1

An Overview

Maxine Eskénazi
Carnegie Mellon University, USA

In the early days of automatic speech processing, researchers dealt with relatively small sets of speech data. They used them mainly to build small automatic systems and to test the systems' validity. The data was often obtained by recording speakers in an anechoic chamber on magnetic tape. It was manually sent to a freestanding spectrogram machine in the 1960s and 1970s or input to a computer in the late 1970s and thereafter. Getting speakers (other than colleagues and students) took time, and labeling the speech that was acquired took much more time. Both endeavors were very costly. These difficulties entered into consideration every time a researcher planned a project, often imposing limitations on the amount of data collected and the scientific goals.

As time went on, one factor that dramatically increased the need for more data was the success of statistically based methods for automatic speech processing. The models for these systems, which quickly became ubiquitous, needed large amounts of data. The expression "more data is better data" was born. As automatic speech processing researchers switched from one application, like Broadcast News, to another, like Communicator, they found that the data from the former application was not very useful for the new one. As data specific to the new application was collected, processed, and fed into speech systems, results improved.

At the same time, other speech research publications (speech synthesis, spoken dialog systems, perception, etc.) also included some assessment. This required increasing numbers of speakers, callers, and judges, and thus a significant investment in data resources. This investment involved researcher time as they found and recorded subjects, as they trained transcribers to write down exactly what had been said, and as they found subjects to try out the resulting systems. Besides the time of the researcher, the investment also included the payment of the speakers and the transcribers, and sometimes a company was engaged to either recruit speakers or to manage transcription, thus adding to the costs.

Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, First Edition.
Edited by Maxine Eskénazi, Gina-Anne Levow, Helen Meng, Gabriel Parent and David Suendermann.

As Google became a major search engine of choice, it gathered a very large amount of data, larger than any that had ever been used before. Google researchers produced conference papers that demonstrated to their colleagues that some previously unsolved issues in natural language processing became surmountable just by using several orders of magnitude more data (Brants *et al.* 2007). The field became ripe for a solution that would provide more processed data at significantly lower cost. The present estimate of the cost of transcribing 1 hour of speech data by an expert (for ASR training) is 6 hours of transcription time for each actual hour of speech that is processed, at a cost of $90–$150 per hour (Williams *et al.* 2011).

At the same time, linguists and sociolinguists, freed from the use of magnetic tape and the onerous postprocessing that accompanied it, found that recording speech directly on computers enabled them to rapidly obtain large samples of the speech that they wanted to study. Many speakers with more diverse backgrounds could be recorded. Several groups of speakers with varying characteristics could be recorded instead of just one. However, as the need for more speakers and more transcriptions of their speech increased, these communities ran up against the same obstacles that the automatic speech processing community had encountered.

What seems like the answer to these needs has come in the form of crowdsourcing. This technique offers the promise of dramatically lowering the cost of collecting and annotating speech data. Some of the automatic speech processing community has quickly embraced crowdsourcing. This chapter and the next will give a short history and description of crowdsourcing, some basic guidelines, and then review the uses of crowdsourcing for speech that have been published in the past few years.

1.1 Origins of Crowdsourcing

What may be one of the earliest examples of the use of the crowd is the open call that the Oxford English Dictionary (OED) made to the community in the 1800s for volunteers to index all of the words in the English language and to find example quotations for each of the uses of each word (Wikipedia 2012).

More recently, James Surowiecki's (2004) book, *The Wisdom of Crowds,* gives an explanation of the power of the wisdom of the crowd. It maintains that a diverse collection of opinions from people who are making independent decisions can produce some types of decisions better than obtaining them from experts. Surowiecki sees three advantages to what he terms disorganized decisions: **cognition** (thinking and information processing), **coordination** (optimization of actions), and **cooperation** (forming networks of trust with no central control).

A good example of cooperation in a disorganized decision is the US *Defense Advanced Research Projects Agency* (DARPA) experiment in crowdsourcing to mark the 40th anniversary of the Internet. The goal was to locate 10 balloon markers that had been placed in a variety of locations across the United States. Teams were formed, each vying to be the first to find all 10 markers. This required collaborative efforts with networks of informers in many locations across the country. The team from MIT had the shortest time (under 9 hours). Its groups, comprised friends, and friends of friends, signed up to help locate the balloons. This underlines the observation that, in a crowd situation, where each person is independent and fairly anonymous, an individual will give their knowledge and opinions more freely. The success of the endeavor centered on this generous participation. Indeed, authors of crowdsourcing tasks who ask the members of their crowd if they have suggestions on how to improve a task (without giving them additional remuneration) often find that some of the crowd will take the time to make very insightful and helpful suggestions.

