COMPARING TWO METHODS FOR CROWDSOURCING SPEECH TRANSCRIPTION

Rachele Sprugnoli[⋆]*, Giovanni Moretti*[⋆]*, Matteo Fuoli*[⋆]*, Diego Giuliani*[‡]*, Luisa Bentivogli*[‡ ⋆]*,*
Emanuele Pianta[‡ ⋆]*, Roberto Gretter*[‡]*, Fabio Brugnara*[‡]

[⋆] CELCT - Center for the Evaluation of Language and Communication Technologies
Via alla Cascata 56/C, 38123, Povo (TN), Italy
{sprugnoli,fuoli,moretti}@celct.it

[‡] Human Language Technology Research Unit - FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38123, Povo (TN), Italy
{giuliani,gretter,brugnara,bentivo}@fbk.eu

ABSTRACT

This paper presents the results of an experimental study conducted with the aim of comparing two methods for crowdsourcing speech transcription that incorporate two different quality control mechanisms (i.e. explicit *versus* implicit) and that are based on two different processes (i.e. parallel *versus* iterative). In the *Gold Standard* method the same speech segment is transcribed in parallel by multiple contributors whose reliability is checked with respect to some reference transcriptions provided by experts. On the other hand, in the *Dual Pathway* method two independent groups of contributors work on the same set of transcriptions refining them in an iterative way until they converge, and thus eliminating the need to have reference transcriptions and to check transcription quality in a separate phase. These two methods were tested on about half an hour of broadcast news speech and for two different European languages, namely German and Italian. Both methods obtained good results in terms of Word Error Rate (WER) and compare well with the word disagreement rate of experts on the same data.

Index Terms— Crowdsourcing speech transcription, Mechanical Turk, CrowdFlower, automatic speech recognition

1. INTRODUCTION

Recently, crowdsourcing has emerged as a promising alternative solution to the employment of well-trained expert transcribers for the creation of large corpora of transcribed speech at a relatively lower cost and turnaround time.

Crowdsourcing refers to the process of segmenting a complex task into smaller work units and distributing these among a large pool of non-expert workers, usually via the web. Recent years have witnessed a proliferation of online crowdsourcing platforms and services, the most popular being the Amazon Mechanical Turk (AMT). AMT is an online crowdsourcing marketplace where requesters distribute small tasks (Human Intelligence Tasks or HITs) to a large number of anonymous non-expert contributors [1, 2], who work in parallel on different portions of the same task, greatly speeding up its completion. Another popular provider of crowdsourcing services is CrowdFlower (CF). Differently from AMT, CF is not a crowdsourcing marketplace but a broker that aggregates a variety of web platforms, including AMT itself. CF provides requesters with an integrated and intuitive web interface for the design and distribution of tasks among AMT workers as well as through other online crowdsourcing marketplaces. The main strengths of CF reside in the possibility of launching tasks onto many different crowdsourcing platforms and a native automatic quality control mechanism based on gold standard units (see Section 3.2). Moreover CF is the main channel through which the AMT marketplace can be accessed by requesters that do not have an US credit card.

This paper presents the results of experiments conducted comparing two different methods for crowdsourcing speech transcription that incorporate two different quality control mechanisms. The first method is based on the iterative *Dual Pathway* process [3, 4], by which transcriptions are iteratively refined by two independent groups of annotators until the transcriptions made by each group converge. The second method relies on the automatic quality control mechanism based on *Gold Standard*, which is included in CF. These methods were tested with two languages, namely German and Italian, and under two different settings, that is *(i)* asking contributors to edit the transcriptions produced by an Automatic Speech Recognition (ASR) system and *(ii)* asking them to produce the transcriptions from scratch.

The main aim of this study is to test and compare the two methods and identify the one that achieves the best results in terms of transcription quality and cost.

The remainder of the paper is structured as follows: Section 2 presents a brief review of related work explaining how this contribution expands on prior studies in the field while Section 3 describes the two crowdsourcing methods tested in the experiments. Experimental setups and results are reported and discussed in Sections 4 and 5 respectively; conclusions and future perspectives are outlined in Section 6.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 287532, TOSCA-MP Task-oriented search and content annotation for media production (http://www.tosca-mp.eu).

This paper is in memory of our colleague and friend Emanuele Pianta.

