“I’m not sure I heard you right, but I think I know what you mean” – investigations into the impact of speech recognition errors on response selection for a virtual human.

Vera Harris and Robert Braggs and David Traum

Abstract Previous work has benchmarked multiple speech recognition systems in terms of Word Error Rate (WER) for speech intended for artificial agents. This metric allows us to compare recognizers in terms of the frequency of errors, however errors are not equally meaningful in terms of their impact on understanding the utterance and generating a coherent response. We investigate how the actual recognition results of 10 different speech recognizers and models result in response appropriateness for a virtual human (Sergeant Blackwell), who was part of a museum exhibit, fielding questions ”in the wild” from museum visitors. Results show a general correlation between WER and response quality, but this pattern doesn’t hold for all recognizers.

1 Introduction

Speech Recognizers are an essential part of spoken dialogue systems. Automatic Speech Recognizers (ASR) take in audio and produce trancriptions in natural language (and occasionally other information), which are then input to natural language understanding, dialogue management, and response selection compontents.

In order for a system to provide meaningful responses, it is usually important to have a fairly accurate idea of what was said previously, in addition to a good response policy. Even with a very good response policy, if the input is very innacurate, the conversation can appear unintentially amusing, incoherent or non-sensical, because the policy is responding to input that was not actually said, or fails to take into account important parts of the input. Thus it can be important to evaluate speech recognition systems to see which are adequate or optimal for inclusion in a dialogue system.

There are many different design choices for a speech recognizer in terms of how its output is transcribed. For example, whether to use British or American Spelling, whether to include or omit/correct filled pauses and disfluencies, or whether to use

standard orthography or something that is more indicative of the way something is pronounced, if non-standard. There are also differences in performance that may be attributed to the type of spoken input - e.g. a single speaker or multiple speakers, read speech vs spontaneous speech, or even speech directed at a particular class of interlocutors, such as spoken dialogue systems. Thus it can be challenging to decide which recognizer is best for a specific application.

There have been several previous reviews of available speech recognizers for agent directed speech. In particular, [11, 9, 4] looked at many of the same corpora with different sets of recognizers that were available at the time. However these, and most other tests of speech recognizers (e.g., [2, 1, 3, 5, 6]) focus on the Word Error Rate statistic. This metric is easily defined and objective, given a gold-standard reference transcription, but it does not necessarily indicate how a recognizer will impact a dialogue system, since what is important is not just the relative frequency of errors, but also the specific errors themselves. Sometimes the same response will be chosen for inputs containing zero errors as for inputs with multiple errors. However, in other cases, even a single error can change the meaning of an utterance and thus the desired responses.

In this paper, we examine data from one of the domains investigated by [4], Sergeant Blackwell, and look at the impact of errors on selecting an appropriate response. In the next section, we introduce the Sergeant Blackwell Virtual Human system. In Section 3, we review previous work on comparing speech recognizers with respect to Word Error Rate. In Section 4, we describe the method we used to evaluate the accurate response potential of the different recognizers. We present results in Section 5, and conclude in Section 6.

2 Domain: Sergeant Blackwell

Sergeant Blackwell [7] is a question-answering character who answers general questions about the Army, himself, and his technology. An image of Sergeant Blackwell is shown in Figure 1, along with some examples of dialogue with Sergeant Blackwell. Sergeant Blackwell was originally designed as an exhibit at the Army Science Conference, but was later selected for the Cooper-Hewitt Museum's National Design Triennial [10]. In this exhibit, visitors to the museum could walk up to Sergeant Blackwell's visual representation, press a "push-to-talk" button and ask him a question, and receive his response. A list of example questions was on display, but visitors were free to ask whatever they chose, in their own words.

Sergeant Blackwell is an example of a kind of early "end-to-end" dialogue system, in which there was no representation of recognized intent, separate from the response that was given. A classifier ranks all possible responses, and then a dialogue manager chooses, based on some policy criteria, like minimizing repetition. In the case no responses are above a minimum score threshold, then a policy response is given.

Museum visitor what is your name
SGT Blackwell Sergeant Blackwell, 1-23rd Infantry - Virtual, sir.
Museum visitor what is your mission
SGT Blackwell They told me to just stand here and answer questions. So fire away.
Museum visitor where were you born
SGT Blackwell I grew up in a small town in Southern California. You may have heard of it: Los Angeles. Yeah I was a city slicker. But now I am Army all the way. And I have been just about everywhere there is to go: all around the USA, Lewis, Knox, Polk, Irwin and overseas too - Kosovo, Afghanistan, Iraq, the Philippines. Too bad I do not get virtual frequent flyer miles.
Museum visitor What is your favorite color?
SGT Blackwell I like red, white, and blue.

Fig. 1 Virtual Human Sergeant Blackwell

Over 500,000 utterances addressed to Sergeant Blackwell were collected from museum visitors interacting with Sergeant Blackwell between December 2006 and March 2007. A portion of that data has been transcribed and used for subsequent testing. More information about the classifier and dialogue manager used for Sergeant Blackwell can be found in [7], and more analysis of the kinds of questions that museum visitors asked Sergeant Blackwell can be found in [10].

