March 2021

Intent Prediction Based On Contextual Factors For Better Automatic Speech Recognition

Diamantino Caseiro

Zelin Wu

Petar Aleksic

Era Jain

Intent Prediction Based On Contextual Factors For Better Automatic Speech Recognition

ABSTRACT

Automatic speech recognition (ASR) machine learning models are used to recognize spoken commands or queries from users. End-to-end ASR models, which directly map a sequence of input acoustic features into a sequence of words, greatly simplify ASR system building and maintenance. This disclosure describes techniques to improve the performance of end-to-end ASR models by providing predicted user intents as additional inputs. Intent prediction vectors or intent embedding is generated based on user-permitted contextual features using a trained intent prediction network (IPN). The IPN can be trained independently from the ASR model or jointly with the ASR model. Training of the IPN can be performed based on training data that includes user-permitted contextual features, even when such data does not include speech data. The IPN can be retrained when the available contextual feature set changes.

KEYWORDS

- Intent prediction
- Query context
- Voice query
- Spoken query
- Smart speaker
- Virtual assistant
- Automatic Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Recurrent Neural Network Transducer (RNN-T)

BACKGROUND

Automatic speech recognition (ASR) machine learning models are used to recognize spoken commands or queries from users in hardware products such as smartphones, smart speakers/displays, or other appliances, as well as applications that enable speech interaction, e.g.,

virtual assistant applications. Conventional ASR systems typically include acoustic, pronunciation, and language model components, which are trained independently. End-to-end ASR models fold the acoustic, pronunciation, and language models of a conventional speech recognition model into a single neural network and optimize them jointly. End-to-end ASR models greatly simplify building and maintaining an ASR system. Variants of end-to-end ASR models use algorithms such as connectionist temporal classification (CTC), attention based models, and recurrent neural network transducer (RNN-T)

Current ASR systems do not take into account the intent of the user at the time the user issues a spoken query or command. If the user permits, performance of end-to-end ASR models can be improved by taking into account user-permitted contextual signals as additional inputs. Some examples of contextual signals that can be used include the user's contact list, location, the system dialog state, time of day etc. Contextual signals, when modelled correctly, can help narrow the search space of the next query or command that the user is likely to issue. Coarse level contextual signals such as the dialog state can be used directly in an end-to-end ASR system as additional inputs using a multistate Recurrent Neural Network Transducer (multistate RNN-T). However, this approach has some drawbacks such as:

- Training the multistate RNN-T with contextual signals requires the availability of speech data with associated contextual information. Often, in existing training databases that include speech data, such contextual information is incomplete or unavailable. Also, obtaining such training data is difficult since storing contextual information in association with speech data can negatively impact privacy and may therefore not be permissible.

- A trained multistate RNN-T is dependent on the set of contextual factors that are used for training. If the contextual information changes, for example, when new factors are added or removed, the Multistate RNN-T needs to be retrained.

DESCRIPTION

This disclosure describes techniques to improve the performance of machine learning based end-to-end automatic speech recognition by using predicted user intents as additional inputs. With user permission, contextual signals at the time of issuing a spoken query or command are obtained and are utilized to predict the likely intents of the user. The intent prediction is performed using an intent prediction network and predicted intents provided as an additional input to the speech recognition model.

As used herein, an intent is a representation of a user's need, expressed directly or indirectly in a query. The intent predictions or embeddings are generated using an intent prediction network (IPN), that takes as input user-permitted contextual information and predicts likely intents for the utterance. The IPN can be a simple feedforward neural network. If the user permits the use of intents from prior utterances as contextual signals, the IPN can be implemented as a Recurrent Neural Network such as LSTM, GRU, etc. To avoid the complexity incurred by the recurrence, the IPN can be a time-aware bag-of-words model. The IPN is trained to generate as output an intent prediction vector that includes the conditional probability of one or more intents given the contextual information provided as input, denoted as P (intent | context). Alternatively, the output of the IPN can also be in the form of an intent embedding.

