Motivational Feedback in Crowdsourcing: a Case Study in Speech Transcription

G. Riccardi, A. Ghosh, S.A. Chowdhury, A.O. Bayer

DISI - University of Trento, 38050 Povo (Trento), Italy

{riccardi,aghosh,sachowdhury,bayer}@disi.unitn.it

Abstract

A widely used strategy in human and machine performance enhancement is achieved through feedback. In this paper we investigate the effect of live motivational feedback on motivating crowds and improving the performance of the crowdsourcing computational model. The provided feedback allows workers to react in real-time and review past actions (e.g. word deletions); thus, to improve their performance on the current and future (sub) tasks. The feedback signal can be controlled via clean (e.g. expert) supervision or noisy supervision in order to trade-off between cost and target performance of the crowd-sourced task. The feedback signal is designed to enable crowd workers to improve their performance at the (sub) task level. The type and performance of feedback signal is evaluated in the context of a speech transcription task. Amazon Mechanical Turk (AMT) platform is used to transcribe speech utterances from different corpora. We show that in both clean (expert) and noisy (worker/turker) real-time feedback conditions the crowd workers are able to provide significantly more accurate transcriptions in a shorter time.

Index Terms: speech transcription, crowdsourcing, feedback systems

1. Introduction

Crowdsourcing platforms introduce a new computational model to distribute task executions to crowds of workers. Such a computational framework is innovative with respect to the scale of the human population, diversity of demographics, skill sets and fast turnaround times for task completions. Last but not the least such human computational power can be combined with machine-driven computational models and used to train best-of-its-parts human-machine processors. However, human-machine processors are still in its infancy and many issues have to be addressed including the monitoring and control of crowd sourced task performances. Amazon Mechanical Turk [1] is one of the most popular web-based crowdsourcing platforms. It provides a customizable user interface for requesters to post tasks as a unit of work called HIT (Human Intelligence Task). Each HIT is then distributed via the web to be performed by anonymous workers (who are called turkers) for a fixed monetary reward.

This mechanism is usually ideal for performing a variety of short tasks, like image labeling, annotation and language tasks such as speech transcription and translation [2][3][4][5][6]. These tasks are easy for humans to do but in most cases it is difficult or expensive to find or to hire experts on-demand. So, the power of the crowd can be harnessed to perform such tasks. One of the most critical open issues in crowdsourcing is quality monitoring and control. Although crowdsourcing a task, is fast and cheap, most of the time it is not reliable in terms of quality. Needless to say, anonymous turkers on a crowdsourcing platform like AMT belong to different ages, education levels, skill groups and most of the turkers might need training for the target requested task [7]. Beside these issues, we must also account for and filter out spammers. The reward model in a platform such as AMT is task based. Thus turkers' goal is to complete the job in the shortest amount of time.

In this paper, we explore how to motivate turkers to improve their performance by giving them real-time knowledge of their task accuracy. We study this problem in the context of a speech transcription task. The knowledge is provided to the turker via real-time evaluative feedback. The feedback signal may allow them to act in real-time, review past actions (e.g. word deletions) and improve the performance of current and future (sub) tasks. The feedback may be controlled via clean (e.g. expert supervision) or noisy (e.g. turker supervision) signal in order to trade-off cost of the feedback and target performance. In the first case, turkers are provided real time feedback using experts' transcriptions while in the second case the transcriptions of other turkers (the crowd) are elaborated to generate real-time feedback.

In the next sections, we describe the background on the crowdsourcing task management in Section 2, followed by experimental design in Section 3. The evaluation of transcription set quality for real time conditions using clean and noisy feedback signals is mentioned in Section 4. In Section 5, we describe the task analysis descriptors of turkers, followed by the main discussion and conclusion in Section 6 and Section 7.

