January 2022

Automatic Contextual Adjustments for Interpretation of Spoken Queries

Agoston Weisz

Alessandro Agostini

Automatic Contextual Adjustments for Interpretation of Spoken Queries

ABSTRACT

Automatic speech recognition techniques implemented in a virtual assistant or other application can sometimes fail to correctly transcribe a user query, even while utilizing user-permitted contextual information. Query recognition failures lead to misunderstanding of user intent. While increasing the strength of contextual biasing of speech recognition can fix the problem of misrecognition, too strong a bias can hurt queries that don't benefit from such adjustments. This disclosure describes techniques that, upon failure to find suitable transcription, intent, or response, increase contextual bias provided to the ASR and re-run speech recognition to obtain a better transcription of user speech and determination of user intent. Effectively, two or more speech recognition results are obtained, each at differing contextual bias strengths. The result with the best intent is selected and a corresponding response is provided to the query.

KEYWORDS

- Automatic speech recognition (ASR)
- ASR biasing
- Speech biasing
- Voice query
- Recognition context
- User intent
- Virtual assistant
- Natural language processing
- Smart speaker

BACKGROUND

Fig. 1 illustrates an example of intent understanding, implemented, for example, in a virtual assistant. A command spoken by a user (100) is processed by an automatic speech recognition (ASR) module (102), which can render speech to text. The text is analyzed by a natural language understanding (NLU) module (104), which determines user intent. With

specific user permission, the ASR module obtains user context from a contextual bias module (106).

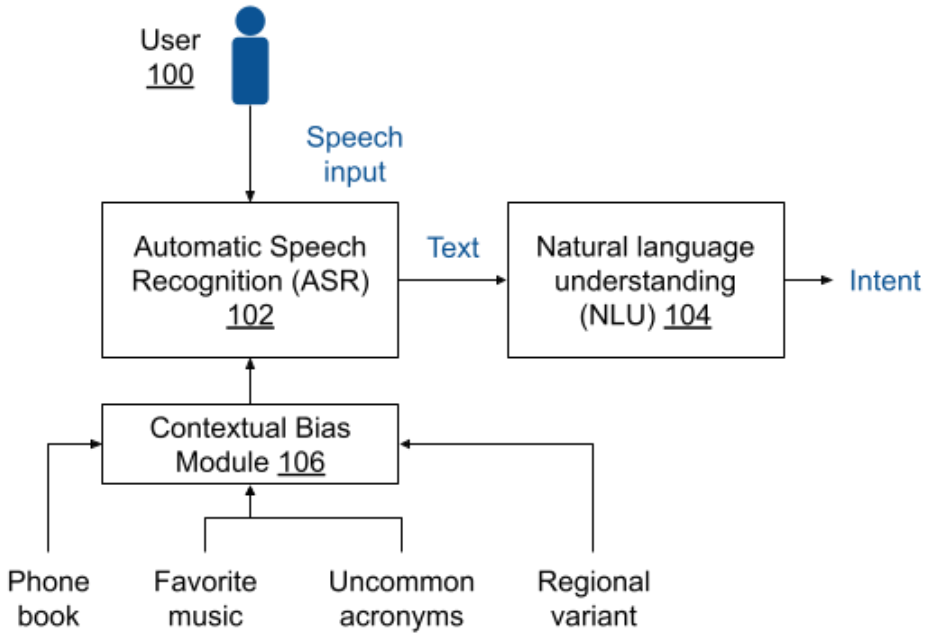

Fig. 1: Intent understanding

The bias module can obtain context from a variety of user-permitted data sources, e.g., the user's phone book, the user's music selections, the user's regional language variants, etc. The bias module can assist in accurate ASR transcription when the user's utterances aren't transcribable into words or phrases of standard language. For example, a user may routinely use a regional variant of English in which the word "creps" is used for shoes. "Creps" being non-standard English, ASR can mistranscribe "creps" as "crêpes" (pancakes), leading to a failure in intent understanding. Knowing the user's context, biasing the ASR to interpret "creps" towards "shoes" can improve intent understanding.

Despite the presence of the biasing module and its access to the user's context, it is sometimes the case that automatic speech recognition misinterprets the user's query. The ASR does get other parts of the query right, but failure to recognize words in their context can lead to

intent misunderstanding. Increasing biasing strength can fix the problem of misrecognition, but too strong a bias can hurt queries that don't benefit from bias (anti-context). It is difficult to find the right biasing strength that achieves a good middle ground.

DESCRIPTION

This disclosure describes techniques that, upon an ASR punt (failure to find suitable transcription, intent, or response), increase the bias to the ASR module and re-run speech recognition to obtain a contextually appropriate transcription of user speech or finding of user intent. Effectively, two or more speech recognition results are obtained, each at differing contextual bias strengths. The result with the best intent is selected.

