情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-SLP-87 No.9
2011/7/22

Amazon Mechanical Turk を用いた音声データ収集による音声検索システムの評価

李　清宰[†1]　　河原　達也[†1]　　Rudnicky Alexander[†2]

Amazon Mechanical Turk (MTurk) を用いて効率的に音声データを収集する方法について述べる。音声検索の評価用セットのための音声データを収集するタスク (HIT)、及び収集されたデータの品質を検証するタスクを設計した。1000 以上の発話をきわめて効率的に収集することができた。そのうち 90%以上は正しい書き起こしがある有用なデータであり、妥当な音声認識精度が得られた。このデータを用いて、音声により書籍を検索するシステムの評価を行った。その結果、意味スロット毎に用意したベクトル空間モデルを組み合わせる提案手法が、従来の単純なベクトル空間モデルに比べて、高い検索性能を実現することを確認した。

Collecting Speech Data using Amazon's Mechanical Turk for Evaluating Voice Search System

Cheongjae Lee,[†1] Tatsuya Kawahara[†1]
and Alexander Rudnicky[†2]

This paper describes a crowd-sourcing method to collect speech data using Amazon's Mechanical Turk (MTurk). We designed a task (HIT) to collect speech data as an evaluation set for voice search and another task to verify the quality of the collected speech data. More than a thousand utterances are collected very efficiently. It turned out that more than 90% of them are valid with correct transcript, and reasonable recognition accuracy is achieved. Using the data, we conducted evaluation of the voice book search system, and confirmed that the combination of slot-based vector space models provides higher search accuracy than the conventional single vector space model.

1. Introduction

Recently, many spoken dialog systems provide users with various information in a large-scale database when they request with a spoken query 13). The system needs to first identify a likely set of candidates for the target item in a large database, then efficiently reduce this set to match that item or items originally targeted by the user. This part of the process is characterized as "voice search" and several such systems have been developed for various applications: automated directory assistance system 14), consumer rating system 15), multimedia search 11), and book search 7), 9). Among them, the automated directory assistance system is one of the most popular voice search systems in which the system returns the phone number and address information of a business or an individual by interpreting a spoken query input 14).

One of the major challenges for voice search systems is how to evaluate the search performance on large-scale evaluation sets produced by real humans. It is difficult to collect real spoken queries during the early stage of the development because we do not have a real system deployed in the real world. For instance, Lee et al. 7) preliminarily evaluated the system performance using the synthetic queries by incorporating simulated ASR errors to investigate the degradation through speech recognition and understanding errors. However, these queries might not include the characteristics of real spoken queries and real ASR errors because they were artificially generated through an ASR simulator.

In this paper, we investigate the use of Amazon's Mechanical Turk (MTurk) to collect spoken queries as an evaluation set for voice book search as the MTurk service provides the opportunity for efficient data collection.

The remainder of this paper is organized as follows: Section 2 introduces related work on the use of MTurk in natural and spoken language communities. Next, Section 3 presents how to collect our speech data as well as to evaluate the quality of the data using MTurk. Then, we present an evaluation of the voice search system using the speech data in Section 4. Finally, we conclude with a brief summary in Section 5.

[†1] Academic Center for Computing and Media Studies, Kyoto University, Japan.
Cheongjae Lee is a research fellow of the Japan Society for Promotion of Science (JSPS).
[†2] Computer Science Department, Carnegie Mellon University, USA

2. Amazon's Mechanical Turk

Amazon's Mechanical Turk (MTurk) [⋆1] is an on-line marketplace for human workers who perform various tasks called "human intelligence tasks" (HITs) published by others in exchange for an amount of money. In MTurk, workers are referred to as "turkers" and people designing the HITs are called "requesters." Requesters design and publish HITs which are simple for humans but difficult for computers (e.g. survey, annotation, evaluation, translation). Turkers can freely select HITs presented in the web pages and perform the described task. They are given a reward (roughly $0.01 to $0.1 USD per task) when they complete the task.

There are several recent studies on collecting and evaluating speech and language data with MTurk 4). First, Snow et al. 10) examined the effectiveness of non-expert labeling using MTurk for a variety of natural language processing tasks such as affect recognition, word similarity, and word sense disambiguation. In general, corpora have been manually annotated by expert annotators to minimize annotation errors. However, it is very costly to scale up the data. This motivated the use of MTurk. Although the annotation could be performed by non-experts, it is shown that the average of four non-expert labels per an item could emulate expert-level label quality.