2. RELATIONS TO PRIOR WORK

Crowdsourcing can contribute to substantially reducing the cost and time required for the creation of large corpora of transcribed speech. Recent studies demonstrate that transcriptions can be obtained for a fraction of the cost and processing time of conventional methods [5, 1, 6, 2]. However, one of the major challenges connected with crowdsourcing is quality control [6, 2], that is, ensuring that the transcriptions produced by non-expert contributors are accurate and complete. Several techniques for the control of the quality of crowdsourced transcriptions have been proposed. One of the most widely-used methods [1, 5, 7, 8, 9] consists of collecting a certain number of redundant transcriptions for each audio clip and combining them using string merging algorithms such as ROVER [10]. Some authors have developed a corrective workflow, whereby the same transcription is checked and iteratively refined by multiple contributors [11, 4, 2]. Parent and Eskenazi [6] employ an automatic quality control mechanism based on the concept of gold standard, whereby one utterance transcribed by an expert is inserted in each work unit and contributors' performance is evaluated in terms of how similar their transcriptions are to those produced by the experts. Other means of quality control proposed in the literature include the automatic poor quality transcript detection system developed by Lee and Glass [12], and the statistical regression model used by Williams et al. [2] to predict transcription reliability.

As regards language coverage of crowdsourcing experiments on speech transcription, most of the researche conducted so far focuses on English with some exceptions, for example on Mexican Spanish [8, 9] and on less-resourced languages such as Korean, Hindi and Tamil [1], and Amharic and Swahili [13].

With respect to previously cited works, this paper aims to contribute to the advancement of research on crowdsourcing techniques for speech processing by a) implementing and testing, for the first time, the iterative transcription workflow proposed by Liem et al. [4] in an online open crowdsourcing scenario, b) by assessing the viability of crowdsourcing for the collection of transcribed speech in languages other than English, and finally c) by assessing the viability of using the crowdsourcing platform CF, which is the main channel through which the AMT marketplace can be accessed by requesters that do not hold a bank account in the United States. In this way, this study aims to contribute to closing a growing gap in the speech research community between researchers working on English and holding a US bank account and all the others [6].

3. METHODS FOR CROWDSOURCING SPEECH TRANSCRIPTION

This section outlines the two methods that were tested for crowdsourcing speech transcription.

3.1. The iterative dual pathway method

This method is based on the iterative dual pathway algorithm presented in Liem et al. [4], by which transcriptions are iteratively refined by two independent groups of contributors until the transcriptions made by each group converge. The procedure works as follows (see Figure 1). The audio to be transcribed is partitioned into short clips. These clips are randomly assigned to contributors who are distributed into two independent transcription pathways (P1 and P2 in Figure 1). Contributors are asked to listen to an audio clip and edit the transcription made in the previous step (S1, S2, S3, S4 in Figure 1) of the same pathway. Transcriptions in one pathway are compared to those produced in the other pathway. When four transcriptions - two from each pathway - match each other, the audio clip is considered to have been transcribed correctly and is removed from the pool of clips to be processed. The assumption underlying this mechanism is that, since the transcription pathways are independent, the higher the convergence between the two pathways, the more accurate and reliable a transcriptions is. The key advantage of this method is that it enables to correctly evaluate transcription accuracy without having an explicit quality control, thus without the need of transcribing any clips in advance as in Method 2 below.

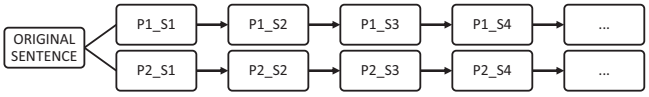

Fig. 1. Dual pathway scheme.

3.2. The gold standard method

This method is based on the gold standard quality control system embedded in CrowdFlower, and therefore it requires that at least 10% of the clips have been previously transcribed by an expert. The expert-made transcriptions are included in the transcription task as gold units. Gold units allow to distinguish between trusted contributors (those who correctly replicate the gold units) and untrusted contributors (those who fail the gold units). Gold units are included in the task in the form of a Boolean question, where contributors are asked to listen to an audio clip and judge whether the transcription provided is correct or not. Half of the transcriptions provided are correct and half are not. If contributors fail to provide a correct judgment for at least 70% of the gold units, they are considered unreliable and thus automatically excluded from the task. Only the transcriptions produced by reliable contributors are considered. Similarly to the dual pathway method, contributors are asked to transcribe several audio clips but following a parallel, non iterative, process. Finally, all transcripts are collected and merged using the ROVER algorithm, which allows to improve the accuracy of the final transcriptions through word-level voting.

