Last Words

Revisiting the Boundary between ASR and NLU in the Age of Conversational Dialog Systems

Manaal Faruqui
Google Assistant
mfaruqui@google.com

Dilek Hakkani-Tür
Amazon Alexa AI
hakkanit@amazon.com

As more users across the world are interacting with dialog agents in their daily life, there is a need for better speech understanding that calls for renewed attention to the dynamics between research in automatic speech recognition (ASR) and natural language understanding (NLU). We briefly review these research areas and lay out the current relationship between them. In light of the observations we make in this article, we argue that (1) NLU should be cognizant of the presence of ASR models being used upstream in a dialog system's pipeline, (2) ASR should be able to learn from errors found in NLU, (3) there is a need for end-to-end data sets that provide semantic annotations on spoken input, (4) there should be stronger collaboration between ASR and NLU research communities.

1. Introduction

More and more users every day are communicating with conversational dialog systems present around them like Apple Siri, Amazon Alexa, and Google Assistant. As of 2019, 31% of the broadband households in the United States have a digital assistant.[1] Henceforth, we refer to these systems as **dialog agents** or simply **agents**. A majority of queries issued to these dialog agents are in the form of speech as the users are directly talking to these agents hands-free.

This is in contrast to a few years ago, when most of the traffic to search engines like Google Search, Yahoo!, or Microsoft Bing was in the form of text queries. The natural language understanding (NLU) models that underlie these search engines were tuned

1 https://www.statista.com/statistics/791575/us-smart-speaker-household-ownership/.

Submission received: 26 April 2021; revised version received: 20 August 2021; accepted for publication: 4 December 2021.

https://doi.org/10.1162/COLI_a_00430

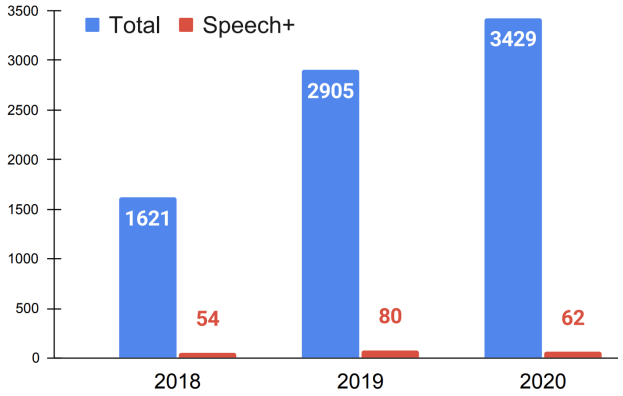

Figure 1
The number of submitted papers in the speech processing (+ multimodal) track vs. total in ACL conference from 2018–2020.

to handle textual queries typed by users. However, with the changing nature of query stream from text to speech, these NLU models also need to adapt in order to better understand the user.

This is an opportune moment to bring our attention to the current state of automatic speech recognition (ASR) and NLU, and the interface between them. Traditionally, the ASR and NLU research communities have operated independently, with little cross-pollination. Although there is a long history of efforts to get ASR and NLU researchers to collaborate, for example, through conferences like HLT and DARPA programs (Liu et al. 2006; Ostendorf et al. 2008), the two communities are diverging again. This is reflected in their disjoint set of conference publication venues: ICASSP, ASRU/SLT, and Interspeech are the major conferences for speech processing, whereas ACL, NAACL, and EMNLP are the major venues for NLU. Figure 1 shows the total number of submitted papers to the ACL conference, and in the speech processing track from 2018–2020 at the same conference.[2] The number of submitted speech-related papers at the conference constitute only 54 (3.3%), 80 (2.7%), and 62 (1.8%) in 2018, 2019, and 2020, respectively, showing the limited amount of interaction between these fields.

In this article, we analyze the current state of ASR, NLU, and the relationship between these two large research areas. We classify the different paradigms in which ASR and NLU operate currently and present some ideas on how these two fields can benefit from transfer of signals across their boundaries. We argue that a closer collaboration and re-imagining of the boundary between speech and language processing is critical for the development of next generation dialog agents, and for the advancement of research in these areas.