2. Background

In the literature various approaches are proposed in order to increase the quality of transcriptions on crowdsourcing platforms such as AMT [5][8][9][10][11][12]. A comprehensive study of NLP tasks on Amazon Mechanical Turk shows that using high agreements among non-expert turkers can reach the experts' performance [13]. Collecting several judgments and selecting the transcription with the highest agreement for each of utterances can result in a decrease in the Word Error Rate (WER) [9][10]. However, this requires collecting a high number of judgments, which is expensive in terms of time and cost.

In [14], a Recognizer Output Voting Error Reduction (ROVER) system is introduced to improve the quality of transcription. M. Marge et. al [8] have shown that the ROVER system improves the quality of tasks by combining judgments that are collected independently from distinct crowd workers. Another approach in [10] uses Gold Standard (GS) references to avoid collecting poor judgments. Therefore, transcriptions performed by turkers with scores lower than a predefined threshold, for these Gold Standard utterances, can be rejected. In methods that use unsupervised Gold Standard quality control, transcripts with the highest agreement among the turkers could be considered as Gold Standard reference. More recently Lee et. al [11] have explored the effect of feedback for collecting high quality transcriptions in a two-stage transcription quality control process.

In experimental psychology and organizational science, the role and effect of *knowledge of results* have been studied in the context of single user and/or group of users' task management. Feedback is the information provided to an individual for the purpose of improving performance [15]. Feedback can provide information about the type and extent of errors, which can be corrected. Feedback allows individuals to compare the result of their actions with predefined goals and helps them to adjust their actions or their targets [16]. Feedback has also been shown to have motivational effect by providing a sense of competence and achievement in workers [17]. According to Locke et. al [19], the motivational effect and subsequent performance improvement provided by feedback should actually be attributed to the goal-setting effect of feedback. Most of these experimental studies have been conducted in controlled settings.

In [18], to achieve the high quality product review, three types of feedback mechanisms, such as self-assessment, external assessment and expert assessment, are provided for turkers. The results show task-specific feedback can help and train the turkers to produce better results over time. However, the use of an external expert in this experiment highly limits the scope and increases the cost of the experiment.

3. Experimental Design

The basic idea behind providing feedback for turkers is to discover how the power of the crowd itself can be used to generate the (on-line) training signals for turkers. The provided signal, in the form of feedback helps turkers to improve their performance in real time conditions. In our transcription task, the posted task is distributed among workers via the available crowdsourcing framework and the annotated output is collected. Then, by using consensus-based algorithms, such as ROVER, in our case, the provided output is enhanced automatically. This human-machine computational architecture aims at improving the quality of the *teaching* signal. Such teaching signal is processed and presented to the crowd via textual and visual realizations. The conceptual flow of our model is illustrated in Fig. 1. For the task we have asked turkers to transcribe speech utterances in order to explore and compare the effects of no feedback and live (expert and turker generated) feedback conditions on turkers' performance. Although we apply it in the transcription domain, the computational architecture is general and applicable to most crowdsourcing tasks.

3.1. Corpora

To evaluate the proposed model, utterances are selected from two publicly available corpora: the Air Travel Information System (ATIS) which includes 981 speech utterances from a flight information-seeking task (Dec. 1994 Test-Set), and the Wall Street Journal (WSJ) which includes 7500 speech utterances collected from spontaneous dictation of journalists with different narration skill levels. For our experiments, we have selected a total of 1000 utterances from the ATIS and WSJ (500 from each set) sets with similar characteristics in terms of speaking rates, duration and number of tokens per utterance. All speech files from these corpora include utterances spoken by native speakers. Table 1 shows statistics of the utterances used in our experiments.

