View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Technical Disclosure Common

Technical Disclosure Commons

Defensive Publications Series

October 2020

Improving Automatic Speech Recognition by Co-embedding Voice Queries and Voice Query Refinements

Sukhdeep Sodhi

Ankit Kumar

Sarvjeet Singh

Tameen Khan

Ajit Apte

See next page for additional authors

Follow this and additional works at: https://www.tdcommons.org/dpubs_series

Recommended Citation

Sodhi, Sukhdeep; Kumar, Ankit; Singh, Sarvjeet; Khan, Tameen; Apte, Ajit; and Jeje, Ayooluwakunmi, "Improving Automatic Speech Recognition by Co-embedding Voice Queries and Voice Query Refinements", Technical Disclosure Commons, (October 12, 2020)
https://www.tdcommons.org/dpubs_series/3666

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.

Inventor(s)

Sukhdeep Sodhi, Ankit Kumar, Sarvjeet Singh, Tameen Khan, Ajit Apte, and Ayooluwakunmi Jeje

This article is available at Technical Disclosure Commons: https://www.tdcommons.org/dpubs_series/3666

Improving Automatic Speech Recognition by Co-embedding Voice Queries and Voice Query Refinements

ABSTRACT

Automatic speech recognition (ASR) models are used to recognize voice commands or queries from users in hardware products such as smartphones, smart speakers/displays, as well as applications that enable speech interaction, e.g., virtual assistant applications. However, the query abandonment rates for voice queries continue to be much higher than text queries which is often due to incorrect interpretation of the spoken query. This disclosure describes techniques to improve the performance of recognition of spoken queries by combining user specific phonetic variations and session specific contextual signals, obtained with specific user permission.

KEYWORDS

- Automatic Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Phonetic refinement
- Query rewriting
- Collaborative filtering
- Deep Neural Network (DNN)
- Virtual assistant
- Smart speaker
- Voice query
- Spoken query
- Speech transcription

BACKGROUND

Automatic speech recognition (ASR) models are used to recognize voice commands or queries from users in hardware products such as smartphones, smart speakers/displays, as well as applications that enable speech interaction, e.g., virtual assistant applications. However, the query abandonment rates for voice queries continue to be much higher than text queries which is often due to incorrect interpretation of the spoken query. For example, the voice query 'Play tutorial of Hidden Markov Model' might get recognized as 'Play tutorial of hiccup half a poodle,' resulting in the user abandoning the voice query and switching to text input.

Automatic speech recognition is a complex problem that requires correct processing of the acoustic and semantic signals from the voice input. A typical ASR system includes an acoustic model that transforms the audio signal into a feature vector, and a language model that uses the features to obtain a meaningful utterance by referring to a language and context specific vocabulary. The acoustic model is designed to account for the variation in pronunciation across users and languages, while the language model is designed to account for the semantic differences across contexts within a language.

Prior queries made by a user within a session are useful in determining the context within which the subsequent queries are likely to be made. However, such information is hard to incorporate in conventional ASR models.

DESCRIPTION

This disclosure describes techniques to improve the performance of recognition of spoken queries by combining user specific phonetic variations and session specific contextual signals, obtained with specific user permission. For example, contextual signals can include the successful and abandoned queries earlier in the session. The solution described in this disclosure employs a two tower deep neural network architecture, which is a generalization of collaborative filtering.

One of the DNN towers represents the user, while the other tower represents the current session. The user tower includes successful queries issued by the user in previous sessions and multiple phonetic refinements of these queries. The query tower includes data from the queries issued by the user in the current session. For all queries in the current session, the query tower captures whether the query was successful or abandoned. For an abandoned query, the query tower also captures the successful query that follows it.