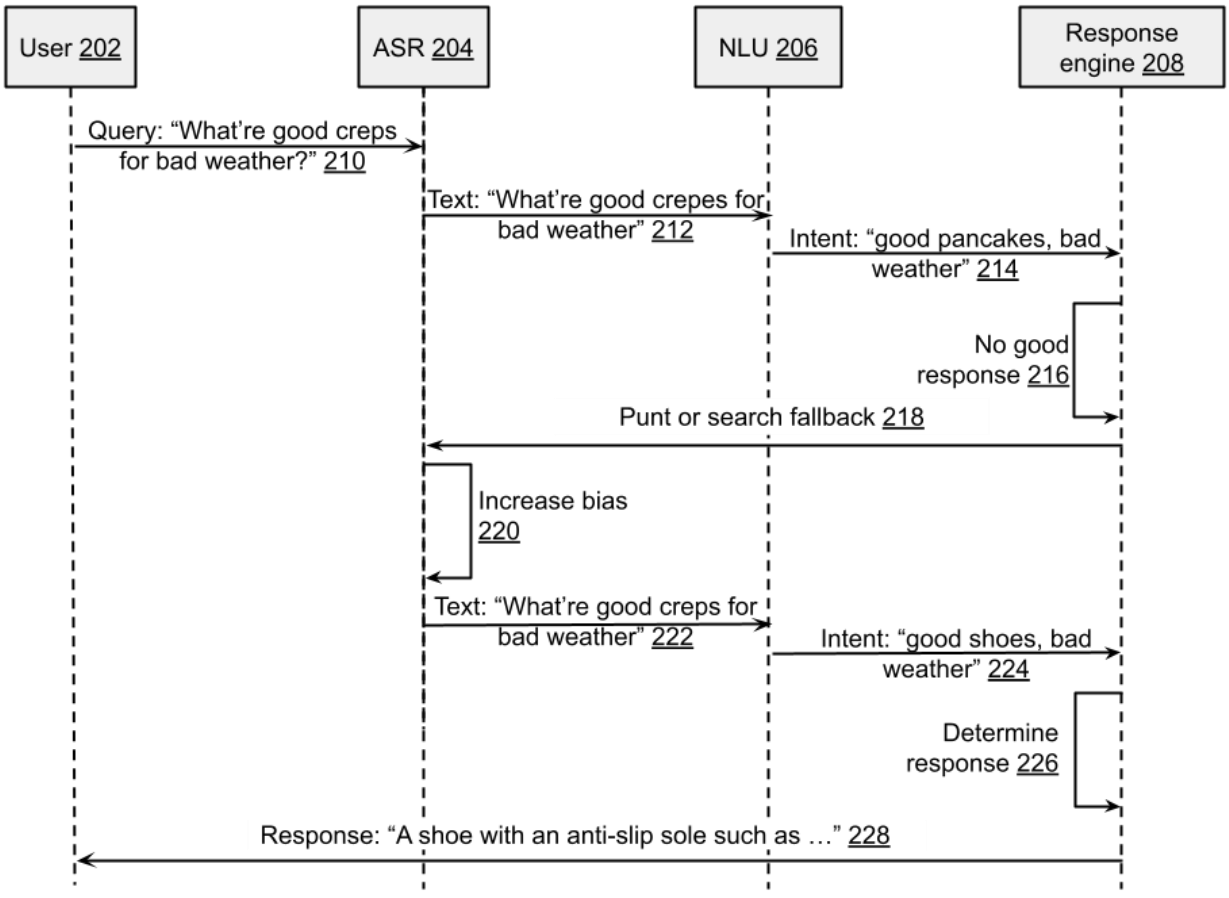

Fig. 2: Stronger biasing upon ASR punt or search fallback

Fig. 2 illustrates an example of stronger biasing in interpreting a user query to a virtual assistant upon ASR punt (inability to understand the query) or search fallback (where the best available response to the query is a search results page, rather than a direct response from the virtual assistant specific). A user (202) utters a query (210) "what're good creps for bad weather?" The ASR (204), under standard biasing, transcribes the query (212) to "what're good *crêpes* for bad weather?" The NLU module (206) obtains an intent (214) "good pancakes, bad weather," for which the response engine (208) finds no high-scoring response (216).

Having punted (218), e.g., failed in transcription, intent discovery, or search response, the contextual bias to the ASR is increased (220), resulting in a transcription (222) "what're good creps for bad weather?" The response engine receives a user intent (224) "good shoes, bad weather" to which it can associate a high-scoring intent (226). An appropriate response (228) "A shoe with an anti-slip sole …" is delivered to the user.