Machine translation (MT) is one of the most popular tasks in MTurk because traditional MT systems rely on huge parallel corpora to achieve a sufficient performance. Some groups in the MT community have shown that MTurk can be useful for creating parallel corpora 1) and for evaluating MT systems manually 2). For example, Burch et al. 3) have investigated how to manually evaluate translation quality at a low cost. They found that combining non-expert judgements shows a high-level of agreement with the existing gold-standard judgements of MT quality.

More recently, MTurk has been used for spoken language processing tasks such as recording audio data 6) and transcribing speech data 8). For instance, Lane et al. 6) have developed tools to collect speech data for speech-to-speech translation systems on mobile devices or MTurk. To collect speech data, a prompt is presented to a remote user and the user will utter the speech by reading the sentence aloud. They investigated the quality of the recorded audio files by experts and concluded that a basic tutorial would be helpful to yield high quality recordings. Marge et al. 8) also explored whether MTurk can be used for transcribing spoken language data. They examined the transcription quality according to the payment amount, the turkers' gender and speaker native language. They could also improve the accuracy by combining multiple transcriptions of an utterance.

In this paper, we demonstrate that MTurk can be used to collect as well as to evaluate a number of spoken queries of voice search. MTurk has two advantages for collecting an evaluation set: 1) collecting spoken queries cheaply and remotely and 2) collecting queries produced by various speakers in real environments.

3. Speech Data

3.1 Data Collection via MTurk

Many previous studies mentioned in Section 2 have suggested requesters should describe the tasks clearly and design HITs carefully to yield reliable results. This section describes the design of our HITs for collecting speech data and presents the statistics of the collected data.

The overall structure of the HITs we designed for collecting speech corpora consists of instructions, profile questions, recording, and transcription. One of the most important issues when collecting speech data in MTurk is that most turkers are non-experts who might have no knowledge and experiences about what they are doing. To address this problem, we should clarify the goal of our task and how to perform the task successfully. For instance, to ensure that the recording condition, turkers were instructed to minimize background noise and to play back the recorded audio file before submitting the current HIT. To encourage careful work, we included this note in our HITs: "Your HITs will be evaluated manually by different turkers. If your recorded speech and transcript are of poor quality constantly, you will be blocked." Next, our HITs also contain a survey to gather some profile information such as age, gender, native language, and microphone type. Turkers should answer these questions before recording an utterance. We created a HIT that elicits utterances by providing metadata consisting of a book title, authors, and a category. Turkers were asked to formulate a response to the question "how can I help you?" posed by a hypothetical human bookstore clerk. A typical query might be "I AM LOOKING FOR ALICE IN WONDER-LAND BY CARROLL". After making up what they would speak, they clicked the button to show the recording interface. We used the web-based recording interface in Java applet

[⋆1] http://www.mturk.com/

provided by VIMAS [*1]. This enables us to record audio data, which are uploaded to the web server via HTTP. After the utterance, turkers were asked to transcribe what they had said just now. Finally, we should design the HITs in a way that deter cheating because they sometimes attempt to complete tasks as quickly as possible without commitment. To address this problem, the HIT was validated when the turkers answered all questions and uploaded the recorded file successfully before submitting HIT results. We also manually monitored for blocking the cheating turkers who might speak no words or transcribe the utterances incorrectly. In MTurk, requesters can specify who is eligible to accept the HITs. We restricted eligibility to turkers who are located in the USA because they should be fluent in speaking English. Moreover, only turkers with at least a 85% HIT approval rating[*2] were allowed access to this task.