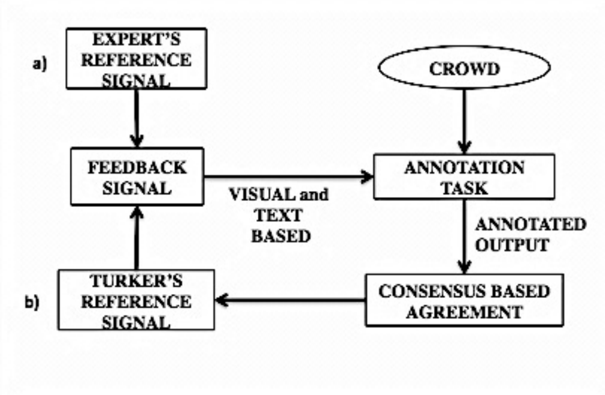

Figure 1: Turkers are provided different reference signals and they receive feedback instantaneously from a) experts or b) turkers.

	ATIS	WSJ
Length of audio files (sec)	7.3	6.6
No. of tokens per utterance	14.6	18.1
Speaking rate (token/sec)	2.2	2.8

Table 1: Average characteristics of selected speech corpora for the experiments

3.2. Task Preparation

To evaluate the quality of transcription done by turkers in various feedback conditions, we have designed and posted tasks containing short speech segments. In the task, each HIT is composed of 4 pages, each page having 8 such small speech files, including 1 gold. The transcription tasks have been submitted in a way that each turker has transcribed a maximum of 32 speech utterances per HIT where each task is split into 5 HITs. All the HITs have been restricted to US.

Each utterance transcription was rewarded with $0.02. Experimental experience shows that this amount is a reasonable amount to pay for such a small task and no further performance improvement is found with higher rewards.

To conduct this experiment at least three transcriptions per utterance have been requested and collected. We posted our tasks to AMT through the Crowdflower platform. Crowdflower allows us to provide a set of Gold Standard utterances and it automatically filters out turkers unable to transcribe the Gold Standard transcriptions properly. The use of Gold Standard utterances has been shown to be an effective method to eliminate spammers. All the parameters and conditions are assumed to be the same in all the experiments to make them comparable.

3.3. Feedback Signal

We have performed two distinct categories of experiments: with feedback and with no feedback. In the "with feedback" scenario, a performance meter is placed below the text area for transcription to provide live feedback (LV) to the turker. The turker is provided both visual and textual feedback regarding the transcription as he types. The visual feedback consists of a performance bar as shown in Fig. 2, which changes color and size as a function of the Word Accuracy (WA) of turker's transcription.

We compute the Word Accuracy as a keystroke event is detected and update the performance bar to provide visual feedback. The partial string word accuracy is binned into five intervals: 0-20%, 21-40%, 41-60%, 61-80% and 81-100%. The quantized value is coded into the color values ranging from red to green and is applied to the performance meter on the turker's client. The textual feedback is also displayed on the turker's client, for each word accuracy intervals: "Very Poor" to "Very Good".

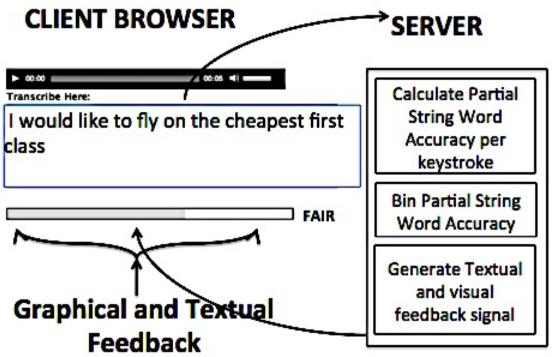
CLIENT BROWSER
SERVER
00:00
00:05
Transcribe Here:
I would like to fly on the cheapest first class
FAIR
Graphical and Textual Feedback
Calculate Partial String Word Accuracy per keystroke
Bin Partial String Word Accuracy
Generate Textual and visual feedback signal

Figure 2: Real time visual and textual feedback mechanism given to the turker

3.4. Quality of On-line Feedback

We have used two types of signals for estimating the motivational feedback:

1. Experts' Reference: The reference transcriptions transcribed by experts are used to compute the word accuracy. This is high quality feedback signal, although expensive. This scenario provides an upper bound on the crowdsourcing system performance in relation to the quality of the reference signal. It may be generally extended to trade-off reference quality and task costs by sampling expert signal over time.