Our call is specially aimed at the computational linguistics community to consider peculiarities of spoken language, such as disfluencies and prosody that may carry additional information for NLU, and errors associated with speech recognition as a core part of the language understanding problem. This change of perspective can lead to the

2 The speech processing area also included papers from other related areas like multimodal processing, so the numbers presented here are an upper bound on speech processing papers.

creation of data sets that span across the ASR/NLU boundary, which in turn will bring the NLU systems closer to real-world settings as well as increase collaboration between industry and academia.

2. Changing Nature of Queries

As users move from typing their queries to conversing with their agent through dialog, there are a few subtle phenomena that differ in the nature of these queries. We briefly discuss these in the following section.

2.1 Structure of the Query

User-typed queries aren't always well-formed nor do they always follow the syntactic and semantic rules of a language (Bergsma and Wang 2007; Barr, Jones, and Regelson 2008; Manshadi and Li 2009; Mishra et al. 2011). This is not surprising because users often want to invest as little effort as they can in typing a query. They provide the minimum required amount of information in the form of words that they think can return the desired result, and, hence, typed search queries are mostly a bag of keywords (Baeza-Yates, Calderón-Benavides, and González-Caro 2006; Zenz et al. 2009).

On the other hand, spoken utterances addressed to dialog agents are closer to natural language, contain well-formed sequence of words that form grammatical natural language sentences (though spoken language can also be ungrammatical, cf. §2.2), and are more complex and interactive (Trippas et al. 2017, 2018). For example, sometimes a user utterance might need to be segmented into multiple parts:

Agent: *"What movie genre do you like?"*
User: *"I don't know <pause> anime"*

Here, if we don't explicitly use the <pause> marker to segment the user utterance, we might understand the user intent as *"I don't know anime"* instead of the selected genre being *"anime"*. Thus, traditional techniques of information retrieval are not effective in understanding such conversational utterances which need deeper understanding. Table 1 shows how the same intent is surfaced in typed vs. spoken utterances.

Another common phenomenon in spoken utterances is disfluencies. When a user stammers, repeats, corrects themselves, or changes their mind in between the utterance, they introduce disfluency in the utterance (Schriberg 1994). For example, as shown in

Table 1
Examples of typed and spoken queries that have the same user intent. Spoken queries have phenomenon like disfluencies which are not part of typed queries.

Typed	Spoken
barack obama age	what is the age of barack obama
boston denver flight	book a flight *from london to umm no* from boston to denver
scooby doo breed	tell me what's the breed of scooby doo

Table 1, *"book a flight from london to umm no from boston to denver"* is a disfluent query. There has been limited NLU research done on correcting speech recognition errors or handling disfluencies. For example, in the last 27 years there has been only one major data set available containing annotated disfluencies in user utterances (Godfrey and Holliman 1993; Schriberg 1994).

2.2 Errors in the Query

Not only do spoken vs. typed queries differ in structure and style, they also vary in the kind of noise or anomalies in the input. Whereas typed queries can contain spelling errors, spoken queries can contain **speech recognition errors**, and **endpointing** issues, which we discuss below.

While there has been extensive research on correcting spelling errors (Hládek, Staš, and Pleva 2020) including state-of-the-art neural machine translation models being launched in products (Lichtarge et al. 2019),[3] there has been limited NLU research done on correcting speech recognition errors.

Handling speech recognition errors is crucial for downstream NLU models to work effectively because an error in the transcription of a single word can entirely change the meaning of a query. For example, a user utterance: *"stair light on"* is transcribed by the cloud Google Speech Recognizer[4] as: *"sterilite on"*. In this example, if the NLU model is given the query *"sterilite on"*, it is very hard for it to uncover the original intent of the user, which was to turn on the stair lights, unless we force the NLU model to learn to correct/handle such systematic ASR errors. Such error analysis is often done for industrial query logs (Shokouhi et al. 2014) but these data sets are not publicly available for academic research purposes.