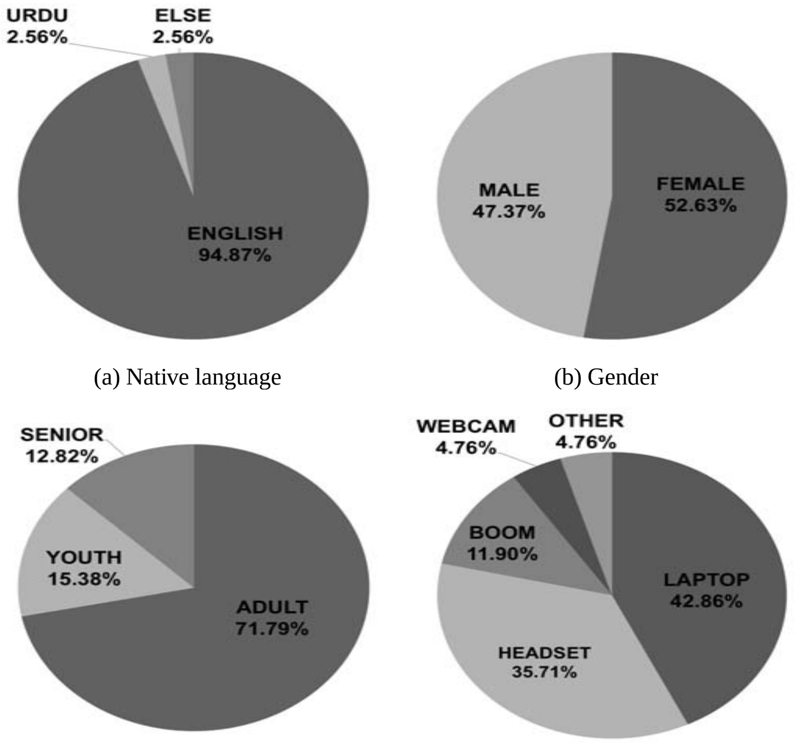
We have collected 1011 utterances from 38 different turkers using MTurk. Each HIT was paid $0.1 USD and we gave a bonus ($0.5 USD) to those who completed many HITs successfully. Totally, it costed approximately $120 USD to complete 1011 HITs including additional small Amazon surcharges and any taxes that apply.

We show the distributions of the speaker characteristics (native language, age, gender, microphone type) based on their answers (Fig. 1). First, most turkers (94.87%) were native English speakers. However, in our investigation on the accent of the speech, some turkers might have lied about their true native language. We noted in the task description, “We requires your English is fluent”. That is why they were afraid that the work would be rejected if they answered non-native.

3.2 Evaluation of the Recorded Audio

The main purpose of the collected data is an evaluation of the book search performance on real spoken queries. To that end, the data should be proofed before conducting evaluation experiments. Therefore, we employed another MTurk to evaluate the quality of the collected speech data by combining multiple non-expert judgements per a spoken query.

We published HITs for evaluating the recording quality and the transcription quality. Turkers listened to the recording of each utterance by using an audio player embedded in the task web page and they were asked to answer the questions about the quality of the recorded audio and the transcript listed in Table 1. They checked one of the values shown in the question.



(a) Native language (b) Gender

(c) Age (d) Microphone type

図 1 Distributions of speaker characteristics.

To control the quality of the HIT results, we restricted eligibility to turkers who are located in the USA and achieved at least 85% approval rating. Each HIT is assigned to three different turkers and paid $0.01 USD. The judgement is validated if two of the three turkers agreed (majority vote); otherwise, it is marked as an invalid data (R0, T0).

Table 2 shows the evaluation results of the recording and transcription quality by non-experts and an expert. As a result, 8.80% of the recorded audio files were determined that they have low quality, and 7.71% of the transcripts were not correct, and thus they were corrected manually. We also verify the non-expert turker judgements by measuring their inter-annotator agreement with

[*1] http://www.vimas.com/
[*2] This means the fraction of HITs approved by requesters in the past. Qualifying the approval rating is one of the best quality control strategies.

表 1 Values used to label recording quality and transcription quality. Minor errors in transcripts include spelling errors and delimiting errors.

Recording quality	R1	Recorded audio is empty
	R2	Recording is clipped
	R3	Recording contains audible echo or noise
	R4	I can't hear it because the volume is too low
	R5	Recording is clear
Transcription quality	T1	There are no words
	T2	It is entirely different from the recorded file
	T3	There are some minor errors in transcript
	T4	The recorded audio is perfectly transcribed

表 2 Evaluation of the speech data via MTurk. R0 and T0 represent invalid data.

