Evaluation of Crowdsourced User Input Data for Spoken Dialog Systems

Maria Schmidt, Markus Müller, Martin Wagner, Sebastian Stüker and Alex Waibel
Karlsruhe Institute of Technology
Karlsruhe, Germany
maria.schmidt@kit.edu

Hansjörg Hofmann and Steffen Werner
Daimler AG
Sindelfingen, Germany
hansjoerg.hofmann@daimler.com
steffen.s.werner@daimler.com

Abstract

Using the Internet for the collection of data is quite common these days. This process is called crowdsourcing and enables the collection of large amounts of data at reasonable costs. While being an inexpensive method, this data typically is of lower quality. Filtering data sets is therefore required. The occurring errors can be classified into different groups. There are technical issues and human errors. For speech recording, technical issues could be a noisy background. Human errors arise when the task is misunderstood. We employ several techniques for recognizing errors and eliminating faulty data sets in user input data for a Spoken Dialog System (SDS). Furthermore, we compare three different kinds of questionnaires (QNRs) for a given set of seven tasks. We analyze the characteristics of the resulting data sets and give a recommendation which type of QNR might be the most suitable one for a given purpose.

1 Introduction

Similar to research in other areas, Automatic Speech Recognition (ASR) systems and SDSs are facing the challenge how to get new training data, e. g., if there is the urge to cover new domains. Until several years ago, a common procedure was to record the required audio samples in an anechoic chamber and let experts (e. g., linguistics students) create the transcriptions. Although the data collected via this method is of high quality and can be used as a gold standard, researchers found that this approach is very time-consuming and results in quite little data related to the effort.

A few years ago, companies like Amazon Mechanical Turk started to offer so-called crowdsourcing approaches, which meant that Human Intelligence Tasks (HITs) are performed by a group of non-experts. Furthermore, these tasks are open calls and are assigned to the different crowdsource workers. Especially in industrial contexts, crowdsourcing seems to be the means to choose because development cycles are short and much data for ASR or SDS development can be generated right as it is needed, although the data collected needs to be checked for quality (Snow et al., 2008).

Our work analyzes crowdsourced data collected by the company Clickworker (Eskenazi et al., 2013, ch. 9.3.4). The collected data consists of user input to an in-car SDS, where the crowdworkers had to input one German utterance for each of the seven tasks, after which they had to transcribe the utterance themselves. This procedure was conducted for three different types of QNRs: pictures, semantics, and text. We show the differences among these QNRs as well as an overall quality evaluation of the collected data. For this purpose, we make use of Natural Language Processing (NLP) tools for tackling the preprocessing of the data.

This paper is structured as follows: In section 2, we look at other work in this area. Following that, we explain our approach in section 3 and have a closer look at the underlying data in section 3.1. We continue with the presentation and an analysis of the results (section 4). We finish with a conclusion in section 5.

2 Related Work

In this section, we describe the research background of the collection of speech data via crowdsourcing as well as using automatic methods for quality improvement.

2.1 Collection of Speech Data via Crowdsourcing

Crowdsourcing is a common part for collecting speech data nowadays. Eskénazi defines it as "a crowd as a group of non-experts who have answered an open call to perform a given task" in (Eskenazi et al., 2013). Such a call will be advertised using special platforms on the Internet. Even though the participants are called "nonexperts", they are skilled enough in general to perform these tasks. For collecting speech data, recording audio from a variety of different speakers helps in building better systems. Different speakers have a different background. This in reflected in their speaking style and choice of words (Hofmann et al., 2012). These aspects are key for training a speaker independent system. The choice of participants should reflect the target audience of the system. Using untrained workers also has the advance of being cheaper than to hire experts.

2.2 Using ASR to Improve the Quality

The use of an ASR system is nowadays an integral part of the collection of annotated speech data. Such sys-

tems are being used of optimize the collection methods. (Williams et al., 2011) have shown how to process HITs for difficult speech data efficiently. One approach to do this is to first create a transcription and let it then correct by crowdworkers. Since humans are optimistic about correcting errors (Audhkhasi et al., 2012), a two step approach was proposed in (Parent and Eskenazi, 2010): Let the workers first rate the quality / correctness of transcriptions and perform the corrections in a separate step.

Another approach (van Dalen et al., 2015) deals with the combination of automatic and manual transcriptions. Errors produced by these two methods are to some degree orthogonal: While humans tend to introduce spelling errors or skip words, automatic transcriptions feature wrong words, additional words and also missing words. The usual approach for combining multiple transcriptions is ROVER (Fiscus, 1997). It requires an odd (typically three) amount of different transcriptions to be merged. Otherwise it is difficult to break the tie. By the use of an ASR system, van Dalen et. al have shown that two manual transcriptions are sufficient to produce high quality transcriptions.