Another common error that affects NLU is speech endpointing. If a user utterance is pre-planned (such as, *"play music"*), the user does not hesitate, but if the utterance is complex or the user is responding to a request from the agents, such as, *"do you prefer 5 or 7?"*), the user may hesitate when responding and pause, saying *"oh <long pause> 7"* causing the ASR model to assume that the user stopped after saying *"oh"* which leads to incomplete queries being transmitted to NLU. On the other hand, if we can learn to provide a signal to endpointing from the dialog manager that the recognized utterance is missing a value (in this case, probably an integer) according to the conversation context, we can improve endpointing, and hence recognition and downstream NLU. Data sets pertaining to endpointing phenomenon are currently available in very limited domains (Raux and Eskenazi 2008) and there is a need for such experimentation in broader domains and the impact analysis of such errors on overall user experience.

Thus, it is imperative for NLU models to be aware of the fact that there could be speech recognition errors in the input that might need to be corrected.

3. Current State of Research

We now describe the current core areas of research related to understanding the meaning of a written piece of text or spoken utterance. Figure 2 shows a general spoken language understanding pipeline.

3 https://cloud.google.com/blog/products/productivity-collaboration/using-neural-machine-translation-to-correct-grammatical-in-google-docs.

4 https://cloud.google.com/speech-to-text.

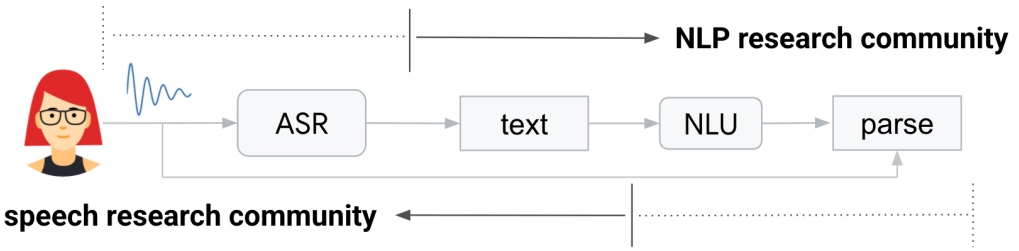

Figure 2
The current focus of speech and NLU research community (dark lines) and preferred focus of speech and NLU community (dotted lines) in future.

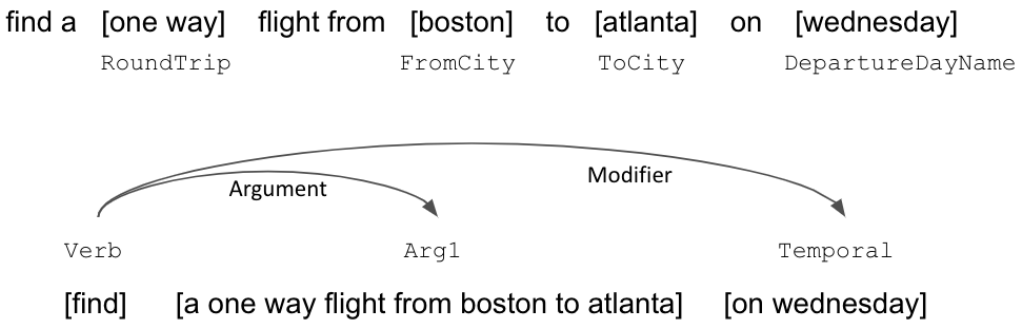

Figure 3
SLU annotation (top) and NLU semantic role labeling annotation (bottom) on a sentence from the English ATIS corpus (Price 1990), a popular SLU benchmark.

3.1 Natural Language Understanding

Natural language understanding refers to the process of understanding the semantic meaning of a piece of text. More concretely, understanding the semantic meaning often implies semantic parsing in NLP academic research and industry. Semantic parsing is the task of converting a natural language utterance to a logical form: a machine-understandable representation of its meaning. Semantic parsing is a heavily studied research area in NLP (Kamath and Das 2019).