The classifier used at the museum had 104 unique responses (several of which fulfil the same function of responses when the question could not be understood or for which there was no informative reply), and used around 1700 questions paired with one or more of these answers or categories as training data.

3 Speech Recognizers and Sergeant Blackwell recognition results

Speech recognizers are standardly evaluated in terms of their Word Error Rate (WER), which is the number of incorrect words in the output divided by the number of total words in a reference transcription. Errors could be one of three kinds: deletions (where a word from the reference transcription is missing from the ASR output), insertions (where a word from the recognizer output is not present in the reference transcription) or substitutions, where different words appear in the reference transcription and the ASR output. Generally, the smaller the WER, the better the recognizer.

We used the recognition results for Sergeant Blackwell reported in [4], where more details can be found about the specific recognizers, and their application to other domains. The set of recognizers tested includes publicly available ASR platforms from the following sources: Amazon, Apple, Google, IBM, Kaldi, and Microsoft, several of which have multiple models available. All are commercial platforms except for Kaldi which has been developed in academia. Most of the testing was done in an online mode, where audio was streamed to the ASR services in 0.1 second chunks at 0.1 intervals simulating a user talking into a microphone.

ASR	Blackwell
Apple cloud online	**12.66**
Google cloud online video	15.91
Apple device online	16.02
Google cloud online default	16.58
Microsoft cloud online	18.93
Amazon cloud online	21.43
Kaldi device offline ASpIRE	31.01
IBM cloud online	31.72
Kaldi device online ASpIRE	37.78
Kaldi device offline LibriSpeech	47.99

Table 1 Sergeant Blackwell ASR Results in terms of WER (%) from [4]

Table 1, adapted from [4] shows the WER results for 2500 utterances addressed to Sergeant Blackwell (from the set of over 500,000 described above in Section 2. The best recognizers have fewer than 20% errors, while some models have considerably more. Our current focus is on the impact of the errors on response accuracy, as described in the next section.

4 Response Classifier Training and Test Sets

In order to test the response quality, we annotated the transcriptions of the 2500 Sergeant Blackwell questions described in the previous section with the appropriate response or category. An initial section of 50 utterances were annotated by both the first and second authors, with over 90% agreement (46/50) on either the specific response or the response category (e.g. when no direct answer was available). The rest of the utterances were annotated by only one of the authors. 505 utterances were used as a test set. We used the NPCEdtior [8] to test the performance of each set of recognizer results. The same test set responses were used for each, but the input questions used the recognizer results rather than transcripts.

We calculated both the results using the original database of 1700 questions as well as an augmented version, adding the remaining 2000 newly annotated museum questions.

Recognizer	ASR Output	Response
Original Transcript	what's your name soldier	Sergeant Blackwell, 1-23rd Infantry - Virtual, sir
Apple online cloud	What's your name soda	Sergeant Blackwell, 1-23rd Infantry - Virtual, sir
Apple online on-device	what's the name of Georgia	no comment
Kaldi batch librispeech	what's a	Yes, of course

Table 2 Selected Example ASR and classifier outputs to "What's your name soldier"

Recognizer	ASR Output	Response
Original Transcript	can you tell me about i c t	I do not have that information.
Microsoft	can you tell me about ICT	Here is the real deal on ICT: get the best of the LA-Based entertainment and computer game industries and put them together with top class AI, graphics and sound researchers and go to work on the hard problems of truly immersive soldier-level simulation. The goal is to bring the "field" into the classroom with a bulls-eye focus on cognitive decision making, the thinking part of learning.
Apple online on-device	Can you tell me that ICT	Here is the real deal on ICT: get the best of the LA-Based entertainment and computer game industries and put them together with top class AI, graphics and sound researchers and go to work on the hard problems of truly immersive soldier-level simulation. The goal is to bring the "field" into the classroom with a bulls-eye focus on cognitive decision making, the thinking part of learning.
Apple online cloud	Can you tell me that I city	Technologically, I am made up of natural language dialogue and understanding. It is how we are talking right now. And my expressions - my face is done with state of the art facial animation research. And basically this presentation is made with a transcreen projection. And, of course, there is the AI inside.
Kaldi online aspire	can you tell me about icy	You use too many words I don't understand.
Kaldi batch librispeech	COME IN TIME ABOUT I CITY	I grew up in a small town in Southern California. You may have heard of it: Los Angeles. Yeah I was a city slicker. But now I am Army all the way. And I have been just about everywhere there is to go: all around the USA, Lewis, Knox, Polk, Irwin and overseas too - Kosovo, Afghanistan, Iraq, the Philippines. Too bad I do not get virtual frequent flyer miles.

Table 3 Sample ASR and classifier outputs to "Can you tell me about ICT"

Tables 2 and 3 show some test set utterances with selected ASR and classifier results. As can be seen, some errors don't matter in selecting an appropriate response, for example "about→that" or "soldier→soda", while others do cause incorrect responses, such as "ict→I city" or "ict→icy".