1.2 Operational Definition of Crowdsourcing

The operational basis of crowdsourcing rests on the idea that **a task** is to be done, there is a means to **attract many nonexperts** to accomplish this task, and that some **open call** has gone out to advertise the task to the nonexperts (Wikipedia 2012). The presence of the Internet and cellphones facilitates not only the open call for nonexperts but also the presentation of the task, its accomplishment, and the accumulation of the nonexperts' opinions. The nonexperts also possess some relevant knowledge, be it only that they are native speakers of a given language. From these assumptions, it is believed that the aggregate opinion of many nonexperts will approach the quality of the opinion of an expert. It is also believed that the use of nonexperts in this manner will be less onerous and more rapid than the use of experts. Given this economy of means, it is understandable that the speech and language processing communities have seen crowdsourcing as a possible solution to their large data dilemma. To illustrate the operational aspects of crowdsourcing, consider a task that comes up at barbeques and other social events. A total of 672 jellybeans have been put into a clear jar. Each person attending the barbeque is asked to estimate how many jellybeans are in the jar. They are aware that there is something that the organizer of the barbeque wants them to do (the open call for nonexperts). They are also aware that they are to provide an estimate of the number of jellybeans in the jar (the task), and they know how to count jellybeans or make estimates (have some expertise). They put their answers on a piece of paper and put that into a box. The organizer looks at all of the tickets in the box and finds answers like 300, 575, 807, 653, 678, 599, and 775. The aggregate answer, such as the average (626), or the median (653), is very close to the real number of jellybeans in the jar. Thus, the conditions that characterize crowdsourcing are:

- A task.
- An open call.
- Attracting many nonexperts.

So we will define a crowd as a group of nonexperts who have answered an open call to perform a given task.

1.3 Functional Definition of Crowdsourcing

A functional view of crowdsourcing, from Surowiecki, defines four characteristics of the *wise* crowd. First, the members of any crowd have a **diversity of opinions**. The opinions may be only slightly different from one another, and some may be correct while others are wrong. Second, each member of the crowd has an opinion that is **independent** of all of the other members of the crowd. No member's opinion is influenced by that of any other member. Third, information that the crowd may have is **decentralized**. Everyone has some local information, but no one in the crowd has access to all of the information that may be pertinent to the task. Finally, the opinions of the members of the crowd can be merged to form an **aggregate**, one collaborative solution. To illustrate this, here is an example where the crowd has the task of translating some text. If we have the following text in French,

> *Je pense qu'il est temps de partir. On prendra congé de ma mère et de ma sœur et on se mettra en route au plus tard à neuf heures.*

we can ask a crowd, that is, English and French speaking, for its translation. Some of its members may offer these four solutions:

S1: I think it's time to leave. We will say goodbye to my mother and my sister and get going at 9 a.m. at the latest.
S2: The time has come to leave. Let's say goodbye to my mother and my sister and be on our way by 9 a.m.
S3: We need to go. We'll say goodbye to my mother and my sister and leave by 9 a.m.
S4: Let's go. Take leave of my mother and my sister and be on our way by 9 a.m.

The four aspects of the functional nature of crowdsourcing are illustrated here. We can see that the four solutions offered by members of the crowd (S1–S4) reflect *diverse opinions* on exactly what the right translation is. We also can imagine that these opinions have been arrived at *independently* from one another. Each member of this crowd possesses some *individual pieces of information* that they are using when forming their opinion. S1, for example, may reflect the idea that "il est temps de partir" should be translated literally as "it's time to leave" while S4 may reflect a broader definition, which, in this context, results in the expression "let's go." Finally, we can *merge* these opinions to form one solution by, for example, asking the members of another crowd to vote on which one they like the best, by choosing the one that is most frequently produced, or by using a string-merging algorithm such as Banerjee and Lavie (2005). There has also been work (Kittur *et al.* 2011; CastingWords 2012) on having the crowd collaborate with one another to make the translation evolve into something on which they can all agree.

Thus, according to Surowiecki, the four characteristics of a wise crowd are:

- Has a diversity of opinions.
- Each individual works independently of the others.
- The information is decentralized.
- An aggregate solution can be formed.

1.4 Some Issues

While crowdsourcing seems to be a remarkable solution to the problems plaguing the speech and linguistics communities, it must be approached with care since misleading or incorrect results can also easily be obtained from crowdsourcing. Several issues should be kept in mind to prevent this.

The first issue concerns **the amount of information** given to the crowd. The crowd should be given just enough information to be able to complete the task, but not enough to influence their decisions. For the translation task above, for example, although the creator of the task could ask for a translation that is as literal and close to the original text as possible, this additional information may make the final result less desirable. The workers' opinions should not be influenced by information from the task creator. The second issue is having a crowd that is **too homogeneous**. A crowd that is too homogeneous will not give a superior result. Oinas-Kukkonen (2008) has found that the best decisions come when there is disagreement and contest within the crowd. Note that giving too much information is one way that a crowd may be rendered too homogeneous. He mentions another issue that contributes to homogeneity—**too much communication**. When members of the crowd have less anonymity and the creator of the task has more communication with the crowd, too much information may gradually be transmitted. Linked to the issue of communication is that of **imitation**. If participants are given access to the opinions of other workers, they may be influenced by them and, consciously or not, imitate what they have seen (thus leading us back to a more homogeneous crowd). While some tasks are given to one crowd and then the result is sent to another crowd for verification, some mechanism should be in place to avoid having this influence.