2. Turkers' Reference: The transcriptions provided by the turkers who have passed the Gold Standard (GS) transcription test are used as reference when computing the word accuracy. In this scenario, we have collected the output of ROVER when the feedback is not provided and then these transcriptions are used as reference to provide feedback signal for turkers. This is a reasonable reference since we have used GS utterances to remove spammers and low-quality turkers.

4. Transcription Quality Evaluation

In two distinct categories of experiments with and without feedback, a total of 159 turkers participated, where 5 turkers participated in 3 of the tasks and around 20 turkers participated in 2 common tasks with a negligible overlapping between the audio files. For each task a minimum of 3 transcribed utterances per audio is collected. The evaluation of the collected utterances is done comparing with expert reference transcription. The results in terms of Word Error Rate (WER), for different experimental conditions, are shown in Table 2.

Referring to Table 2. It is observed that the overall WER of a task with no feedback is 7.65%, whereas a decrease in WER to 5.99% occurs when a live feedback, with turkers' reference, is provided. A further improvement in WER is seen when expert's reference is used as a live feedback signal, which is 4.18%.

The reason behind this decrease in WER with feedback can be explained as an encouragement strategy that motivates a turker to revise, correct and improve over the mistakes in his transcription or give assurances to a turker that his work maintained a certain quality. This is in agreement with the goal-setting phenomenon of feedback mentioned by Locke [19]. The goal of the turker in this case is to achieve a high accuracy while performing the transcription task. And by providing continuous knowledge of results in the form of feedback, a turker becomes aware of the quality of his transcription, and hence tries to improve it by correcting his mistakes, if needed.

A difference of 1.81% of WER is observed between the two feedback tasks. It is seen that this difference is due to the fuzziness in the feedback signal of live feedback with turkers' reference, where there is a small number of misleading signals that confuse a turker and also leads to a slight increase in the time to complete his job.

In order to show that the results in Table 2 are statistically significant, we used the software developed by Sebastian Pado (http://www.nlpado.de/~sebastian/software/sigf.shtml). The algorithm is based on the computationally intensive randomization test presented in [20]. It shows that in the three given transcriptions sets (no feedback vs. experts' feedback, no feedback vs. turkers' feedback and experts' feedback vs. turkers' feedback) the improvement is statistically significant ($p \leq 0.05$). The result is presented in Table 3.

Method	Overall	ATIS	WSJ
No Feedback	7.65	5.67	9.21
Live Feedback (Turkers' ref.)	5.99	3.74	7.76
Live Feedback (Experts' ref.)	4.18	2.02	5.88

Table 2: Word Error Rate (%) for different feedback conditions

Method	Overall
No Feedback vs. Experts' Feedback	9.99E-05
No Feedback vs. Turkers' Feedback	9.99E-05
Turkers' Feedback vs. Experts' Feedback	9.99E-05

Table 3: The p-value comparison for different feedback conditions

Method	Overall	ATIS	WSJ
No Feedback	5.27	3.70	6.51
Live Feedback (Turkers' Ref.)	4.47	2.57	5.96
Live Feedback (Experts' Ref.)	3.01	0.93	4.66

Table 4: Table 4: WER (%) following ROVER in different feedback conditions.

4.1. ROVER Accuracy

As mentioned earlier, a minimum of 3 transcriptions per utterance are requested and collected from turkers. This allows for an automated computation of a consensus-based hypothesis to improve the quality of our transcription tasks by applying ROVER [14]. The obtained results show a considerable improvement in WER. Table 4 shows results of combining collected judgments and applying ROVER for different types of feedback condition in different corpora.

5. Task Analysis

We have analyzed selected sub-task statistics to evaluate the impact of feedback on *how* the task was accomplished. We have computed the typing rate and deletion rate observed as the turkers completed their speech transcription tasks.

