Towards an Automatic Assessment of Crowdsourced Data for NLU

Patricia Braunger, Wolfgang Maier, Jan Wessling, Maria Schmidt
Daimler AG
Sindelfingen, Germany
{patricia.braunger, wolfgang.mw.maier, jan.wessling, maria.m.schmidt}@daimler.com

Abstract

Recent development of spoken dialog systems has moved away from a command-style input and aims at allowing a natural input style. Obtaining suitable data for training and testing such systems is a significant challenge. We investigate with which methods data elicited via crowdsourcing can be assessed with respect to its naturalness and usefulness. Since the criteria with which to assess usefulness depend on the application purpose of crowdsourced data we investigate various facets such as noisy data, naturalness and building natural language understanding (NLU) models. Our results show that valid data can be automatically identified with the help of a word based language model. A comparison of crowdsourced data and system usage data on lexical, syntactic and pragmatic level reveals detailed information on the differences between both data sets. However, we show that using crowdsourced data for training NLU services achieves similar results as system usage data.

Keywords: crowdsourcing, spoken dialog system, natural language understanding, training data

1. Introduction

A major goal of current spoken dialog systems (SDS) development is to obtain a more natural and human-like style of communication. Such a communication style presupposes an understanding of all utterances which are associated with a specific semantic meaning. Thanks to the advent of statistical data-driven approaches, recent systems have been able to interpret freely spoken user input. However, obtaining suitable data for their training and evaluation is a significant challenge.

In order to increase the amount of data for training and evaluation of natural language understanding (NLU), researchers are turning to crowdsourcing instead of collecting data from Wizard of Oz experiments or using handcrafted grammars (see, e.g., Callison-Burch and Dredze (2010)). While the benefits of crowdsourcing can include expedition, individuality of people, and low costs (see, e.g., Eskenazi et al. (2013)), it is often criticized for poor standards as it is difficult to control the quality of work when requesting complex tasks (Eskenazi et al., 2013). Since the community of crowd workers is acknowledged for being Internet savvy young adults, chiefly between ages 18 and 35, the findings are not representative for any target group. In addition, crowdsourced data intended to be used for improving SDS do not necessarily reflect real system usage. Therefore we investigate to which extent crowdsourced data are suitable for the development of human-like SDS. Specifically, we want to provide answers to the following questions:

- Do data elicited via crowdsourcing reflect real system usage and naturalness?
- Are data elicited via crowdsourcing suitable to encourage the development of high-quality NLU modules?

Thereby, we aim to propose methods with which to assess crowdsourced data for NLU. In order to establish a baseline of what requirements the utterances collected via crowdsourcing must fulfill, we conduct a study in which free user utterances are collected for an actual in-car SDS. By means of a comparative analysis with crowdsourced data, we first demonstrate to which extent real system usage utterances and crowdsourced utterances share properties in terms of naturalness and variety. Additionally, we address application-related questions that arise with making use of crowdsourcing for training and assessing NLU modules. We examine the impact of noisy data and we aim to answer the question how noisy data can be identified automatically.

The remainder of the paper is structured as follows: In section 2, we review previous literature which aims to evaluate crowdsourcing methods for different application purposes. Next, in section 3, we introduce our crowdsourced and system usage data and the design of our study. Section 4 presents and discusses the experiments and results. Lastly, section 5 serves as the coda of the article.

2. Related Work

In the past years, the speech processing communities have recognized that crowdsourcing is a promising solution to their strong need for data (Eskenazi et al., 2013). In the field of spoken dialog systems, crowdsourcing has been used for evaluation (Yang et al., 2010; Komarov et al., 2013) and acquiring training data for system components such as automatic speech recognition (ASR) (Rothwell et al., 2015; McGraw et al., 2010), natural language understanding (NLU) (Braunger et al., 2016; Wang et al., 2012; Misu, 2014), dialog management (Manuvinakurike et al., 2015; McGraw et al., 2010) and natural language generation (NLG) (Novikova et al., 2016; Mitchell et al., 2014).

The evaluation of the crowdsourced data sets differs depending on research questions and application purposes. The literature that evaluates crowdsourcing approaches for acquiring training and test data can described as follows.