3 Analysis of Crowdsourced User Input Data for Spoken Dialog Systems

In this section, we describe our approach to analyze the given corpus containing crowdsourced user input data for a goal-oriented spoken dialog system deployed in a car. Section 3.1 summarizes the characteristics of the corpus, while Section 3.2 and 3.3 describe the actual analysis steps for audio and transcript data, respectively.

3.1 The Corpus

The underlying German utterances for our analysis were collected by the German company Clickworker (similar to Amazon Mechanical Turk; see http://www.clickworker.com/en). The participants were asked to invoke seven specific actions of an imaginary SDS deployed in a car. First, they got a task description, then they should record an audio of their input via a browser-based application on their own PC incl. microphone at home. After that the subjects were asked to transcribe their own utterance without hearing or seeing it again. In the following we describe the tasks 1, 4 and 5 exemplarily: In task 1, the imaginary user tells the system that he/she wants to listen to a certain radio station. Task 4 comprises the navigation to the address "Stieglitzweg 23, Berlin". In task 5, the user should call Barack Obama on his cell phone.

There were three different QNRs, where each of them asked for all seven tasks named above. The QNRs differed in the way how the tasks were presented to the subject: by means of pictures, text, or semantics (see Figure 1). In the QNR of type pictures, the participants were shown one or more pictures depicting the task they should perform. Without any written text, this type of task description does not imply the use of specific terms.

As for the type text, the participants were presented a few lines describing the situation they are in and the actions they should perform. This form of textual representation of the objects is more influencing towards the use of specific terms. In the third form of representation (semantics), the participants are influenced the most, as they get presented a few keywords. This does not favor the use of different words.

Each participant answered all seven tasks, but was presented only one type of task description across them. Each type of QNR was assigned to approximately 1,080 users resulting in 22,680 utterances (34.7 hours) in total, i. e., roughly 7,560 per QNR. Most subjects were between 18 and 35 years old, a smaller number of subjects was up to 55 years old. 90% of the subjects were between 18 and 35 years old, 8% between 36 and 55. The smallest group was aged over 55 which resulted in 2% of the data. Our participants were 60% men and 40% women.

3.2 Evaluation of Self-Entered Transcripts

In order to be able to tell the overall quality of the underlying corpus, we had to analyze the self-entered transcripts, too. For this purpose, we developed an NLP analysis chain which contains a large part of preprocessing (i. e. mainly cleaning the text) apart from the actual analysis part.

Concerning the preprocessing, we first applied a basic tokenizer to split the punctuation marks from the rest of the text. Second, we went over the transcripts with a spell checker called LanguageTool (https://www.languagetool.org/). For all apparently misspelled words, we checked whether it equals one of the predefined, special keywords which should be entered for the current task (e. g., "Michael Jackson", "Stieglitzweg"). If such a keyword was found, we processed the next word; if not, we checked which of the correct alternatives proposed by LanguageTool is most similar to one of the words on our "synonymously used words" list by using the Levenshtein distance. Third, after deciding which spelling is the most appropriate one for each word in an utterance, we store the corrected utterances and use them for further analysis.

The analysis as such included Part-of-Speech (POS) Tagging with the TreeTagger (Schmid, 1994) to investigate, which and how many different POS patterns, i.e. types of sentence patterns, occur in the corpus and how the QNRs differ from each other on this level. Further, we investigated the most frequent words used in each task, and how many words in total are used in a specific task and in a specific QNR.

With our analysis, we provide answers to the following questions: (a) How large is the linguistic variation in the data set (on sentence and word level)? (b) Which pros and cons do the presented QNRs have? (c) Which QNR is the right one for a certain purpose? The results

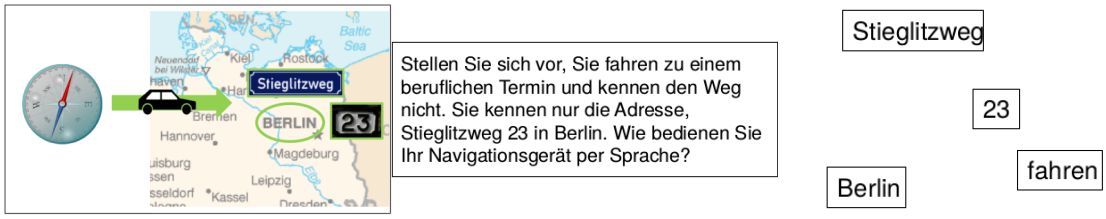

Figure 1: Instructions for task 4 in form of pictures, text and semantic entities

are presented in Section 4.2.