In its simplest form, a semantic parse of a given query can contain a label identifying the desired *intent* of the query and the *arguments* that are required to fulfill the given intent. However, a more informative semantic parse can contain edges describing relations between the arguments, as in the case of abstract meaning representations (AMR) (Banarescu et al. 2013), or a graph describing how different words join together to construct the meaning of a sentence, as in the case of combinatory categorial grammar (CCG) (Steedman 1987).[5] Figure 3 shows a parse containing semantic role labels of a given query obtained using the AllenNLP tool (Gardner et al. 2018).

5 Reviewing all different semantic parsing formulations is beyond the scope of this paper.

In general, all the semantic parsing research done in NLU assumes the input to the parser as a piece of text (and available world context). Thus, the training/evaluation data set of all available semantic parsing data sets are devoid of any speech phenomenon like the presence of speech recognition errors, annotation of pauses, or disfluencies.

3.2 Spoken Language Understanding

Spoken language understanding (SLU) refers to the process of identifying the meaning behind a spoken utterance (De Mori et al. 2008; Tur and De Mori 2011). In that regard, the end goal of NLU and SLU is the same but the input to NLU and SLU components are different: text in the former, and speech input in the latter. The most common SLU task is intent prediction and slot-filling, which involves classifying the intent of the utterance and identifying any required arguments to fulfill that intent (Price 1990). Figure 3 shows the SLU annotation for a slot-filling task. We now review the main approaches used to solve SLU.

3.2.1 Speech → Text → Semantic Parse. The traditional way of performing speech understanding is to use a pipeline approach: First use an ASR system to transcribe the speech into text and then run NLU on the transcribed text to result into a semantic parse. Using a pipeline approach has its own set of pros and cons. Using a 2-step pipeline approach is modular. The first component is an ASR system and the second component is an NLU system. The errors of each module can be independently analyzed and corrected (Fazel-Zarandi et al. 2019; Wang et al. 2020). The training of both models is also independent, which makes it easier to use off-the-shelf start-of-the-art ASR and NLU models during inference.

The obvious disadvantage of this method is that the NLU and ASR models are unaware of each other. Because the models are not trained jointly, the ASR model cannot learn from the fact that the downstream NLU model could have failed on an erroneous ASR prediction. Similarly, at inference time the NLU model relies only on the best prediction of the ASR model and cannot exploit the uncertainty in ASR's prediction. However, this can be fixed to a good extent by forcing the ASR model to propagate the n-best list of speech transcript hypotheses to the NLU system and letting the NLU model use all the hypotheses together to make the semantic parse prediction (Hakkani-Tür et al. 2006; Deoras et al. 2012; Weng et al. 2020; Li et al. 2020b) or using a lattice or word-confusion network as input (Tür, Deoras, and Hakkani-Tür 2013; Ladhak et al. 2016).

3.2.2 Speech → Semantic Parse. There is a renewed focus of attention on approaches to directly parse the spoken utterance to derive the semantics by making a deep neural network consume the speech input and output the semantic parse (Haghani et al. 2018; Chen, Price, and Bangalore 2018; Serdyuk et al. 2018; Kuo et al. 2020). This is an end-to-end approach that does not rely on the intermediate textual representation of the spoken utterance produced by an ASR system. Thus, this system can be trained end-to-end with direct loss being optimized on the semantic parse prediction. However, such end-to-end models are data-hungry and suffer from lack of training data (Lugosch et al. 2019; Li et al. 2020a). Even though such models often have better performance on benchmark data sets, deploying such models in a user-facing product is difficult because of the lack of ease of debugging and fixing errors in output (Glasmachers 2017).

4. Reimagining the ASR–NLU Boundary

In Section 3 we saw that parsing the user query is a step in SLU. And thus SLU is a large umbrella utilizing NLU techniques to parse spoken language. Even though both SLU and NLU at some level are solving the same problem, there is a clear disconnect between the way problems are formulated and the way solutions are devised for these problems. On one hand, in NLU, a lot of emphasis is laid on understanding deep semantic structures in text formulated in the tasks of semantic parsing, dependency parsing, language inference, question answering, coreference resolution, and so forth. On the other hand, SLU is mainly concerned with information extraction on the spoken input formulated in the tasks of slot filling, dialog state modeling , and so on.