Most of the research works compare different meaning representation modalities with which to elicit data for given meanings or tasks. Depending on the application of the acquired data, e.g., NLG or NLU, different metrics are used in order to find the best elicitation method among the proposed. Novikova et al. (2016), e.g., compare two meaning

representation modalities, namely text and pictures, with which to elicit NLG training data. The measures they use to assess the effect of the meaning representation method include time taken to collect data, average length of utterances, average number of sentences per utterance, semantic similarity, informativeness and naturalness. Naturalness is measured by human evaluation by asking whether the utterance could have been produced by a native speaker.

Braunger et al. (2016) evaluate whether pictures, semantic entities or textual descriptions are suitable for collecting natural language input for a given semantic form in an in-car SDS context. They compare the collected corpora in terms of semantic correctness and linguistic variance. They conclude that a text-based method is suitable for a myriad kinds of tasks, produces a high linguistic variance, and yields a high rate of usable data.

Yang Wang et al. (2012) compare three different crowdsourcing elicitation methods for the collection of utterances which correspond to a given semantic form, namely sentence-based, scenario-based and list-based method. They analyze the acquired data in terms of semantic correctness and the biases that the methods create. Their hypothesis is that if the method creates a bias the crowd workers follow the same slot order as presented. The authors found that where a natural ordering exists, it is captured. However, since slot ordering is only one of the criteria which can be affected by crowdsourcing methods, analyzing slot ordering is not enough to evaluate the utility of crowdsourced data in terms of naturalness. Other criteria that can be affected include wording, sentence types, politeness etc.

While the works described thus far use different criteria for what is natural, other authors evaluate their methods with help of real system-interaction data. Manuvinakurike and DeVault (2015), e.g., conduct their experiment each in the lab and online. They compare their crowdsourced dialog data set to a smaller lab-based data set. However, in terms of naturalness they do not compare the collected dialogs but rather subjective questionnaire ratings. As an example, a question related to naturalness was "I talked to my partner in the way I normally talk to another person".

Misu (2014) investigate crowdsourced ASR and NLU data for situated dialog systems. They collect information seeking queries that contain points-of-interests (POI) in the participants' surroundings such as "What is that blue building on the corner?". They compare the crowdsourced data to utterances generated by a handcrafted grammar in terms of similarity to data collected from users interacting with a situated dialog system in a moving car. With the help of BLEU score (Papineni et al., 2012) they show that crowdsourced data is closer to queries produced in real driving situations than manually created utterances. In addition, their evaluation with test set perplexity indicates that crowdsourced data improves the performance of language models compared to manually created utterances. Nevertheless, it is difficult to decide whether the scores they report are satisfactory for productive use.

For the development of human-like SDS, it is important that the data used for training and testing reflect real system usage. Within most crowdsourcing elicitation methods crowd workers have to fulfill imaginary tasks. Due to a lack of imagination and indirectness data collected via crowdsourcing might produce other kinds of utterances than data collected from real system usage. In order to approve the utility of crowdsourced data real system-interaction should be therefore utilized as a baseline. We aim to examine to which extent crowdsourced queries share linguistic properties with queries collected from real system usage and how well crowdsourced data perform compared to naturally spoken utterances. In addition, application-related questions such as the impact of noisy data have not been addressed thus far.

3. Data Collection Setup

One of our goals is to evaluate the usefulness of crowdsourced NLU data. Therefore, we evaluate our German data set acquired from crowdsourcing against data collected by an experimental study, in order to gain a baseline of what constitutes naturalness in user input. In the following, we explain the experimental setup and procedure of our investigation.

3.1. Crowdsourced Data

As crowdsourced data, we use a subset of the data we collected in a previous study (cf. Schmidt et al. (2015) and Braunger et al. (2016)). In a previous work, Braunger et al. (2016) compares three data elicitation methods which differ in how the tasks are presented to the participants, namely by pictures, semantic entities and text. Considering that a text-based method was found to suite best[1], we apply this method within the development process of an actual in-car SDS. For this reason, our experiments are based on utterances which are elicited via text descriptions following Braunger et al. (2016).

Our crowdsourced data is collected using the German crowdsourcing platform Clickworker[2]. First, crowd workers are asked to spontaneously formulate and record a voice command directed at solving an assigned problem, such as entering an address in the navigation application. Second, the crowd workers are asked to transcribe their utterances into a textual form[3]. The transcription has to be an exact match of their spoken utterances.