A fifth issue that should be addressed concerns the **prerequisites of the crowd**. We have seen that the members of the crowd are presumed to have some local knowledge. It is not evident that everyone who responds to a call has that knowledge. For example, in the above translation task, the creator of the task will assume that the participants speak both French and English. Since there may be some impostors in the crowd, it is wise to give some sort of pretest. We will discuss this further in Chapter 2. Pretests of performance on the specific task at hand are a reasonable way to winnow out those who may not be able to perform the task. However, creation and checking of the pretest is onerous in itself and it may be more time- and cost-saving to let all who respond complete the task and then eliminate outlier answers later on.

Another issue is **motivation**. Why should someone participate in a task? Are they learning something, playing a game, being remunerated? There should be some reason for an individual to not only sign up to work on a task but also want to continue to work on it. This is linked to a seventh issue, keeping a **reasonable expectation of the work effort**. If members of the crowd are lead to believe that there is less work than what is actually expected, especially in the case of remunerated work, they will quit the task and recommend to others (via worker forums and blogs) that they also avoid this task.

Finally, as we will see in several chapters in this book, it is absolutely necessary to carry out some form of **quality control**. This control can come in many forms and it is meant to weed out the work of poor workers (who have good intentions, but who furnish work that is not of good quality) and malicious workers (those who randomly enter answers or automated bots).

When reading research papers that incorporate crowdsourcing results, it is wise to determine whether these issues have been dealt with since this may affect the wellfoundedness of a paper.

Therefore, before proposing a task, researchers should deal with the following issues:

- Giving the crowd too much information.
- A crowd that is too homogeneous.
- Having too much communication with the crowd.
- Avoiding the possibility of imitation.
- Requesting prerequisites from the crowd.
- Maintaining crowd motivation.
- Presenting a reasonable expectation of workload.
- Conducting quality control.

1.5 Some Terminology

At this point, a short discussion of terminology is useful. The person who is creating the task and who submits it is called (at Amazon Mechanical Turk, MTurk, in this book, a platform that is used for crowdsourcing) the requester. This person may be called the client at other crowdsourcing sites. Herein we will use the term *requester*. The person in the crowd who does the work is appropriately called the *worker* (some also say *turker*) at MTurk and other sites, a *freelancer* at MiniFreelance, and a *contributor* at CrowdFlower and elsewhere. We will use the term *worker*. The individual task itself is called a *Human Intelligence Task* or *HIT* at MTurk, a *mission* at AgentAnything.com, a *microjob* at MicroWorkers, and a *task* at CrowdFlower. We will use the term *task* here, but the reader will also see this term broken down into three types of tasks, according to granularity:

- *Set of tasks* is the complete set of items that the requester wants to have done. For example, the transcription of 2000 hours of speech.
- *Unit task*, for example, transcribing one utterance out of the 2000 hours of speech in the set of tasks above.
- *Assignment* is one piece of work within the unit task; that is, the transcription of the above unit task of one utterance may be assigned to three workers, thus there would be three assignments for one unit task.

Also, when referring to the number of unit tasks completed per hour, we will use the term *throughput*. When referring to when the requester makes a set of tasks available to the workers, we will use the term *submission*. When talking about the agreement between multiple workers in an annotation task, we will use the term *interannotator agreement* (ITA).

1.6 Acknowledgments

This work was supported by National Science Foundation grant IIS0914927. Any opinions, findings, and conclusions and recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

References

Amazon Mechanical Turk—Artificial Artificial Intelligence. http://mturk.com (accessed 9 July 2012).

Brants T, Popat AC, Xu P, Och FJ and Dean J (2007) Large language models in machine translation. *Proceedings of the Conference on Empirical Methods on Natural Language Processing(EMNLP-2007).*

Banerjee S and Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.*

CastingWords. http://castingwords.com (accessed 9 July 2012).

Kittur A, Smus B and Kraut R (2011) CrowdForge: crowdsourcing complex work. *Proceedings of the ACM 2011 Annual Conference on Human Factors in Computing Systems.*

Oinas-Kukkonen H (2008) Network analysis and crowds of people as sources of new organizational knowledge, in *Knowledge Management: Theoretical Foundation* (eds A Koohang *et al.).* Informing Science Press, Santa Rosa, CA, pp. 173–189.

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday Anchor.

Wikipedia—Crowdsourcing. http://en.wikipedia.org/wiki/Crowdsourcing (accessed 9 July 2012).

Williams JD, Melamed ID, Alonso T, Hollister B and Wilpon J (2011) Crowd-sourcing for difficult transcription of speech. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*.