We define “Typing Time” as the time taken by the turker from the moment he makes the first keystroke until the last keystroke for a single utterance. It should be kept in mind that this time interval might contain other events such as the turker replaying the audio file or searching online for the correct spelling of a Named Entity. So at best this might lead to an underestimation of the true Typing Rate. We observe that in a no feedback condition the average Typing Time of a turker per utterance is 119.5 seconds for ATIS and 163.3 seconds for WSJ. When expert feedback is provided, this time is considerably reduced to 61.7 and 101.4 seconds for ATIS and WSJ respectively. Fuzzy feedback from turkers’ reference leads to a slight increase in the time to 66 seconds and 97.6 seconds for ATIS and WSJ respectively. Based on this we report the Typing Rates (in characters per minute) in Table 5. And in Figure 3, a simple distribution of turkers is given, on the basis of their average Typing Rate. The figure gives a brief overview to support the evidence of Typing Time analysis, showing that for no feedback condition, maximum number of turkers have a Typing Rate of 85 char per min, whereas for feedback condition with Experts’ reference most of the turkers are distributed around 260 char per min and above. In case of live feedback with turkers’ reference, the distribution is evenly distributed between the no feedback and feedback with experts’ reference case.

6. Discussion

Providing feedback helps turkers to complete tasks more accurately and faster. This may be attributed to the fact that feedback enables the turkers to have a continuous knowledge of results. In the experiments where feedback is provided, it is observed that turkers are able to achieve a higher accuracy as well as high typing rates. From tables 4 and 5 we can see that the best performances are obtained when experts' feedback is provided. We see low Word Error Rates: ATIS (0.93) and WSJ (4.66) as well as high typing rates: ATIS (293.75) and WSJ (222.64). When turkers’ references are used to provide feedback, the accuracy and typing rates show a slight decrease. However, we find that the Typing rates for ATIS and WSJ are still significantly higher in case of turkers’ feedback (255.04 and 215.11 respectively) as compared to when no feedback is provided (167.40 and 138.28 respectively).

Considering the number of deletes per character for the final submitted utterance, we observe that the turkers delete more when no feedback is provided. This again can be attributed to uncertainty. In most cases, they are often unsure about whether they have written a certain word (mostly out of dictionary words) correctly, and hence edit more. This effect decreases upon providing feedback. We observe that when no feedback is provided, some of the common mistakes are misspelling of named entities and wrong way of writing abbreviations and numbers. These mistakes are avoided when any type of live feedback (expert’s or turker’s) is provided to the turkers. Thus feedback acts as a teaching signal for the turker, where he is able to learn from the feedback to perform a task better and faster.

From our experiments we can conclude that live feedback helps the turker to do proper goal setting, and hence perform the transcription task better. In most practical scenarios it is not possible to provide large-scale feedback from experts. We have shown that the feedback signal may be computed from turkers’ outputs and that is an acceptable strategy in terms of efforts and resource trade-off.

Corpora	**ATIS**			**WSJ**		
	NFB	**TFB**	**EFB**	**NFB**	**TFB**	**EFB**
TR (char/min)	167.4	255.0	293.8	138.3	215.1	222.6
Del./char	0.3	0.2	0.1	0.3	0.2	0.2

Table 5: Typing Rate (char/min) and Deletion per character for different feedback conditions. : TR - Typing Rate, Del. - Deletion, NFB - No Feedback, EFB – Experts’ Feedback, TFB – Turkers’ Feedback

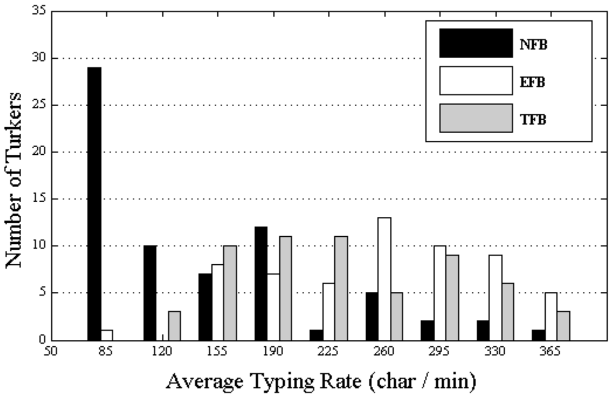

Figure 3: Distribution of turkers over Average Typing Rate (in char per min). NFB – No Feedback condition, EFB – Live Feedback condition (Experts’ ref), TFB - Live Feedback condition (turkers’ ref).