> "Imagine that you are travelling by car.
> For a little change you would like to switch to
> the radio station SWR3. What would you say?"

Figure 1: Textual Task Description (Task 1).

[1] The task presentation methods were compared in terms of linguistic variety, number of valid utterances and priming effects. Overall, the authors recommend making use of the text presentation method since it is a good compromise between a high rate of valid utterances, linguistic variety and the possibility of creating very specific tasks.

[2] http://www.clickworker.com/

[3] The analysis of the recordings themselves is beyond the scope of this paper. The quality of the audio data has been examined by Schmidt et al. (2015).

The tasks are presented to the crowd workers by means of textual descriptions of the situation in which they are in, as well as the actions they should perform. An example is given in Fig. 1.
With the help of the textual descriptions we requested speech input for seven tasks typically performed in a car:

1. Listen to radio station SWR3
2. Play Michael Jackson Greatest Hits
3. Next Shell gas station
4. Navigate to Stieglitzweg 23 in Berlin
5. Call Barack Obama on mobile phone
6. Set temperature to 23 degrees
7. Send text message to brother

For each of the seven tasks we collect 1,080 spoken and transcribed utterances. The 1,080 crowd workers are German native speakers. 40% of the crowd workers are female and 60% are male. 90% of the crowd workers are between the ages of 18 and 35.8% are between the ages of 36 and 55, and 2% are above the age of 55.

3.2. System Usage Data

As we want to examine to which extent utterances used for system interaction in a realistic situation share properties with artificial crowdsourced data, we use data from a previous study in which users had to freely speak to an actual in-car SDS within a Wizard of Oz (WOZ) experiment (Braunger et al., 2017). These WOZ data serve as gold standard data in the experiments below.
Among the tasks the participants have to solve are the same aforementioned seven tasks. In order to avoid bias, i.e. offering the participants verbal examples, the tasks are presented by pictures as opposed to textual descriptions. These pictures are first pre-tested with friendly users to evaluate if the desired situation was put in the user's mind. An example of the task description is given in Fig. 2.

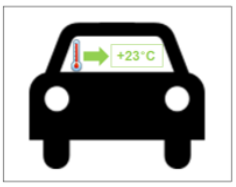

Figure 2: Graphical Task Description (Task 6).

The system behavior is simulated with the SUEDE tool (Klemmer et al., 2000) and designed such as available in a current Mercedes-Benz E-class. The participants are told that the system is able to understand any utterance in the given context. To activate the speech recognition engine, the participants have to speak the phrase "Hallo Auto" (eng. "Hello car"). After speaking their request the system directly activates the appropriate function or provides the requested information. For example, when given user input for Task 1, the radio program begins to play and the screen provides information for the current radio station.

Considering that we wish to discover how users naturally interact with a SDS while driving, we place the participants in a simulated driving situation. The car in which the participants are placed is situated in front of a canvas onto which the driving simulation is projected, as done by Hofmann et al. (2014). In the driving simulation, the participants drive behind another vehicle and they have to brake if and only if the preceding vehicle brakes. The setup is illustrated in Fig. 3.

Figure 3: Driving Simulation Setup.

The overall procedure of the experiment is as follows. First, the participants are shown the pictures which they have to interpret verbally. In order to prevent incorrect interpretations, the participants are offered assistance when necessary. Second, the participant is quickly acclimated to the driving simulation through a three minutes test drive. The instructor, who sits in the passenger seat, shows the pictures arbitrarily while the participant operates the vehicle. The tasks are permuted to avoid order effects. More details are described in Braunger et al. (2017).
Since 45 subjects participated in the study, 45 utterances per task are collected respectively (in total: 315 utterances). 46% of the participants are female and 54% are male. The average age is 39.5 years (standard deviation SD: 13.5). 27% of the participants are experienced in the use of voice-controlled devices while 74% have little to no experience with SDS.

4. Experiments and Results

As normalizing preprocessing steps, we automatically spell-check the data, standardize the spelling, lowercase it, and eliminate all punctuation. For further analysis we POS-tag and parse the data using *SpaCy*[4]. The POS tagger uses the Google Universal POS tag set of Petrov et al. (2012).