4. EXPERIMENT SETUP

In the experiments, roughly half an hour of speech data were used for each of the two languages: German (about 4,700 words) and Italian (about 5,700 words). For each language, audio recordings were taken from television news broadcasts. These data were manually partitioned by an expert in segments ranging from 1 to 12 seconds in length: in particular, German speech data was split into 288 segments while Italian data was split into 313 segments. These segments were given as input to contributors on crowdsourcing platforms via CF: beyond AMT, for German transcriptions we also tested Crowd Guru[1], a medium-scale channel that reaches contributors mostly in Germany. Regional qualifications were applied to all HITs: countries that have contributors who tend to produce the majority of spam answers were excluded.

Both methods described in Section 3 were tested under two different settings, namely:

[1] http://www.crowdguru.de/

Results of Iterative Dual Pathway Method on German Data								
	Step 1		**Step 2**		**Step 3**		**Step 4**	
	ASR	SCRATCH	ASR	SCRATCH	ASR	SCRATCH	ASR	SCRATCH
#Conv.(%)	-	115(40.0%)	138(48.0%)	115(40.0%)	156(54.2%)	146(50.7%)	178(61.8%)	167(58.0%)
WER(%) Conv.	-	1.2	3.0	1.2	2.8	1.4	2.8	1.7
Global WER(%)	-	6.5	5.3	5.4	4.5	5.0	4.4	4.7

Table 1. Number, percentage and WER of converged segments together with the global WER achieved at each step for both settings of the dual pathway method on German data.

Results of Iterative Dual Pathway Method on Italian Data								
	Step 1		**Step 2**		**Step 3**		**Step 4**	
	ASR	SCRATCH	ASR	SCRATCH	ASR	SCRATCH	ASR	SCRATCH
# Conv. (%)	-	163(52.1%)	186(59.4%)	163(52.1%)	229(73.2%)	241(77.0%)	252(82.7%)	271(86.6%)
WER(%) Conv.	-	1.4	1.6	1.4	1.8	2.1	2.1	2.6
Global WER(%)	-	3.6	3.7	3.6	3.4	3.2	3.4	3.1

Table 2. Number, percentage and WER of converged segments together with the global WER achieved at each step for both settings of the dual pathway method on Italian data.

1. by asking contributors to correct the transcriptions produced by an ASR system;
2. by asking contributors to produce the transcriptions from scratch.

Automatic transcription of audio data was performed by using two transcription systems developed by Fondazione Bruno Kessler for German and Italian. This resulted in 17.1% and 10.4% WER for German and Italian, respectively. The difference in performance between languages is partly explainable by the different level of maturity of the transcription systems used. While the transcription system for Italian is well established and has been widely used over the years [14], the German transcription system was developed specifically for this work and needs further refinements.

For the gold standard method, the native CF interface was used and five transcriptions were collected for each segment in both the settings: all the transcripts of the same segment produced by multiple contributors were combined using the ROVER voting scheme. We tested the ROVER algorithm on various number of transcriptions per segment: the best results in terms of WER were achieved applying it on at least 5 contributors' transcripts, confirming the findings of Marge et al. [5] about the use of multiple transcriptions to improve ROVER accuracy.

As regards the dual pathway method, a dedicated database infrastructure and a web-based GUI for collecting transcriptions were created. The interface, available as an external HIT in CF, includes two automatic quality checks, namely a script that prevents contributors from submitting transcriptions before each audio has been played until the end, and a mechanism that forbids the submission of empty values. The infrastructure has built-in matching controls in order to quickly and easily manage the pathways: in particular, these controls apply filters to normalize insignificant differences before checking if transcriptions match each other. Filters, for example, were used to normalize numbers written with words or digits, double spaces, acronyms written with all letters in upper case or capitalizing only the first letter. On the other hand, differences in capitalization of proper nouns and in transcription of partial words and disfluencies were considered relevant so no normalization filters were applied.

WER	
ASR	**SCRATCH**
5.8%	3.8%

(a) Results on German

WER	
ASR	**SCRATCH**
3.1%	2.9%

(b) Results on Italian

Table 3. Results for the gold standard method on German (a) and Italian (b) data, starting from ASR transcriptions and from scratch.

In the case of setting (1) a transcription was considered correct when four transcriptions, i.e. two from each pathway, matched each other. In the case of setting (2), if at the first step two transcriptions produced from scratch by two independent contributors matched each other they were accepted as correct, without further iterations. This is justified by the fact that the first step is a generation task for which chance agreement is highly unlikely. Starting from the second step, contributors are asked to edit the transcriptions produced by the previous contributor and the standard criteria (i.e. four matching transcriptions) were applied. In both settings, transcriptions which did not converge after four steps were merged with ROVER. We decided to stop after four steps as preliminary experiments on German data with eight step pathways showed that after the fourth step the number of converged transcriptions and the improvement on global WER were not relevant.