Quality	Turkers (%)	Expert (%)
R0	0.6924	0.0000
R1	0.2967	0.0000
R2	0.0000	0.2967
R3	7.6162	12.6607
R4	0.1978	0.2967
R5	91.1968	86.7458

(a) Recording quality

Quality	Turkers (%)	Expert (%)
T0	0.1978	0.0000
T1	0.2967	0.0000
T2	0.0000	0.2967
T3	5.5391	7.4184
T4	93.9664	92.2849

(b) Transcript quality

表 3 Statistics of the collected queries.

Type	#Queries	Avg. Words	Avg. Slots
Typed (training)	1574	14.00	2.01
Spoken (test)	1011	15.39	2.22

the expert judgements. This yields a κ value of 0.9185 in the recording quality and 0.9377 in the transcription quality, which can be considered as good agreement.

4. Experiments and Results

4.1 Experiment Setup

PocketSphinx has been used as a speech recognizer configured with a general acoustic model (AM) and a statistical n-gram language model (LM). The AM used was the freely available US English broadcast news acoustic model. We did not adapt the models to our task.

We used all the collected 1011 spoken queries described in the previous sectionfor evaluation. In addition to the spoken queries, we had collected about 1500 typed queries via MTurk 7), which were used as a training set in this work. Table 3 shows the statistics of the collected queries. Unfortunately, they are not sufficient to provide a good coverage for language model training. Therefore, we built a trigram LM by generating sampled queries. All typed queries were manually annotated with semantic slots (e.g. book title, author's name, book category), and then 10K synthetic sentences were automatically generated from the annotated queries and the metadata of book items. Original slot values in the query were replaced with another metadata randomly selected in the book database containing 15088 book items.

表 4 ASR evaluation on the test set.

Voca. Size	WER (%)	PP	OOV (%)
10483	35.07	92.09	1.67

表 5 WER and PP of partial workers who submitted more than 30 queries.

Worker Id	#Queries	WER (%)	PP
W1	123	19.73	42.76
W2	354	23.84	154.01
W3	40	25.85	140.31
W4	82	29.33	134.50
W5	103	30.95	20.64
W6	39	74.60	20.73
W7	78	91.49	138.96

The resulting hypotheses were processed to extract semantic slots such as book title, author's name, book category. These slots are necessary to search for relevant books precisely in our search algorithm. Our SLU module relied on the conditional random field (CRF)-based information extraction trained with the 1574 spoken queries 5).

4.2 Speech Recognition Performance

We first measured word error rate (WER), perplexity (PP), and out-of-vocabulary rate (OOV) for the collected speech data (Table 4). The WER is 35.07%. To analyze ASR errors, we also examined ASR performance according to individual workers. Table 5 shows that WER and PP for individual workers who submitted more than 30 queries. This indicates that the ASR accuracy of individual workers varies. Some (W1-W5) are relatively accurate, while others (W6-W7) are not appropriate for ASR systems regardless of PP. We should investigate which factors such as noises and accents are related with ASR errors.

表 6 Search evaluation on manual transcripts and 1-best ASR outputs.

	WER (%)	CER (%)	SVSM		HVSM	
			$P@100$	MRR	$P@100$	MRR
manual	0.00	10.20	0.9772	0.8505	0.9831	0.8783
ASR	35.07	39.78	0.7633	0.5709	0.7861	0.6005

4.3 Book Search Performance

The system for voice book search has been developed 7). It relies on the vector space model. In the conventional single vector space model (SVSM), all terms in different slots are indexed together over a single vector space where every term is equally weighted regardless of its slot name. SVSM accepts a bag-of-words vector without relying on SLU modules. However, we can search for relevant items more precisely through SLU and slot-specific subspace models because it considers the relationship between slots. Therefore, we have proposed the hybrid vector space model (HVSM) by linearly interpolating SVSM and slot-based vector space models in which each slot is independently indexed over subspaces.

For evaluation, we use two evaluation metrics widely used in information retrieval. One is precision at n ($P@n$), which represents the number of queries having the correct answer in the top n relevant items divided by the total number of queries. The other is mean reciprocal rank (MRR), which indicates the average of the reciprocal ranks of search results for a set of queries 12). In reality there may be multiple correct answers in a list, when users do not have the exact book in their mind. For example, some users can search for any fictions without an exact book in their mind. Because it is difficult to automatically determine the relevance relationship between the queries and the lists, we identified a single correct book corresponding to each query. Note that some of queries (e.g. I AM LOOKING FOR A BOOK IN FICTION CATEGORY) could not get a correct answer.