4.1. Noisy Data

When it comes to assess crowdsourced data many researchers examine rates of unusable data. Previous works report rates of unusable data between 6% and 30% (Wang et al., 2012; Schmidt et al., 2015; Braunger et al., 2016). Yang Wang et al. (2012) find incorrect slot values, missing or added slots and garbage utterances. In addition, Braunger et al. (2016) find technical problems and task misunderstandings that cause faulty data. We also expect our data con-

[4]https://github.com/explosion/spaCy

tains such mistakes. The experiments we have conducted confirm this assumption.
In order to create a baseline of valid utterances we first define "keywords" that have to be named. The keywords are defined such that they represent the minimal information which is required to fulfill the predefined tasks. As an example, at least the radio station name "SWR3" needs to be mentioned to fulfill task 1. The utterances are automatically annotated and manually corrected. This procedure leads to a rate of 4.6% unusable utterances within our crowdsourced data[5].
In order to automatically identify such erroneous utterances we propose an approach based on a maximum likelihood language model (LM) which is also applicable for more complex tasks. We train a word based trigram model since it performed best within our experiments[6]. As training data, we use the hundred most frequent utterances of the preprocessed crowdsourced data. Since we observe that the hundred most frequent utterances are repeated several times by different speakers we expect this data do not contain mistakes. Before we test the model on the whole crowdsourced data set we remove the stop words in both, training set and test set. The score of an utterance x is calculated such that the impact of the utterance length UL is normalized:

$$LMscore(x) = LaplaceSmoothing^{\frac{1}{UL}} \quad (1)$$

The results are filtered by an experimentally determined threshold value. All utterances, whose probabilities are more than 80% lower than the mean score, are classified as faulty. With the help of the LM we are able to identify 83% of the faulty utterances. The remaining utterances contain only 0.8% incorrect utterances which are not captured. Examples of the mistakes we identify are given below.

a) Wrong language (instead of German):
i tell my car to switch on to 23 degrees celsius.

b) Garbage utterance:
telefonieren mit amerika 001 amerikanischer präsident rufnummer suchen barack obama telefonnummer amerika barack obama.
"Call America 001 American president search telephone number Barack Obama phone number America Barack Obama."

c) Technical problem:
aufnahme lässt sich nicht abspielen.
"The sound file cannot be played back."

d) Technical problem:
23 grad wurde wegen technische aufnahmen nicht richtig aufgenommen.
"23 degrees was not correctly recorded due to technical issues."

e) Typing error:
swchlate radio swr3.
"Typing error *radio SWR3.*"

f) Wrong intent:
auto bitte öffne in meinen handy das sms schreiben und suche den kontakt meines bruders.
"Car, please open text messaging function on my mobile and search the entry of my brother."

g) Wrong intent:
auto bitte lass das lied the way you make me feel laufen.
"Car, please play the song The way you make me feel."

h) Missing slot:
spiel das album greatest hits.
"Play the album Greatest Hits."

i) Wrong slot value:
schreibe max mustermann eine sms.
"Write a text message to Max Mustermann."

j) Task misunderstanding:
ich könnte mir vorstellen dass es gut wäre wenn ich das lied summe oder singe wird es gefunden.
"I believe it would be good if the song I am singing can be found."

Even though 15% of the truly valid utterances are classified as faulty, the high precision score for the valid utterance class (99%) shows that the model is able to robustly identify truly valid utterances. Overall, the model achieves an accuracy rate of 85% and an F_1 score of 89%. With task 3 the model performs worst and with task 2 it performs best.

4.2. Linguistic Evaluation

In order to evaluate the usefulness of crowdsourcing we check the data set acquired by a previously evaluated crowdsourcing method against data acquired by an experimental study. Since our system usage data serves as a baseline for naturalness in user input, both data sets are examined and compared in terms of syntactic, lexical, and pragmatic criteria commonly used in literature. The criteria we examine also include those mentioned by literature for natural queries: politeness, full sentences, filler words and a higher number of words (Braunger et al., 2017).
Table 1 presents the lexical properties of both, crowdsourced queries and real system usage queries (referred to as "natural data"). A common measure of lexical diversity is the type-token ratio which is calculated by dividing the number of individual word types (lemmas) by the number of occurring word tokens. The standardized type-token ratio (STTR) (Johnson, 1944) is commonly used to normalize the impact of the size of different data sets. Table 1 shows that there are no significant differences between

[5]Note that the rates of usable data may vary depending on how valid utterances are defined. Following Schmidt et al. (2015) and Braunger et al. (2016) valid utterances must contain given entities in the form of proper names such as "SWR3" and additionally variable units such as "radio", "radio station" or "station". This constraint leads to a rate of 13% unusable utterances.