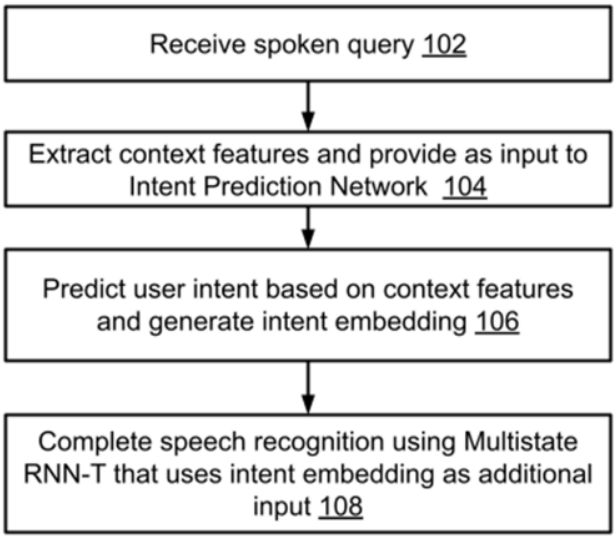

Fig. 1: Process for end-to-end ASR using intent prediction

Fig. 1 illustrates an example process for improvement in end-to-end automatic speech recognition using intent prediction, per techniques of this disclosure. A spoken query is received from the user (102). User-permitted context features such as system dialog state, time-of-day, locale, surface, previous queries, etc. are provided as inputs to a trained intent prediction network (104). A prediction of the user intent is obtained from the IPN and used to generate an intent embedding (106). The end-to-end speech recognition task is completed using a Multistate RNN-T that takes the spoken query and the intent embedding generated by the Intent Prediction Network as an additional input (108). While Fig. 1 shows the blocks 102-108 sequentially, it will be understood that intent prediction (104-106) can be performed prior to receiving the spoken query, or at the same time as receipt of the query (102).

The use of a separate IPN to obtain the intent embedding or intent prediction vector that is then provided as input to a Multistate RNN-T, rather than raw context features is advantageous for reasons such as:

- The IPN can be trained on training data that includes contextual features, but no speech data. For example, with user permission, the IPN can be trained from logs of queries to a virtual assistant that include transcripts and/or contextual factors, and the intent. If speech data with associated contextual features are available, such data can also be used to train the IPN.
- The IPN can be a much simpler and lightweight neural network compared to the multistate RNN-T. This allows fast iterations for improvements, fine tuning, and exploration of various available contextual features without having to retrain the multistate RNN-T.
- When the available contextual features change, or more data becomes available, the IPN can be retrained without having to retrain the multistate RNN-T.
- When speech data with associated contextual features are available, such data can be used to jointly finetune the IPN for the ASR Task. For this, a composite network can be created by connecting a pre-trained IPN to the multistate RNN-T. Then, the speech data can be used as input to the composite network, and the contextual features can be used as input provided to the IPN. To finetune the IPN, gradients (feedback) computed for the ASR task can be back propagated to the IPN to update its parameters, while parameters of the multistate RNN-T can be kept frozen.
- A language agnostic IPN can be trained if speech transcriptions are not used for training. Such models can be trained using a mix of data from different languages, and can improve speech recognition for languages where little or no data is available for training.

The described techniques can be used in any context where spoken input is received. For example, interpretation of queries to a hardware device such as a smart speaker or other smart

appliance, or to a virtual assistant provided via a smartphone or other device can, with user permission, incorporate intent prediction to improve recognition of the spoken input.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's spoken input such as queries or commands, whether a query was successful or not, contextual factors, a user's preferences, or a user's location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques to improve the performance of end-to-end ASR models by providing predicted user intents as additional inputs. Intent prediction vectors or intent embedding is generated based on user-permitted contextual features using a trained intent prediction network (IPN). The IPN can be trained independently from the ASR model or jointly with the ASR model. Training of the IPN can be performed based on training data that includes user-permitted contextual features, even when such data does not include speech data. The IPN can be retrained when the available contextual feature set changes.

REFERENCES

1. Li, Bo, Shuo-yiin Chang, Tara N. Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, and Yonghui Wu. "Towards fast and accurate streaming end-to-end asr." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6069-6073. IEEE, 2020.
2. Hu, Ke, Antoine Jean Bruguier, Tara N. Sainath, Rohit Prakash Prabhavalkar, and Golan Pundak. "Phoneme-based contextualization for cross-lingual speech recognition in end-to-end models." U.S. Patent Application 16/861,190, filed April 28, 2020.
3. Wu, Zelin, Bo Li, Yu Zhang, Petar S. Aleksic, and Tara N. Sainath. "Multistate Encoding with End-To-End Speech RNN Transducer Network." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7819-7823. IEEE, 2020.