Even though there are academic benchmarks available for SLU that aim to extract information from the spoken input, there is an informal understanding between the ASR and NLU communities that assumes that as long as the ASR component can transcribe the spoken text correctly, the majority of the language understanding burden can be taken up by the NLU community. Similarly, there is an implicit assumption in the NLU community that ASR will provide correct transcription of the spoken input and hence NLU does not need to account for the fact that there can be errors in the ASR prediction. We consider the absence of an explicit two-way communication between the two communities problematic.

Figure 2 shows how the NLU research community can expand its domain to also consider spoken language as input to the NLU models instead of pure text. Similarly, the ASR community can also account for whether the text they produced can result in a semantically coherent piece of text or not by explicitly trying to parse the output. That said, there are already a few efforts which have tried blurring the boundary between ASR and NLU. We will first give some examples of how ASR and NLU can learn from each other and then review some existing initial work in this domain aimed at enriching existing academic benchmark data sets.

4.1 ASR → NLU Transfer

A significantly large missing portion of information in understanding a spoken input is the nature of speech, which we often refer to as **prosody**. Whether the user was angry, happy, in a rush, frustrated, and so on, can help us better understand what the user's real intent was. For example, *"no…don't stop"* and *"no don't…stop"* have exactly opposite meanings depending on whether the user paused between first and second words or the second and third words. This information can only be transferred from speech to NLU. Amazon Alexa already has such a tone-detection feature deployed in production.[6] There are academic data sets that map speech to emotion (Livingstone and Russo 2018), but academic benchmarks containing examples of intonation affecting the NLU output do not exist.

An ASR system can provide more information than its best guess to the NLU model by providing a list of n-best speech hypotheses. Unfortunately, most of the state-of-the-art NLU models are trained to only accept a single string of text as input, be it parsing, machine translation, or any other established NLU task. To some extent, SLU has enlarged the domain for understanding tasks by creating benchmark data sets that

6 https://onezero.medium.com/heres-how-amazon-alexa-will-recognize-when-you-re-frustrated-a9e31751daf7.

contain *n*-best speech hypotheses lists—for example, the dialog state tracking challenge data set DSTC-2 (Henderson, Thomson, and Williams 2014). This allows the language understanding model to make use of all the *n*-best instead of just relying on the top ASR output.[7]

4.2 NLU → ASR Transfer

Making sure that the output produced by an ASR model can be understood by an NLU model can help improve transcription quality (Velikovich et al. 2018). For example, trying to boost paths in the ASR lattice that contain named entities as predicted by a named entity recognition (NER) model can help overcome recognition errors related to out-of-vocabulary words (Serrino et al. 2019).

ASR models can also learn from the errors produced in the NLU model. If a downstream NLU model in a conversational agent cannot reliably parse a given ASR output, this might indicate the presence of speech recognition errors. If there is a reliable way to identify the cause for NLU failure as a speech error, then such examples can be provided back to the ASR module to improve itself. In Google Home, Faruqui and Christensen (2020) propose a method for picking out the correct transcription from the *n*-best hypotheses if the top hypothesis does not parse, and explicitly confirm the new hypothesis with the user in a dialog. If the user accepts the selected speech hypothesis, this correction is provided as a training example to the ASR system. In general, if the ASR models start computing error based on whether or not the produced text can be semantically parsed, the performance metric will be more representative of the real-world setting instead of the currently widely used word-error-rate (WER) metric (He, Deng, and Acero 2011).

4.3 Spoken NLU Data Sets

There has already been some progress on the front of enriching existing NLU benchmarks with speech. We would now briefly review these efforts.

4.3.1 Data Sets with Speech Input

Synthesized Speech. Li et al. (2018) presented a new data set called Spoken-SQuAD, which takes the existing NLU data set SQuAD (Rajpurkar et al. 2016) containing textual questions and textual documents. The Spoken-SQuAD data set contains the audio form of the document that has been artificially constructed by using Google text-to-speech system, and then the textual form of the document was generated using the CMU Sphinx speech recognizer (Walker et al. 2004). You et al. (2021) have created the Spoken-CoQA data set from the CoQA data set (Reddy, Chen, and Manning 2019) using the same technique. Both of these systems have shown that the presence of ASR errors has a devastating effect on the quality of the QA system. However, it is worth noting that this speech still does not reflect what people do in spontaneous interactions.