3.3 Evaluation of Self-Recorded Audio Data

We use different measures to determine the usability of the recordings. We compared the length of the recordings and we analyzed them by means of an ASR system. In general, we assume that most of the recordings are done appropriately and that the quality of the recordings resembles a normal distribution. We conducted our analysis using the Janus Recognition Toolkit (JRTk) (Woszczyna et al., 1994) which features the IBIS decoder (Soltau et al., 2001).

For every task, an answer with a certain length is expected. This length may vary to some degree, but an audio file that is significantly shorter or longer is an indication of an error. This may be a technical issue with the recording setup or a misunderstanding of the task description. In both cases, the recording needs to be discarded.

Even if the length is within a suitable range, the transcription of the audio might be wrong. To estimate if the transcription matches the spoken words, we use JRTk to perform a forced alignment the audio to the transcriptions. We use a GMM / HMM based recognizer for German with 6,000 context-dependent quintphone states. For the alignment, a phoneme sequence based on the word sequence is generated. This phoneme sequence is then aligned to the audio using forced Viterbi alignment. If there is a mismatch between audio and transcriptions, there will be phonemes covering unusual long or short parts of the audio.

4 Results & Discussion

4.1 Results of the Audio Data Analysis

We divided the recordings into 21 different sets as there are 3 different QNRs and 7 tasks each. Table 1 shows a detailed overview of the recording lengths for different tasks. It can be seen that while task 4 produces the longest recordings, the QNR "semantics" produces the shortest recordings.

We also performed a forced Viterbi alignment. Having a look at the length of the longest phoneme in an utterance is an indicator whether the recording and the transcription fit together. Figure 2 shows a histogram of the length of the longest phoneme per utterance. Since we do not have multiple transcriptions per utterance,

Task	Pictures	Text	Semantics
1	5.21s	5.04s	5.04s
2	5.75s	6.01s	5.65s
3	4.97s	5.01s	4.56s
4	6.80s	6.79s	6.44s
5	5.45s	5.39s	5.26s
6	5.46s	5.78s	5.51s
7	5.37s	5.21s	4.73s

Table 1: *Average length of recordings.*

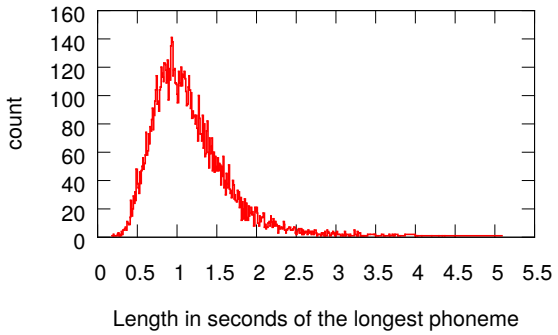

Figure 2: *Longest phoneme length per utterance.*

we could not determine an optimal parameter set for identifying mismatched cases. But our preliminary results indicate that the longer the longest phoneme, the more likely a mismatch.

4.2 Results of the Transcript Analysis

Aiming at answering the questions posed in Section 3.2, we show the results of the transcript analysis in the following paragraphs together with a short discussion. Tables 2, 3, and 4 show the total number of utterances in the whole data set or the different QNRs, respectively. The second line of each table displays how many of the obligatory semantic entities were named, i.e. whether the two main content words (nouns in many tasks) were named, like "Sender, SWR3". The third line displays the number of insufficient utterances according to this criterion. Similarly, line four and five tell how many of the entities, which were actually asked for, i.e. all three (or four) items, were named and how many utterances were dismissed accordingly. One can see that the pictures QNR has to dismiss the most entities, while the semantics QNR dismisses the least.

The values of the text QNR are in between the latter two QNRs. In total we have a dismissal rate of 17% and 37%, respectively.

Table 5 displays the variance of words used for all three QNRs and across tasks 1-7. It is valid for all tasks that the semantics QNR has the lowest number of different words. This is probably caused by displaying three exact semantic items, which is inevitably the corresponding word. For tasks 1, 2, 3 and 7, the text QNR has the highest number of different words, while the pictures QNR leads the number of different words in tasks 4-6.

Table 6 shows the most frequent POS sequences per QNR. One can see that in the semantics QNR, most people used a polite modal construction "Ich möchte den Sender SWR hören" while in the other QNRs "Radio SWR3" is the most common one among finite and infinite constructions.