[6]Within this work several language models have been experimented with: probabilistic language models and recurrent neural network based language models. Each main kind was trained on word and on character basis. The maximum likelihood model trained on word based trigrams performed best concerning the task.

crowdsourced queries and real system usage queries.
The content-function word ratio is an indicator for lexical density. The ratio contains the proportion of content words (open class words) to the number of function words (closed class words). With 68.45% in crowdsourced queries and 60.9% in natural queries people seem to provide less information when speaking to an actual system.
The keyword-to-utterance-length ratio (KLR) is mentioned by Möller et al. (2008). "Keywords" represent obligatory contents to occur in each valid query. The ratio is calculated by the number of keywords in an utterance divided by the number of tokens in an utterance. The KLR computed for both, crowdsourced queries and natural queries, shows that people tend to get to the point when fulfilling the tasks via crowdsourcing.

Lexical criteria	Crowdsourced data	Natural data
STTR	34.93%	35.26%
Keyword-utterance length ratio	37.73%	30.22%
Content-function word ratio*	68.45%	60.90%

Table 1: Lexical properties of crowdsourced data and natural data. (*Differences are significant at $p<0.05$.)

The syntactic properties are presented in Table 2. A comparison of the utterance lengths shows that utterances towards the system are significantly longer than the crowdsourced utterances. Tree depth is used by Pinter et al. (2016) and Guy (2016) as an indicator for syntactic complexity. It is calculated by the number of edges in the longest path from the root node to a leaf. The mean tree depth of the system-directed utterances is 2.94 and of the crowdsourced utterances 2.33. That is, people tend to use more complex syntactic structures when talking to an actual system. In order to conclude the syntactic analysis we examine whether people used full sentences or syntactic incomplete structures. Therefore, we rely on Braunger et al. (2017) who defined the following sentence type categories: Interrogative, Declarative, Imperative, Infinitive and Verbless, whereas an infinite and a verbless construction do not count as full sentences. For annotation, we use a statistical model[7]. The results of the model are inspected and corrected manually. As Table 2 implies, 80% of the natural queries are full sentences but only 70% of the crowdsourced queries. In addition, natural queries contain a lot more declarative constructions such as "I would like to call Barack Obama on his mobile phone" whereas crowdsourced queries contain more imperative constructions.
Besides lexical and syntactic aspects we analyze our data in terms of pragmatic properties. Civility and filler words are mentioned by some authors being salient features of natural language compared to a command or keyword style of speaking to an SDS (Berg et al., 2010; Hofmann et al., 2012). As for politeness, we analyze the occurrences of the

[7]The statistical model bases on POS tag unigrams and bigrams and achieves an accuracy of 91% on similar crowdsourced data.

Syntactic criteria	Crowdsourced data	Natural data
Mean query length	5.72	6.76
Mean tree depth	2.33	2.94
Proportion of Interrogative	3%	7%
Proportion of Declarative*	5%	27%
Proportion of Imperative*	62%	46%
Proportion of Infinitive	19%	15%
Proportion of Verbless*	11%	5%

Table 2: Syntactic properties of crowdsourced data and natural data. (*Differences are significant at $p<0.05$.)

particle "please" and we additionally rely on the findings of Danescu-Niculescu-Mizil et al. (2013) who characterize politeness indicators (cf. Schmidt and Braunger (2018)). Out of the 14 strategies they mention as being polite, we examine the following:

- Counterfactual modal: *Could/Would* you
- Indicative modal: *Can/Will* you
- 1st person start: *I* search
- 1st person plural: Could *we* find

The proportion of utterances that contain these politeness markers is shown in Table 3. The proportion of polite queries in the natural data is nearly five times higher than in the crowdsourced data. Additionally, we compare the proportion of utterances containing filler words. Filler words do not contribute to the sentence meaning. The filler words we count include disfluencies such as "ähm" (eng. "uh") and modal particles according to Bross (2012). The modal particles we find include "doch", "einmal", "nochmal", "mal", "vielleicht", "denn", "eigentlich". Table 3 shows that filler words occur more often when people were speaking to an actual SDS.