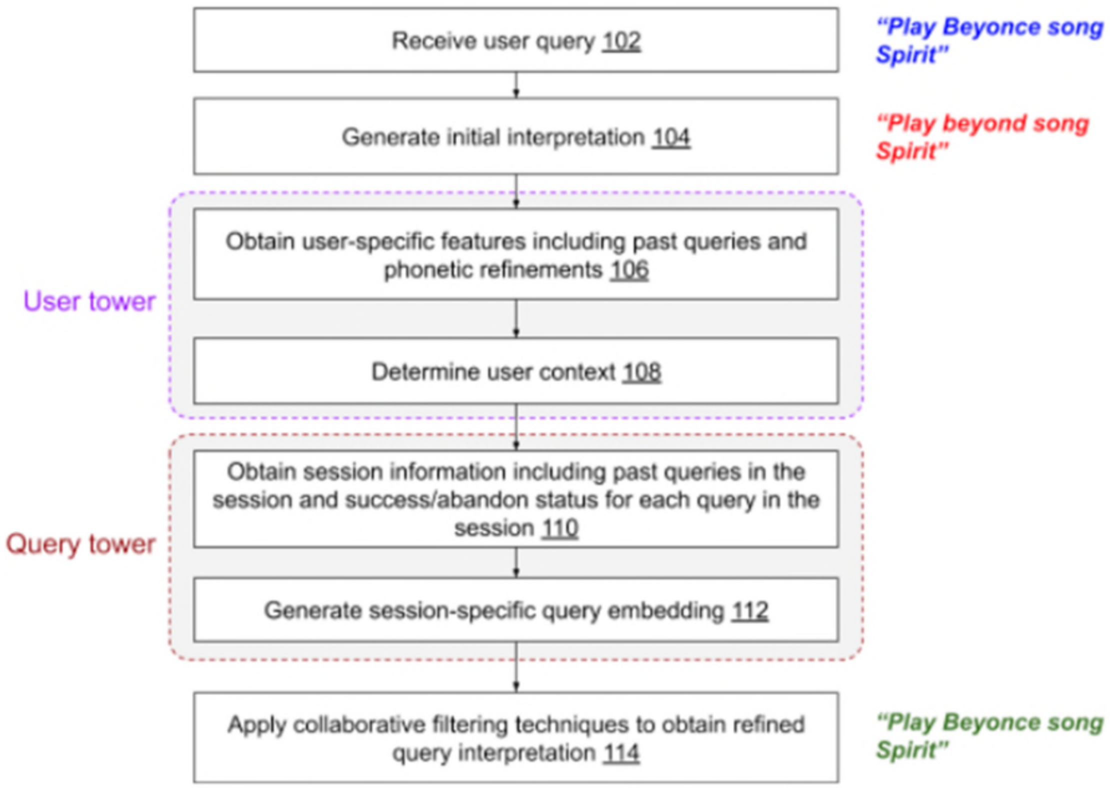
Receive user query 102
"Play Beyonce song Spirit"
Generate initial interpretation 104
"Play beyond song Spirit"
User tower
Obtain user-specific features including past queries and phonetic refinements 106
Determine user context 108
Query tower
Obtain session information including past queries in the session and success/abandon status for each query in the session 110
Generate session-specific query embedding 112
Apply collaborative filtering techniques to obtain refined query interpretation 114
"Play Beyonce song Spirit"

Fig. 1: Example process to correct query interpretation based on phonetic replacement and past ASR data

Fig. 1 illustrates an example process for improvement in automatic speech recognition, per techniques of this disclosure. User permission is obtained to utilize data such as past queries by the user for the specific purpose of query refinement. If the user denies permission, or restricts permission to certain data, only such data as permitted by the user are used for query interpretation.

When a new user query is received (102), an initial interpretation of the audio signal (104) is generated. In the example shown in Fig. 1, the user's query "Play Beyonce song Spirit" (shown in blue) is incorrectly interpreted as "Play beyond song Spirit" (shown in red). The initial interpretation is then further analyzed using a two tower deep neural network (DNN) to generate a refined query interpretation.

User specific features including the user's past queries and their phonetic refinement are obtained (106) from the user tower, to generate the user context (108). Session specific information including prior abandoned queries in the session are obtained (110) from the query tower to generate the session specific query embedding (112).

Using the user context and session specific query embedding, collaborative filtering techniques (e.g., weighted average least squares - WALS) are applied (114) are applied to generate the refined query interpretation. In the illustrated example, the refined query interpretation (shown in green) matches the original user query. Nearest neighbor refinement may be obtained based on past queries in the session and used to rewrite the query. Per the described techniques, the context signals and query refinements are in the same embedding space.

Training of the model is performed using query pairs that include an abandoned voice query that is followed by a successful query. Such query pairs of training data are utilized with specific user permission.

The described techniques for query interpretation can be utilized in any application, e.g., a virtual assistant activated by voice, a media application that accepts voice input, that is provided on a device such as a computer, smart speaker, smart display/TV, smartphone, or other smart appliance. Reduction in speech recognition errors can help improve the user experience and increase user engagement for such applications.

Further to the descriptions above, a user is 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, a user's preferences), 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. 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 recognition of spoken queries by combining user specific phonetic variations and session specific contextual signals, obtained with specific user permission.

REFERENCES

1. Sarma, Arup, and David D. Palmer. "Context-based speech recognition error detection and correction." In *Proceedings of HLT-NAACL 2004: Short Papers,* pp. 85-88. 2004.
2. Skobeltsyn, Gleb, Evgeny A. Cherepanov, and Behshad Behzadi. "Query rewrite corrections." U.S. Patent 9,514,743, issued December 6, 2016.
3. Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).