Natural Speech. The above data sets contain artifically synthesized speech. The OSDQA data set (Lee et al. 2018), on the other hand, was constructed by recruiting 20 speakers

7 Note that lack of diversity in the *n*-best speech hypotheses could be an issue and directly using the recognition lattice produced by ASR might be more informative.

for speaking out the documents from the original QA data set (Shao et al. 2018). This data set is for Chinese QA and contains spoken Chinese documents as audio. In order to accurately model the real-world setting of SLU, we need to construct a data set containing real spoken utterances similar to the approach used in OSDQA.

There are certain drawbacks with both the *artificial* and *natural*-speech style data sets. While artificially generated speech suffers from lack of sufficient speech style variability and the absence of natural speech cues, naturally generated speech comes with strict privacy and scalability concerns, preventing a large-scale collection of human speech. This privacy concern is even more pressing when dealing with utterances that humans issue to dialog agents at home that contain personal information.

4.3.2 Data Sets with Speech Errors. Instead of providing audio input in the data set, another line of effort is about adding speech recognition errors in the transcribed text. For example, RADDLE (Peng et al. 2021) is a benchmark data set and an evaluation platform for dialog modeling where the input text can contain speech phenomena like verbosity, and speech recognition errors. Similarly, the LAUG toolkit (Liu et al. 2021) provides options to evaluate dialog systems against noise perturbations, speech characteristics like repetitions, corrections, and language variety. NoiseQA (Ravichander et al. 2021) contains ASR errors in the questions of the QA data set introduced both using synthetic speech and natural speech.

5. Conclusion

In this article we have argued that there is a need for revisiting the boundary between ASR and NLU systems in the research community. We are calling for stronger collaboration between the ASR and NLU communities given the advent of spoken dialog agent systems that need to understand spoken content. In particular, we are calling for NLU benchmark data sets to revisit the assumption of starting from text, and instead move toward a more end-to-end setting where the input to the models is in the form of speech as is the case in real-world dialog settings.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. We thank Shachi Paul, Shyam Upadhyay, Amarnag Subramanya, Johan Schalkwyk, and Dave Orr for their comments on the initial draft of the article.

References

Baeza-Yates, Ricardo, Liliana Calderón-Benavides, and Cristina González-Caro. 2006. The intention behind web queries. In *String Processing and Information Retrieval*, pages 98–109, Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/11880561_9

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Barr, Cory, Rosie Jones, and Moira Regelson. 2008. The linguistic structure of English web-search queries. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1021–1030. https://doi.org/10.3115/1613715.1613848

Bergsma, Shane and Qin Iris Wang. 2007. Learning noun phrase query segmentation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 819–826.

Chen, Yuan-Ping, Ryan Price, and Srinivas Bangalore. 2018. Spoken language understanding without speech recognition. In *Proceedings of ICASSP*, pages 6189–6193. https://doi.org/10.1109/ICASSP.2018.8461718

De Mori, R., F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur. 2008. Spoken language understanding. *IEEE Signal Processing Magazine*, 25(3):50–58. https://doi.org/10.1109/MSP.2008.918413

Deoras, Anoop, Ruhi Sarikaya, Gökhan Tür, and Dilek Hakkani-Tür. 2012. Joint decoding for speech recognition and semantic tagging. In *INTERSPEECH 2012*, pages 1067–1070. https://doi.org/10.21437/Interspeech.2012-324

Faruqui, Manaal and Janara Christensen. 2020. Contextual error correction in automatic speech recognition. *Technical Disclosure Commons*.

Fazel-Zarandi, Maryam, Longshaokan Wang, Aditya Tiwari, and Spyros Matsoukas. 2019. Investigation of error simulation techniques for learning dialog policies for conversational error recovery. *arXiv preprint arXiv:1911.03378*.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6. https://doi.org/10.18653/v1/W18-2501

Glasmachers, Tobias. 2017. Limits of end-to-end learning. *CoRR*, abs/1704.08305.