5. RESULTS

This section presents the results obtained with the two crowdsourcing methods on Italian and German broadcast news.

Two expert native-speakers for each language transcribed the data: the word disagreement rate [15] between the two experts before the adjudication phase was used as an upper bound of what we could expect from non-experts, while the transcription provided after discrepancy rectification was used as a reference to determine the quality of crowdsourced data in terms of WER. Transcription guidelines for contributors were created based on those provided to the expert

who produced the reference transcripts, with a special emphasis on creating simple and intuitive instructions.

Tables 1 and 2 show the results of the dual pathway method for both settings and languages presenting the number, percentage and WER of converged transcriptions together with the global WER achieved on all segments (i.e. converged transcriptions plus not converged transcriptions merged with the ROVER algorithm). Improvements were registered during the iterative process, obtaining a global WER below 5% for German and below 3.5% for Italian after four steps. WER of converged transcriptions is always low (often below 2.0%): in particular, transcriptions matched in the first step of the dual pathway method starting from scratch show a very high quality with a WER of 1.2% for German and 1.4% for Italian.

Results for both settings of the gold standard method on German and Italian data are reported in Table 3. In most cases, WERs obtained with the gold standard method are lower to the ones obtained with the dual pathway method: in particular, asking contributors to produce the transcriptions from scratch proved to be the most promising setting with a 3.8% WER for German and an 2.9% WER for Italian.

Column charts in Figure 2 and 3 summarize the outcome of our experiments compared to the disagreement achieved by the two experts before the adjudication phase on discrepancies (i.e. 4.2% of WER on German data and 2.4% of WER on Italian data). Transcriptions collected using the methods described in this paper show a level of WER that approaches expert disagreement. The only exception is given by the WER of 5.8% achieved on German data with the gold standard method starting from the transcriptions of the ASR system. Probably contributors have been biased by the provided transcriptions which had a high WER (i.e. 17.1%).

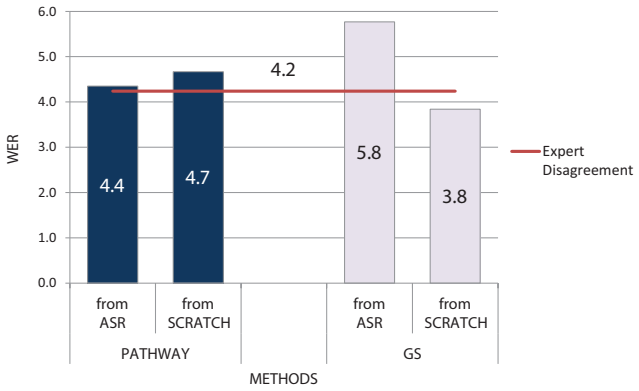

Fig. 2. Summary of results for German.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented the experiments conducted to test and compare two crowdsourcing methods in order to identify the one that achieves the best results in terms of transcription quality. The main differences between these two methods rely in the quality control mechanism they incorporate and in the process on which they are based (i.e. iterative *versus* parallel).

Liem et al. [4] implemented their dual pathway structure in a controlled environment, with a homogeneous group of well-educated participants (i.e. Harvard undergraduate students). In this study,

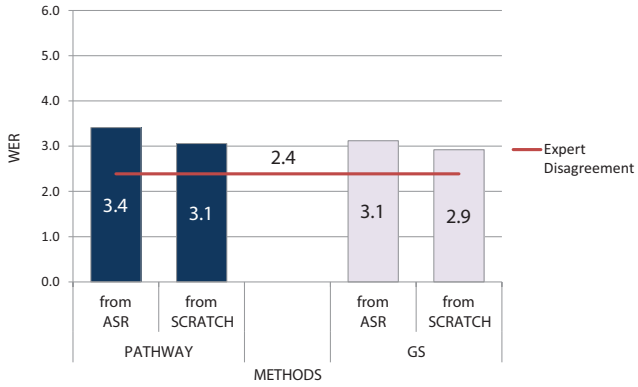

Fig. 3. Summary of results for Italian.

we tested for the first time the dual pathway system in an open online crowdsourcing marketplace, where the population is arguably more heterogeneous, using CF and comparing it to a gold standard method based on the quality control mechanism included in CF. All the works reviewed in Section 2 use AMT and, to our knowledge, no studies have been reported so far that use CF platform for speech transcription tasks. Assessing the viability and costs of CF for this particular application is of special interest to all Europeans who do not hold a bank account in the US and, therefore, cannot access the AMT marketplace directly.