7. Conclusions

This paper presents experiments on crowd-sourced speech transcription task. The effects of different types and quality of feedback signals are evaluated with respect to the transcription quality and turkers’ behavior. The results show that live feedback has significant positive effect on both: quality of the transcription and turkers’ performance (typing rate). Moreover, it is observed that using “cheaper” feedback from turkers has an effect close to that of the expert feedback at a fractional cost. Feedback mechanism is demonstrated to be a promising approach for performance enhancement in crowdsourcing tasks.

8. Acknowledgements

This work was partially funded by PortDial FP7 project n. 296170.

9. Reference

[1] http://www.mturk.com.

[2] A. Gruenstein, I. McGraw, and A. Sutherland, "A self-transcribing speech corpus: collecting continuous speech with an online educational game," in International Workshop on Speech and Language Technology in Education, 2009.

[3] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization," in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, 2010.

[4] K. Evanini, D. Higgins, and K. Zechner, "Using amazon mechanical turk for transcription of non-native speech," in Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, 2010.

[5] J. Williams, I. Melamed, T. Alonso, B. Hollister, and J. Wilpon, "Crowd-sourcing for difficult transcription of speech," in IEEE Workshop on Automatic Speech Recognition and Understanding, Hawaii, USA, IEEE Workshop on Automatic Speech Recognition and Understanding.

[6] C. Callison-Burch, "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, Association for Computational Linguistics,
2009.

[7] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson, "Who are the turkers? Worker demographics in amazon mechanical turk," Department of Informatics, University of California, Irvine, USA, Tech. Rep, 2009.

[8] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010.

[9] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics,
2010.

[10] G. Parent and M. Eskenazi, "Toward better crowd-sourced transcription: Transcription of a year of the let's go bus information system data," in Spoken Language Technology Workshop (SLT), 2010 IEEE, IEEE, 2010.

[11] C. Lee and J. Glass, "A transcription task for crowd-sourcing with automatic quality control," in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[12] P. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in Proceedings of the ACM SIGKDD workshop on human computation, ACM,
2010.

[13] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast-but is it good? Evaluating non-expert annotations for natural language tasks," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,
2008.

[14] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on, IEEE, 1997.

[15] P. Earley, G. Northcraft, C. Lee, and T. Lituchy, "Impact of process and outcome feedback on the relation of goal setting to task performance," Academy of Management Journal, pp. 87–105, 1990.

[16] M. Campion and R. Lord, "A control systems conceptualization of the goal-setting and changing process," Organizational Behavior and Human Performance, vol. 30, no. 2, pp. 265–287, 1982.

[17] J. Hackman and G. Oldham, "Motivation through the design of work: Test of a theory," Organizational behavior and human performance, vol. 16, no. 2, pp. 250–279, 1976.

[18] S. Dow, A. Kulkarni, S. Klemmer, and B. Hartmann, "Shepherding the crowd yields better work," in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12, pp. 1013–1022, ACM, 2012.

[19] E. Locke, N. Cartledge, and J. Koeppel, "Motivational effects of knowledge of results: A goal-setting phenomenon?" Psychological Bulletin, vol. 70, no. 6p1, p. 474, 1968.

[20] A. Yeh, "More accurate tests for the statistical significance of result differences," in Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING '00, (Stroudsburg, PA, USA), pp. 947–953, Association for Computational Linguistics, 2000.