In this manner, the user query is recognized correctly by optimizing contextual bias in response to punting or search fallback from an initial interpretation of the query, without hurting ASR quality or incurring ASR loss. Alternatively, to avoid the possibility of over-biasing, the user can simply be informed of a punt or search fallback without further tuning of contextual bias, referred to as safe ground.

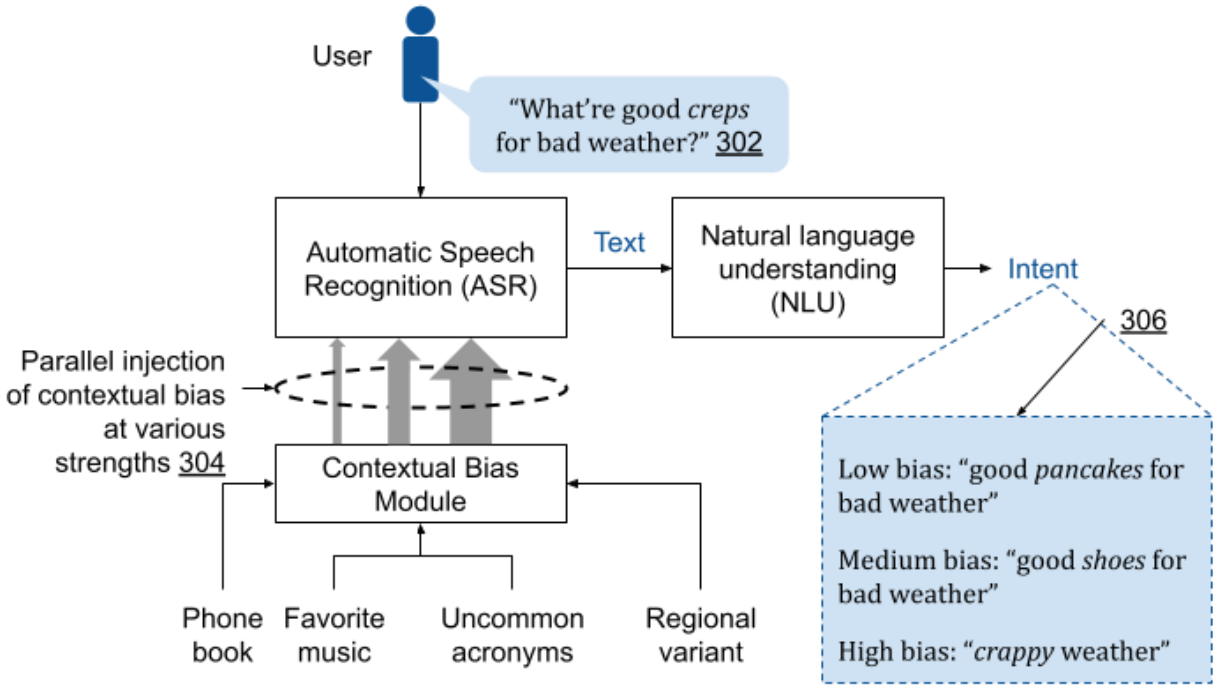

Fig. 3: Parallel contextual biasing

Alternatively, as illustrated in Fig. 3, the ASR unit can be parallelly biased with contextual biases of differing strengths. Upon a user query (302) "what're good creps for bad weather?" the ASR module is parallelly injected with contextual bias of various strengths (304, indicated by arrows of differing thicknesses). Intents corresponding to the differing biasing strengths are obtained simultaneously (306) and the highest-scoring intent is chosen to provide a response to the user. Parallel injection of contextual bias, as described herein, can reduce the latency of response.

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 the collection of user information (e.g., information about a user's query, data indicative of a user's context, social network, social actions or activities, profession, a user's preferences, or a user's current 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 that, upon failure to find suitable transcription, intent, or response, increase contextual bias provided to the ASR and re-run speech recognition to obtain a better transcription of user speech and determination of user intent. Effectively, two or more speech recognition results are obtained, each at differing contextual bias strengths. The result with the best intent is selected and a corresponding response is provided to the query.

REFERENCES

[1] Aleksic, Petar, and Pedro J. Moreno Mengibar. "Dynamically biasing language models." U.S. Patent Application 14/525,826, filed Oct. 28, 2014.

[2] Ingmarsson, Carl-Anton. "Location-based voice query recognition." U.S. Patent Application 15/336,056, filed Oct. 27, 2016.

[3] Zhao, Ding, Bo Li, Ruoming Pang, Tara N. Sainath, David Rybach, Deepti Bhatia, and Zelin Wu. "Using Context Information With End-to-End Models for Speech Recognition." U.S. Patent Application 16/827,937, filed Mar. 24, 2020.