouton De Gruyter (2009)
9. Callison-Burch, C.: Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pp. 286–295. Association for Computational Linguistics (2009)
10. Carpenter, B.: Multilevel bayesian models of categorical data annotation (2008). Available as http://lingpipe.files.wordpress.com/2008/11/carp-bayesian-multilevel-annotation.pdf
11. Caverlee, P.: Eploitation in human computation systems. In: P. Michelucci (ed.) Handbook of Human Computation. Springer (2013)
12. Chamberlain, J., Kruschwitz, U., Poesio, M.: Facebook phrase detectives: Social networks meet games-with-a-purpose (2012). In preparation
13. Chamberlain, J., Kruschwitz, U., Poesio, M.: Methods for engaging and evaluating users of human computation systems. In: Handbook of Human Computation. Springer (2013)
14. Chamberlain, J., Poesio, M., Kruschwitz, U.: Phrase Detectives: A Web-based Collaborative Annotation Game. In: Proceedings of the International Conference on Semantic Systems (I-Semantics'08). Graz (2008)
15. Chklovski, T., Gil, Y.: Improving the design of intelligent acquisition interfaces for collecting world knowledge from web contributors. In: Proceedings of the 3rd international conference on Knowledge capture, pp. 35–42 (2005)
16. Chklovski, T.: Collecting paraphrase corpora from volunteer contributors. In: Proceedings of K-CAP '05, pp. 115–120. ACM, New York, NY, USA (2005). DOI http://doi.acm.org/10.1145/1088622.1088644. URL http://doi.acm.org/10.1145/1088622.1088644
17. Dawid, A.P., Skene, A.M.: Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statistics **28**(1), 20–28 (1979)
18. El-Haj, M., Kruschwitz, U., Fox, C.: Using mechanical turk to create a corpus of arabic summaries. In: Proc. of LREC Workshop on Semitic Languages, pp. 36–39. Malta (2010)
19. Estellés-Arolas, E., González-Ladrón-de Guevara, F.: Towards an integrated crowdsourcing definition. Journal of Information Science **38**(2), 189–200 (2012)
20. Felstiner, A.: Labor standards. In: P. Michelucci (ed.) Handbook of Human Computation. Springer (2013)
21. Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., Dredze, M.: Annotating named entities in twitter data with crowdsourcing. In: Proc. of CSLDAMT '10 - NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 80–88 (2010)
22. Fort, K., Adda, G., Cohen, K.B.: Amazon mechanical turk: Gold mine or coal mine? Computational Linguistics **37**, 413–420 (2011). Editorial

23. Gurevych, I., Zesch, T.: Collective intelligence and language resources: introduction to the special issue on collaboratively constructed language resources. Language Resources and Evaluation **47**(1) (2013)
24. Havasi, C., Speer, R., Alonso, J.: Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In: Proc. of RANLP (2007)
25. Hladká, B., Mírovskỳ, J., Schlesinger, P.: Play the language: play coreference. In: Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, pp. 209–212. Association for Computational Linguistics (2009)
26. Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., Hovy, E.: Learning whom to trust with MACE. In: Proc. of NAACL, pp. 1120–1130 (2013)
27. Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: Ontonotes: the 90% solution. In: Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 57–60 (2006)
28. Howe, J.: Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business. Crown Publishing Group (2008)
29. Kamp, H., Reyle, U.: From Discourse to Logic. D. Reidel, Dordrecht (1993)
30. Koller, A., Striegnitz, K., Gargett, A., Byron, D., Cassell, J., Dale, R., Moore, J., J.Oberlander: Report on the second nlg challenge on generating instructions in virtual environments (give-2). In: Proceedings of the 6th International Natural Language Generation Conference. Dublin (2010)
31. Mainzer, J.E.: Labeling parts of speech using untrained annotators on mechanical turk. Master's thesis, Ohio State University (2011)
32. Mason, W., Watts, D.J.: Financial incentives and the "performance of crowds". Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter **11**, 100–108 (2010)
33. McGraw, I., Lee, C., Hetherington, I.L., Seneff, S., Glass, J.: Collecting voices from the cloud. In: Proc. of LREC (2010)
34. Michelucci, P. (ed.): Handbook of Human Computation. Springer (2013)
35. Mihalcea, R., Strapparava, C.: The lie detector: explorations in the automatic recognition of deceptive language. In: Proc. ACL/IJCNLP, pp. 309–312 (2009)
36. Mohammad, S.M., Turner, P.D.: Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. In: Proc. of CAAGET '10 - the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 26–34 (2010)
37. Munro, R., Bethard, S., Kuperman, V., Lai, V.T., Melnick, R., Potts, C., Schnoebelen, T., Tily, H.: Crowdsourcing and language studies: the new generation of linguistic data. In: Proc. of CSLDAMT '10 - NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 122–130 (2010)
38. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., Biewald, L.: Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In: Proc. of the AAAI Workshop on Human Computation, pp. 43–48 (2011)
39. Passonneau, R.J., Bhardwaj, V., Salleb-Aouissi, A., Ide, N.: Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. Language Resources and Evaluation **46**(2), 219–252 (2012). DOI 10.1007/s10579-012-9188-x
40. Passonneau, R.J., Carpenter, B.: The benefits of a model of annotation. Transactions of the ACL **2**, 311–326 (2014)
41. Poesio, M., Artstein, R.: Anaphoric annotation in the arrau corpus. In: Proceedings of the sixth International Conference on Language Resources and Evaluation. Marrakesh (2008)
42. Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L.: Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. ACM Transactions on Intelligent Interactive Systems **3**(1) (2013)
43. Quinn, A.J., Bederson, B.B.: A taxonomy of distributed human computation. Tech. rep., University of Maryland, College Park (2009)

44. Rafelsberger, W., Scharl, A.: Games with a purpose for social networking platforms. In: Proceedings of the 20th Association for Computing Machinery (ACM) conference on Hypertext and hypermedia, pp. 193–198. ACM (2009)
45. Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. Journal of Machine Learning Research **11**, 1297–1322 (2010)
46. Singh, P.: The public acquisition of commonsense knowledge. In: Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access. Palo Alto, CA (2002)
47. 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 '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics, Morristown, NJ, USA (2008)
48. Uebersax, J.S., Grove, W.M.: A latent trait finite mixture model for the analysis of rating agreement. Biometrics **49**, 832–835 (1993)
49. Venhuizen, N., Basile, V., Evang, K., Bos, J.: Gamification for word sense labeling. In: Proc. of the 10th IWCS, pp. 397–403. Potsdam, Germany (2013)
50. Wang, A., Hoang C. D. V. Kan, M.Y.: Perspectives on crowdsourcing annotation for natural language processing. Language Resources and Evaluation **47**(1), 9–31 (2013)
51. Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In: Advances in Neural Information Processing Systems, vol. 22, pp. 2035–2043 (2009)