Godfrey, John and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62. Philadelphia: Linguistic Data Consortium.

Haghani, Parisa, Arun Narayanan, Michiel Adriaan Unico Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Jose Moreno Mengibar, Delia Qu, Rohit Prabhavalkar, and Austin Waters. 2018. From audio to semantics: Approaches to end-to-end spoken language understanding. In *Spoken Language Technology Workshop (SLT)*, pages 720–726. https://doi.org/10.1109/SLT.2018.8639043

Hakkani-Tür, Dilek, Frédéric Béchet, Giuseppe Riccardi, and Gokhan Tur. 2006. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514. https://doi.org/10.1016/j.csl.2005.07.005

He, Xiaodong, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5632–5635. https://doi.org/10.1109/ICASSP.2011.5947637

Henderson, Matthew, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272. https://doi.org/10.3115/v1/W14-4337

Hládek, Daniel, Ján Staš, and Matúš Pleva. 2020. Survey of automatic spelling correction. *Electronics*, 9(10):1670. https://doi.org/10.3390/electronics9101670

Kamath, Aishwarya and Rajarshi Das. 2019. A survey on semantic parsing. In *Automated Knowledge Base Construction (AKBC)*, https://openreview.net/forum?id=HylaEWcTT7

Kuo, Hong-Kwang Jeff, Zoltán Tüske, Samuel Thomas, Yinghui Huang, Kartik Audhkhasi, Brian Kingsbury, Gakuto Kurata, Zvi Kons, Ron Hoory, and Luis A. Lastras. 2020. End-to-end spoken language understanding without full transcripts. *CoRR*, abs/2009.14386. https://doi.org/10.21437/Interspeech.2020-2924

Ladhak, Faisal, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, and Björn Hoffmeister. 2016. LatticeRnn: Recurrent neural networks over lattices. In *Proceedings of Interspeech*, pages 695–699, https://doi.org/10.21437/Interspeech.2016-1583

Lee, Chia Hsuan, Shang-Ming Wang, Huan-Cheng Chang, and Hung-Yi Lee. 2018. ODSQA: Open-domain spoken question answering dataset. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 949–956. https://doi.org/10.1109/SLT.2018.8639505

Li, Chia-Hsuan, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Proceedings of Interspeech*, pages 3459–3463.

Li, Jinyu, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. 2020a. On the comparison of popular end-to-end models for large scale speech recognition.

arXiv preprint arXiv: 2005.14327. https://doi.org/10.21437/Interspeech.2020-2846

Li, Mingda, Weitong Ruan, Xinyue Liu, Luca Soldaini, Wael Hamza, and Chengwei Su. 2020b. Improving spoken language understanding by exploiting ASR n-best hypotheses. *arXiv preprint arXiv:2001.05284*.

Lichtarge, Jared, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 2019. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3291–3301. https://doi.org/10.18653/v1/N19-1333

Liu, Jiexi, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480. https://doi.org/10.18653/v1/2021.acl-long.192

Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540. https://doi.org/10.1109/TASL.2006.878255

Livingstone, Steven R. and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35. https://doi.org/10.1371/journal.pone.0196391

Lugosch, Loren, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*. https://doi.org/10.21437/Interspeech.2019-2396

Manshadi, Mehdi and Xiao Li. 2009. Semantic tagging of web search queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 861–869. https://doi.org/10.3115/1690219.1690267

Mishra, Nikita, Rishiraj Saha Roy, Niloy Ganguly, Srivatsan Laxman, and Monojit Choudhury. 2011. Unsupervised query segmentation using only query logs. In *Proceedings of WWW*. https://doi.org/10.1145/1963192.1963239

Ostendorf, Mari, Benoît Favre, Ralph Grishman, Dilek Hakkani-Tur, Mary Harper, Dustin Hillard, Julia Hirschberg, Heng Ji, Jeremy G. Kahn, Yang Liu, Sameer Maskey, Evgeny Matusov, Hermann Ney, Andrew Rosenberg, Elizabeth Shriberg, Wen Wang, and Chuck Wooters. 2008. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3):59–69. https://doi.org/10.1109/MSP.2008.918023