We conducted experiments on the spoken queries. We extracted one-best ASR outputs from spoken queries. Table 6 summarizes the search performance together with WER and CER (Concept Error Rate). While the WER was 35%, the search accuracy was degraded only by 20%. This means that the system is robust to ASR errors. We achieve improvement of approximately 2% absolute in precision by using HVSM compared to SVSM.

5. Conclusions

We have explored an efficient data collection method using MTurk. We also used MTurk to measure the quality of the collected data. The quality of recording and transcription was determined by taking a majority voting with multiple non-expert judgements. We found high agreement between non-experts and experts. Finally, we evaluated two different search models on the spoken queries to examine the effectiveness of the proposed search algorithm. We could confirm that HVSM shows higher search accuracy than SVSM.

Some issues have yet to be resolved in the future work. One of them is that the search robustness to ASR errors should be improved by incorporating n-best results and confidence scores into the search algorithm. In addition, SLU should be enhanced by comparing the values in the top-n items. We should also investigate how to collect real dialog data via MTurk, which is more challenging than collecting a one-shot utterance.

参考文献

1) Ambati, V. and Vogel, S.: Can crowds build parallel corpora for machine translation systems?, *Proc. NAACL/HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.62–65 (2010).

2) Bloodgood, M. and Callison-Burch, C.: Using Mechanical Turk to build machine translation evaluation set, *Proc. NAACL/HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.208–211 (2010).

3) Callison-Burch, C.: Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk, *EMNLP*, pp.286–295 (2009).

4) Callison-Burch, C. and Dredze, M.: Creating Speech and Language Data With Amazon's Mechanical Turk, *Proc. NAACL/HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.1–12 (2010).

5) Jeong, M. and Lee, G.G.: Tri-angular chain conditional random fields, *IEEE Transactions on Audio, Speech and Language Processing*, Vol.16, No.7, pp.1287–1302 (2008).

6) Lane, I., Waibel, A., Eck, M. and Rottmann, K.: Tools for Collecting Speech Corpora via Mechanical Turk, *Proc. NAACL/HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.184–187 (2010).

7) Lee, C., Rudnicky, A. and Lee, G.: Let's Buy Books: Finding eBooks using Voice Search, *Proc. IEEE SLT Workshop*, pp.442–447 (2010).

8) Marge, M., Banergee, S. and Rudnicky, A.I.: Using the Amazon Mechanical Turk for transcription of spoken language, *Proc. ICASSP*, pp.5270–5273 (2010).

9) Passonneau, R.J., Epstein, S.L., Ligorio, T., Gordon, J.B. and Bhutada, P.: Learning about Voice Search for Spoken Dialogue Systems, *Proc. NAACL*, pp.840–848 (2010).

10) Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y.: Cheap and fast - but it is good? Evaluating non-expert annotations for natural language tasks, *Proc. EMNLP*, pp.254–263 (2008).

11) Song, Y.-I., Wang, Y.-Y., Ju, Y.-C., Seltzer, M., Tashev, I. and Acero, A.: Voice Search of Structured Media Data, *Proc. IEEE ICASSP*, pp.3941–3944 (2009).

12) Voorhees, E.M. and Tice, D.M.: The TREC-8 question answering track evaluation, *Proc. Text Retrieval Conference TREC-8*, pp.83–105 (1999).

13) Wang, Y.-Y., D.Yu, Ju, Y.-C. and Acero, A.: An Introduction to Voice Search, *IEEE Signal Processing Magazine*, Vol.25, No.3, pp.29–38 (2008).

14) Yu, D., Ju, Y.-C., Wang, Y.-Y., Zweig, G. and Acero, A.: Automated Directory Assistance System, *Proc. INTERSPEECH*, pp.2709–2712 (2007).

15) Zweig, G., Nguyen, P., Ju, Y.-C., Wang, Y.-Y., Yu, D. and Acero, A.: The Voice-Rate Dialog System for Consumer Ratings, *Proc. INTERSPEECH*, pp.2713–2716 (2007).