Position Paper: Indirect Supervision for Dialog Systems in Unwritten Languages

Mark Hasegawa-Johnson, Najim Dehak and Odette Scharenborg

Abstract This paper considers the problem of creating spoken dialog systems for unwritten languages. A dialog system accepts speech as input, and generates speech as output; the user does not use writing to interact with the system. During training time, however, most dialog systems require written labels in order to train speech-to-text and text-to-speech system components. This paper proposes training the components of a dialog system in an unwritten language using auxiliary training signals that replace writing: non-native transcription, chat-alphabet transcription, images, and re-speaking.

Introduction: A dialog system is, essentially, a function $y = f(x)$ that maps from user utterances, x, to system responses, y.[1] The function f is trained to maximize some objective measure of the usefulness of the system response. The set of all possible user utterances is so huge, however, that most dialog systems simplify it in some way: some dialog systems reduce x from speech to text using an automatic speech recognizer, and others reduce it still further using some kind of semantic parse. In an unwritten language, converting x from speech to text is theoretically impossible, therefore other alternatives must be considered.

End-to-End Training of a Sequence-to-Sequence Dialog System: Suppose that there is a large training database of examples from which the dialog system can learn. For example, we might collect recorded speech from a call center: each recorded caller utterance, x, is paired with an example of the appropriate system response, y. By imitating these recorded data, the system can learn the mapping $y = f(x)$. Unfortunately, the training dataset would have to be prohibitively large. Barron demonstrated that the generalization error of a neural network is proportional to N/n, where N is the number of trainable parameters, and n is the number

Mark Hasegawa-Johnson
University of Illinois, e-mail: jhasegaw@illinois.edu

Najim Dehak
Johns Hopkins University, e-mail: ndehak3@jhu.edu

Odette Scharenborg
Delft University of Technology, e-mail: o.e.scharenborg@tudelft.nl

of training tokens[2]. A speech input signal requires about 1000 times as many trainable parameters as a text-based dialog manager, in the following sense: different tokens of the same word in text are exactly the same, but different tokens of the same word in speech are not. Auditory perceptual models [3] can be interpreted to allow about 1000 independent dimensions of variability among different tokens of the same spoken word. If a sequence-to-sequence system is capable of learning to map text inputs to text outputs using a training dataset with N examples, this reasoning suggests that a similar system could learn to map from speech inputs to speech outputs using $1000N$ examples. End-to-end training of a spoken dialog system is not practical if it increases the training data requirement 1000-fold.

Nonstandard Transcription: In an unwritten language, standardized transcription is impossible, but non-standard transcription is still possible. For example, we can hire people who don't speak the language, and ask them to write down what they hear, as if it were nonsense speech in their own language.[4] Although non-native transcriptions are far noisier than native transcriptions, they are also far cheaper.[5] It is also possible to ask native speakers of a language to invent an orthography as they write. Elmahdy found that native speakers of Egyptian Arabic were able to transcribe their own language faster and more easily using an Arabic chat alphabet (ACA), instead of using the Standard Arabic (SA) alphabet.[6]

Images: A more unusual training signal is provided by image2speech and, possibly, speech2image conversion. Harwath and Glass [7] showed that it's possible to train a neural net to accept speech inputs, and look up the corresponding image in a large image database, or vice versa. Hasegawa-Johnson et al. [8] demonstrated that it's possible to train a sequence-to-sequence neural net to generate the spoken description of an image. An image2speech training corpus can be acquired in an unwritten language by simply showing photographs to a native informant, one after the other, and asking him or her to describe each photograph out loud. The image2speech and speech2image neural nets pre-trained from these data might be used to initialize the components of a spoken dialog system.

Re-speaking: A certain amount of training supervision can be acquired by asking native informants to re-speak one another's utterances. Re-speaking has been used to generate closed captioning for real-time television broadcast [9]: the broadcast audio is played to the headphones of a golden speaker, who, rather than typing what he hears, simply repeats the same words into a microphone. Similar methods could be used to acquire speech synthesis training material in un unwritten language: recordings of a large variety of native informants could be played to a golden speaker, who would then repeat exactly the same word sequences.

Conclusion: A dialog system in an unwritten language would need to be trained without text transcriptions. Possible alternatives to text transcription include non-native transcription, chat alphabet transcription, images, and re-speaking.

Acknowledgements The ideas in this paper were developed in response to two ongoing research projects: (1) NSF 1550145, a grant to the first author, and (2) the 2017 Jelinek Speech and Language Technology workshop, funded by grants from Alphabet, Amazon, Apple, Facebook, and Microsoft.

References

[1] H. Niemann, M. Lang, and G. Sagerer, Eds., *Recent Advances in Speech Understanding Systems*. Springer-Verlag, 1988.

[2] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.

[3] H. Fletcher, "A space-time pattern theory of hearing," *J. Acoust. Soc. Am.*, vol. 1, no. 3, pp. 311–343, 1930.

[4] P. Jyothi and M. Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proc. AAAI*, 2015.

[5] M. Hasegawa-Johnson, P. Jyothi, W. Chen, and V. H. Do, "Mismatched crowdsourcing: Mining latent skills to acquire speech transcriptions," in *Proceedings of Asilomar*, 2017.

[6] M. Elmahdy, R. Gruhn, S. Abdennadher, and W. Minker, "Rapid phonetic transcription using everyday life natural chat alphabet orthography for dialectal arabic speech recognition," in *Proc. ICASSP*, 2011, pp. 4936–4939.

[7] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop*, Scottsdale, Arizona, USA, 2015, pp. 237–244.

[8] M. Hasegawa-Johnson, A. Black, L. Ondel, O. Scharenborg, and F. Ciannella, "Image2speech: Automatically generating audio descriptions of images," *Journal of International Science and General Applications*, vol. 1, no. 1, pp. 19–27, 2018, ISSN: 2351-8715.

[9] A. Pražák, Z. Loos, J. Trmal, and J. V. Psutka, "Novel approach to live captioning through re-speaking: Tailoring speech recognition to re-speaker's needs," in *Proc. Interspeech*, 2012, pp. 1372–1375.