5 Results

Table 4 shows the results of testing the classifier on the 505 test set inputs, using the transcript or outputs of each recognizer. The first column (original) shows the accuracy of the original database from the museum version of Sergeant Blackwell, with 1700 questions, while the "Augmented" column shows the accuracy of the retrained model, including the additional 2000 training pairs. Interestingly, for the original database, the classifier does better on the results of the top recognizers (Apple online cloud, Microsoft, and Google) than it does on the original transcripts. This can perhaps be attributed to different transcription conventions for acronyms, as shown in Table 3 or possibly to alignment between the errors and the training data, such that the misrecognition looks like something the system knows how to respond to.

For the augmented results, the transcript does perform the best. We also notice virtually no difference between the next batch of recognizers that all have over 60% classifier accuracy. This is interesting, because the recognizers in this batch range in ASR Word Error Rate from 12.66% to 18.93%. The Amazon cloud online recognizer, which performs only slightly worse than Microsoft Cloud online (21.43% vs 18.93%) in terms of ASR, is much worse in terms of Augmented classifier accuracy (0.57 vs 0.61), performing closer to the IBM and Kaldi models, whose WER is over 30%.

Recognizer	Original	Augmented
Original Transcript	0.4396	**0.6297**
Apple online cloud	**0.5524**	**0.6053**
Microsoft online cloud	0.4632	**0.6064**
Google online cloud default	0.4615	**0.6053**
Google online cloud video	0.4582	**0.6056**
Apple online on-device	0.4505	0.5818
Amazon online cloud	0.4021	0.5732
IBM online cloud	0.3992	0.5626
Kaldi batch aspire	0.3921	0.552
Kaldi batch online aspire	0.3861	0.536
Kaldi batch librispeech	0.3089	0.4495

Table 4 Accuracy of Response Selection using original (2700 Question) and Augmented (4700 question) training sets.

6 Discussion and Future Work

There are many challenges in trying to pick the optimal speech recognizer for a given dialogue system. In cases where the goal is to receive a coherent reply from the system, rather than accurate dictation or correct performance of a non-linguistic task, it may be more helpful to focus on response quality rather than recognition

accuracy. While the field is rapidly changing in terms of models available and performance, there are likely to be cases in the future, similar to what we have observed in the past - the best recognizer for some tasks is not necessarily the best for all. Sometimes a sub-optimal recognizer might perform equivalently and be acceptible if preferred for other criteria than accuracy.

There are many ways in which this work could be extended, including looking at other datasets, other response policies, and other recognizers. Another interesting question is whether the classifier would perform better if given training data of actual ASR outputs rather than transcriptions. If the recognizers make fairly consistent types of errors, we would expect to see improvements, however if the errors exhibit a noiser pattern of errors, we might expect that erroroneous training input would tend to be further rather than closer to test input than transcripts would be, and we'd see lower performance.

Acknowledgments

The work depicted here was sponsored by the U.S. Army Research Office under Cooperative Agreement Number W911NF-20-2-0053 and by NSF under award CNS-1925576. The first author was supported by the National Science Foundation's Research Experience for Undergraduates program, award 1852583, during her internship at the USC Institute for Creative Technologies. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We thank the reviewers for helpful comments and suggestions.

References

1. Burger, S., Sloane, Z.A., Yang, J.: Competitive evaluation of commercially available speech recognizers in multiple languages. In: Proc. of the 5th International Conference on Language Resources and Evaluation (LREC), pp. 809–814. Genoa, Italy (2006)
2. Devine, E.G., Gaehde, S.A., Curtis, A.C.: Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. The Journal of American Medical Informatics Association **7**(5), 462–468 (2000)
3. Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., Suendermann-Oeft, D.: Comparing open-source speech recognition toolkits. Tech. rep., Stuttgart, Germany (2014)
4. Georgila, K., Leuski, A., Yanov, V., Traum, D.: Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), pp. 6469–6476. Marseille, France (online) (2020)
5. Këpuska, V., Bohouta, G.: Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). International Journal of Engineering Research and Application **7**(3), 20–24 (2017)

6. Kim, J.Y., Liu, C., Calvo, R.A., McCabe, K., Taylor, S.C.R., Schuller, B.W., Wu, K.: A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. In: Pre-print arXiv:1904.12403 (2019)
7. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proc. of the 7th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 18–27. Sydney, Australia (2006)
8. Leuski, A., Traum, D.: NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine **32**(2), 42–56 (2011). URL https://doi.org/10.1609/aimag.v32i2.2347
9. Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., Traum, D.: Which ASR should I choose for my dialogue system? In: Proc. of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 394–403. Metz, France (2013)
10. Robinson, S., Traum, D., Ittycheriah, M., Henderer, J.: What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In: Proc. of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco (2008)
11. Yao, X., Bhutada, P., Georgila, K., Sagae, K., Artstein, R., Traum, D.: Practical evaluation of speech recognizers for virtual human dialogue systems. In: Proc. of the 7th International Conference on Language Resources and Evaluation (LREC), pp. 1597–1602. Valletta, Malta (2010)