Pragmatic criteria	Crowdsourced data	Natural data
Proportion of polite queries*	13.7%	60.9%
Proportion of queries that contain filler words	4.4%	9.0%

Table 3: Pragmatic properties of crowdsourced data and natural data. (*Differences are significant at $p<0.05$.)

Overall, we find quite striking differences between crowdsourced data and real system usage data. This arises the question whether crowdsourced data contribute to the goal of enabling a more human-like SDS interaction and thus whether crowdsourced data are as useful as real system usage data. In order to answer the question, we conduct an experimental evaluation which is described in the following section.

4.3. Application-Related Evaluation

Since NLU services are already widely used in both, academia and industry, we rely on NLU services to evalu-

ate the performance of our two data sets. We use Microsoft LUIS[8] and RASA[9] since Braun et al. (2017) find that those achieve the best results compared to other popular NLU services. The data sets were semi-automatically labeled by the authors.

First, the NLU services are trained on crowdsourcing data as well as on 60% of the system usage data. As crowdsourced data, we use the utterances which are classified as valid utterances (cf. section 4.1.). The labels created by the services are then compared against 40% of our system usage gold standard data. Fig. 4 shows that the average F_1 scores are very similar between crowdsourced training data and system usage training data. The models that are trained on crowdsourced data achieve very good results when tested on gold standard data.

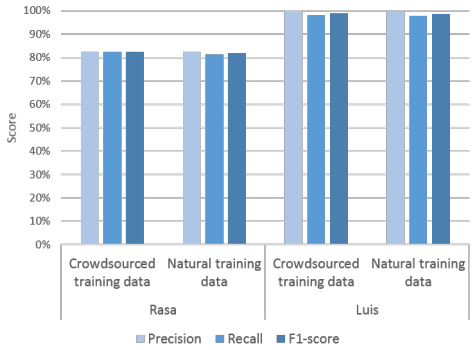

Figure 4: Average scores of intent classification and entity recognition: crowdsourced vs. natural data.

Since we identified that 4.6% of the crowdsourced data contain errors we additionally investigate the impact of such noise on performance results. In a further experiment the NLU services are therefore trained on both 100% of crowdsourced data and cleaned-up crowdsourced data (95.4%). The test set in that case consists of all gold standard utterances (100%). Fig. 5 displays the results. It is shown that the noisy data perform as well as the cleaned-up data. This can be due to the fact that the amount of training utterances is increased in this case. However, this result should be viewed with caution. Further experiments have to be conducted taking into account a more difficult task, e.g., more overlapping use cases.

5. Conclusion

We propose different methods of how to evaluate usefulness and naturalness of crowdsourced data. Therefore, we create a baseline by means of collecting real system usage data.

In order to automatically assess the quality of the crowdsourced data, we first train a word based language model on the hundred most frequent utterances and test the model on all crowdsourced utterances. We show that this method is able to successfully identify valid data. Since many works report high proportions of faulty utterances within crowdsourced data the amount of mistakes is an important criterion when it comes to assess crowdsourced data.

[8]https://www.luis.ai

[9]https://www.rasa.ai (Open source service).

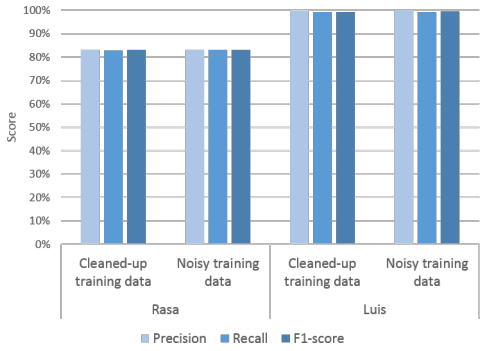

Figure 5: Average scores of intent classification and entity recognition: cleaned-up crowdsourced data vs. noisy crowdsourced data.