Results show that the gold standard method starting from scratch produced the best quality transcriptions for both languages but also that transcriptions with near-expert quality in term of WER can be obtained through the iterative process. This outcome proves particularly useful in case no gold standard data is available. In general, crowdsourcing methods generate transcriptions with a much lower WER with respect to automatic transcriptions: more than 12 and 7 percentage points for German and Italian data, respectively.

As far as cost and time are concerned, it is important to note that the actual response of crowdsourced workforce to HITs is beyond the requester's control so it is not always predictable: cost, time and number of contributors usually change over time and are also correlated to the language under analysis. In particular, we noticed a great variability in terms of completion time while crowdsourcing costs ranged from 30 to 75 dollars per hour of speech. An additional cost required by the gold standard method is related to the production of reference transcriptions. In our study, the required effort was minimized reusing gold units; furthermore, it can also be reduced by utilizing crowdsourced transcriptions on which non-expert contributors obtained a perfect agreement. Overall, crowdsourcing costs are notably lower than costs for professional transcriptions, which can reach the average cost of 150 dollars per hour of speech [6].

As regards future efforts, we are now testing the two methods described in this paper on English and Flemish broadcast news. We also plan to work on talk show TV programs, which are characterized by conversational style. The ultimate goal is the transcription of data to be used for testing and training ASR systems.

7. REFERENCES

[1] S. Novotney and C. Callison-Burch, "Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription," in *Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, USA, 2010, pp. 207–215.

[2] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon, "Crowd-sourcing for difficult transcription of speech," in *Proceedings of IEEE ASRU workshop*, Hawaii, USA, 2011, pp. 535–540.

[3] G. Little, L.B. Chilton, M. Goldman, and R.C. Miller, "Exploring iterative and parallel human computation processes," in *Proceedings of HCOMP '10*, Washington DC, USA, 2010, pp. 68–76.

[4] B. Liem, H. Zhang, and Y. Chen, "An Iterative Dual Pathway Structure for Speech-to-Text Transcription," in *Proceedings of Human Computation AAAI Workshop*, Toronto, Canada, 2011.

[5] M. Marge, S. Banerjee, and A.I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *Proceedings of ICASSP*, Dallas, USA, 2010, pp. 5270–5273.

[6] G. Parent and M. Eskenazi, "Toward better crowdsourced transcription: Transcription of a year of the Let's Go Bus Information System data," in *Proceedings of SLT*, Berkeley, USA, 2010, pp. 312–317.

[7] K. Evanini, D. Higgins, and K. Zechner, "Using Amazon Mechanical Turk for transcription of non-native speech," in *Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, USA, 2010, pp. 53–56.

[8] K. Audhkhasi, P. G. Georgiou, and S. S. Narayanan, "Reliability-Weighted Acoustic Model Adaptation Using Crowd Sourced Transcriptions," in *Proceedings of INTERSPEECH*, Florence, Italy, 2011, pp. 3045–3048.

[9] K. Audhkhasi, P.G. Georgiou, and S.S. Narayanan, "Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics," in *Proceedings of ICASSP*, Prague, Czech Republic, 2011, pp. 4980–4983.

[10] J. Fiscus, "A post-processing system to yield reduced error rates: recognizer output voting error reduction (ROVER)," in *Proceedings of IEEE ASRU workshop*, 1997, pp. 347–354.

[11] M. Marge, S. Banerjee, and A.I. Rudnicky, "Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization," in *Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, USA, 2010, pp. 99–107.

[12] C. Lee and J.R. Glass, "A Transcription Task for Crowdsourcing with Automatic Quality Control," in *Proceedings of INTERSPEECH'11*, Florence, Italy, 2011, pp. 3041–3044.

[13] H. Gelas, S.T. Abate, L. Besacier, and F. Pellegrino, "Quality Assessment of Crowdsourcing Transcriptions for African Languages," in *Proceedings of INTERSPEECH'11*, Florence, Italy, 2011, pp. 3065–3068.

[14] F. Brugnara, D. Falavigna, D. Giuliani, and R. Gretter, "Analysis of the Characteristics of Talk-show TV Programs," in *Proceedings of INTERSPEECH'12*, Portland, OR, 2012.

[15] M. L. Glenn, S. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription Methods for Consistency, Volume and Efficiency," in *Proceedings of LREC'10*, Valletta, Malta, 2010.