Peng, Baolin, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. RADDLE: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4418–4429. https://doi.org/10.18653/v1/2021.acl-long.341

Price, P. J. 1990. Evaluation of spoken language systems: The ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop*. https://doi.org/10.3115/116580.116612

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264

Raux, Antoine and Maxine Eskenazi. 2008. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 1–10. https://doi.org/10.3115/1622064.1622066

Ravichander, Abhilasha, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W. Black. 2021. NoiseQA: Challenge Set Evaluation for User-Centric Question Answering. In *Conference of the*

European Chapter of the Association for Computational Linguistics (EACL). https://doi.org/10.18653/v1/2021.eacl-main.259

Reddy, Siva, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266. https://doi.org/10.1162/tacl_a_00266.

Schriberg, Elisabeth. 1994. Preliminaries to a theory of speech disfluencies. Technical report, University of California at Berkeley.

Serdyuk, Dmitriy, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. 2018. Towards end-to-end spoken language understanding. *CoRR*, abs/1802.08395. https://doi.org/10.1109/ICASSP.2018.8461785

Serrino, Jack, Leonid Velikovich, Petar Aleksic, and Cyril Allauzen. 2019. Contextual recovery of out-of-lattice named entities in automatic speech recognition. In *Proceedings of Interspeech*, pages 3830–3834. https://doi.org/10.21437/Interspeech.2019-2962

Shao, Chih Chieh, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: A Chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Shokouhi, Milad, Rosie Jones, Umut Ozertem, Karthik Raghunathan, and Fernando Diaz. 2014. Mobile query reformulations. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1011–1014. https://doi.org/10.1145/2600428.2609497

Steedman, Mark. 1987. Combinatory grammars and parasitic gaps. *Natural Language and Linguistic Theory*, 5:403–439. https://doi.org/10.1007/BF00134555

Trippas, Johanne R., Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the design of spoken conversational search: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, CHIIR '18, pages 32–41. https://doi.org/10.1145/3176349.3176387

Trippas, Johanne R., Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 325–328. https://doi.org/10.1145/3020165.3022144

Tur, G. and R. De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, Wiley. https://doi.org/10.1002/9781119992691

Tür, Gökhan, Anoop Deoras, and Dilek Hakkani-Tür. 2013. Semantic parsing using word confusion networks with conditional random fields. In *INTERSPEECH*, pages 2579–2583. https://doi.org/10.21437/Interspeech.2013-580

Velikovich, Leonid, Ian Williams, Justin Scheiner, Petar Aleksic, Pedro Moreno, and Michael Riley. 2018. Semantic lattice processing in contextual automatic speech recognition for Google Assistant. In *Proceedings of Interspeech*, pages 2222–2226. https://doi.org/10.21437/Interspeech.2018-2453

Walker, Willie, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition.

Wang, Longshaokan, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos. 2020. Data augmentation for training dialog models robust to speech recognition errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 63–70. https://doi.org/10.18653/v1/2020.nlp4convai-1.8

Weng, Yue, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, and Gökhan Tür. 2020. Joint contextual modeling for ASR correction and language understanding. *CoRR*, abs/2002.00750. https://doi.org/10.1109/ICASSP40776.2020.9053213

You, Chenyu, Nuo Chen, Fenglin Liu, Dongchao Yang, Zhiyang Xu, and Yuexian Zou. 2021. Towards data distillation for end-to-end spoken conversational question answering. *arXiv preprint arXiv:2010.08923*. https://doi.org/10.21437/Interspeech.2021-120

Zenz, Gideon, Xuan Zhou, Enrico Minack, Wolf Siberski, and Wolfgang Nejdl. 2009. From keywords to semantic queries—Incremental query construction on the semantic web. *Journal of Web Semantics*, 7(3):166–176. https://doi.org/10.1016/j.websem.2009.07.005