In order to assess the naturalness of the voice input elicited via crowdsourcing, we examine different linguistic criteria on a lexical, syntactic and pragmatic level. Our comparative analysis reveals the differences between crowdsourced data and system usage data. However, we show that when training NLU services on crowdsourced data the scores achieved are as good as system usage data, even when the test set contains faulty utterances. We conclude that for the purpose of training NLU services crowdsourced data is at least as suitable as system usage data.

6. Bibliographical References

Berg, M., Gröber, P., and Weicht, M. (2010). User study: Talking to computers. In *Proceedings of the Third Workshop on Inclusive eLearning*.

Braun, D., Mendez, A. H., Matthes, F., and Langen, M. (2017). Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Braunger, P., Hofmann, H., Werner, S., and Schmidt, M. (2016). A comparative analysis of crowdsourced natural language corpora for spoken dialog systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Braunger, P., Maier, W., Wessling, J., and Werner, S. (2017). Natural language input for in-car spoken dialog systems: How natural is natural? In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Bross, F. (2012). German modal particles and the common ground. *Helikon. A Multidisciplinary Online Journal*, 2:182–209.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors.

In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Eskenazi, M., Levow, G.-A., Meng, H., Parent, G., and Suendermann, D. (2013). *Crowdsourcing for Speech Processing: Application to Data Collection, Transcription and Assessment*. John Wiley & Sons.

Guy, I. (2016). Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hofmann, H., Ehrlich, U., Berton, A., and Minker, W. (2012). Speech interaction with the internet - a user study. In *Proceedings of the Eighth International Conference on Intelligent Environments*.

Hofmann, H., Hermanutz, M., Tobisch, V., Ehrlich, U., Berton, A., and Minker, W. (2014). Evaluation of in-car sds notification concepts for incoming proactive events. In *Proceedings of 5th International Workshop on Spoken Dialog Systems*.

Johnson, W. (1944). Studies in language behavior: I.a program of research. *Psychological Monographs*, 56:1–15.

Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., and Wang, A. (2000). Suede: A wizard of oz prototyping tool of speech user interfaces. In *Proceedings of the 13th Annual ACM Symposium on User interface Software and Technology*.

Komarov, S., Reinecke, K., and Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

Manuvinakurike, R. and DeVault, D., (2015). *Natural Language Dialog Systems and Intelligent Assistants*, chapter Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection, pages 189–201. SpringerLink.

Manuvinakurike, R., Paetzel, M., and DeVault, D. (2015). Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection. In *Proceedings of the Workshop on Semantics and Pragmatics of Dialogue 2015*.

McGraw, I., Lee, C., Hetherington, L., Seneff, S., and Glass, J. R. (2010). Collecting voices from the cloud. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*.

Misu, T. (2014). Crowdsourcing for situated dialog systems in a moving car. In *Proceedings of Interspeech*.

Mitchell, M., Bohus, D., and Kamar, E. (2014). Crowdsourcing language generation templates for dialogue systems. In *Proceedings of the Eighth International Natural Language Generation Conference (INLG)*.

Möller, S., Gödde, F., and Wolters, M. (2008). Corpus aanalysis of spoken smart-home interactions with older users. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

Novikova, J., Lemon, O., and Rieser, V. (2016). Crowdsourcing nlg data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation Conference*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2012). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of 8th International Conference on Language Resources and Evaluation (LREC)*.

Pinter, Y., Reichart, R., and Szpektor, I. (2016). Syntactic parsing of web queries with question intent. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rothwell, S., Elshenawy, A., Carter, S., Braga, D., Romani, F., Kennewick, M., and Kennewick, B. (2015). Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proceedings of Interspeech*.

Schmidt, M. and Braunger, P. (2018). Towards a speaking-style adaptive assistant for task-oriented applications. In *Proceedings of 29. Elektronische Sprachsignalverarbeitung (ESSV)*.

Schmidt, M., Müller, M., Wagner, M., Stüker, S., Waibel, A., Hofmann, H., and Werner, S. (2015). Evaluation of crowdsourced user input data for spoken dialog systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Wang, W. Y., Bohus, D., Kamar, E., and Horvitz, E. (2012). Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *Proceedings of the IEEE Workshop on Spoken Language Technologies (SLT)*.

Yang, Z., Li, B., Zhu, Y., King, I., Levow, G., and Meng, H. (2010). Collection of user judgements on spoken dialog system with crowdsourcing. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT)*.