Table 7 displays the most common sentence for each task. As you can see, there is a wide variety of linguistic patterns in each task.

total number of utterances	22,701	
number of obligatory entities	18,935	83%
number of insufficient utterances	3,766	17%
number of asked for entities	14,371	63%
number of insufficient utterances	8,330	37%

Table 2: *Overall data dismissal across all QNR types.*

total number of utterances	7,546	
number of obligatory entities	5,420	72%
number of insufficient utterances	2,126	28%
number of asked for entities	3,033	40%
number of insufficient utterances	4,513	60%

Table 3: *Picture QNR data with its dismissal rate.*

total number of utterances	7,581	
number of obligatory entities	6,947	92%
number of insufficient utterances	634	8%
number of asked for entities	6,126	81%
number of insufficient utterances	1,455	19%

Table 4: *Semantics QNR data with its dismissal rate.*

5 Conclusion

We have presented various methods for evaluating the collected data set. We have shown that different types of QNRs lead to different styles in performing the tasks. With respect to the actual application scenario, the way in presenting the task to the participants has to be chosen in the correct manner.

The semantics QNR is precise by using three semantic items and is the best choice for generating exact phrases. Furthermore, it generates very few utterance dismissals. But at the same time it displays the words themselves. To avoid the mere usage of these words, one approach for future studies would be to display the semantic items in English. Simultaneously, this would have the advantage that the QNR would be reusable also for the generation of data from other languages.

The pictures QNR is optimal to generate a very high linguistic variance in the data. The downside of this approach is the high dismissal rate, if one aims at generating specific utterances.

The text QNR is a good compromise between the two other QNR types characterized above. According to the data analyzed in this word, the text QNR has a lower priming effect on formulations than the semantics QNR.

Future work includes the evaluation of politeness of the prompts and a more profound recommendation about the required size of data collection to provide maximal coverage of linguistic constructions. From an economical point of view, the prediction of the appropriate size for the data set is important to reduce costs. This includes the estimation of the amount of unusable data.

# of words used	pictures	semantics	text	avg.
Task 1	199	176	237	204
Task 2	216	206	256	226
Task 3	279	225	326	277
Task 4	327	260	309	299
Task 5	266	179	253	233
Task 6	297	188	264	250
Task 7	340	229	377	315
Average	275	209	289	258

Table 5: *Variance of used words across all QNR types.*

QNR	POS sequence	%
pictures	NN NN	12.5%
	NN NN NN	6.0%
	NN NN VVINF	6.0%
semantics	PPER VMFIN ART NN NN VVINF	24.8%
	NN NN VVINF	16.5%
	NN VVINF NN	3.7%
text	NN NN	11.0%
	NN NN VVINF	9.0%
	NN NN NN	5.4%

Table 6: *Three most common POS sequences for each QNR – exemplary: task 1 (valid data sets).*

Task	Most frequent sentences
1	Ich möchte den Sender SWR3 hören
4	Navigiere [mich] zu[m] Stieglitzweg 23 in Berlin
5	Barack Obama [auf [dem]] Handy anrufen

Table 7: *Most common sentences for task 1, 4 and 5.*

References

Kartik Audhkhasi, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2012. Analyzing quality of crowd-sourced speech transcriptions of noisy audio for acoustic model adaptation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4137–4140. IEEE.

Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann. 2013. *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*. John Wiley & Sons.

Jonathan G Fiscus. 1997. 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*, pages 347–354. IEEE.

Hansjörg Hofmann, Ute Ehrlich, André Berton, and Wolfgang Minker. 2012. Speech interaction with the internet–a user study. In *Intelligent Environments (IE), 2012 8th International Conference on*, pages 323–326. IEEE.

Gabriel Parent and Maxine Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. 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*, pages 254–263. Association for Computational Linguistics.

Hagen Soltau, Florian Metze, Christian Fugen, and Alex Waibel. 2001. A one-pass decoder based on polymorphic linguistic context assignment. In *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 214–217. IEEE.

R van Dalen, K Knill, P Tsiakoulis, and M Gales. 2015. Improving multiple-crowd-sourced transcriptions using a speech recogniser. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Institute of Electrical and Electronics Engineers*.

Jason D Williams, I Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. 2011. Crowdsourcing for difficult transcription of speech. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 535–540. IEEE.

Monika Woszczyna, N. Aoki-Waibel, Finn Dag Buø, Noah Coccaro, Keiko Horiguchi, Thomas Kemp, Alon Lavie, Arthur McNair, Thomas Polzin, Ivica Rogina, Carolyn Rose, Tanja Schultz, Bernhard Suhm, M. Tomita, and Alex Waibel. 1994. Janus 93: Towards spontaneous speech translation. In *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia.