March 2022

Use of Language Models to Improve Automatic Speech Recognition

Pawel Janus

Agoston Weisz

Aurelien Boffy

Miroslaw Michalski

Use of Language Models to Improve Automatic Speech Recognition

ABSTRACT

Speech recognition enables users to interact with devices via their voice. However, errors in speech recognition during a user's interaction with such devices can be problematic and lead to a less than satisfactory user experience. This disclosure describes the use of language modeling to recover from automatic speech recognition (ASR) errors by identifying broken queries. The full natural language understanding (NLU) stack is executed to obtain a coherent, alternative, speech recognition. The alternative recognition (or query) runs in parallel to the original, misrecognized query. The potential actions triggered by the misrecognized and the NLU-augmented queries are compared to pick the query interpretation that is more likely to be correct.

KEYWORDS

- Automatic speech recognition (ASR)
- Language model
- Language representation model
- Speech biasing
- Virtual assistant
- Virtual assistant correction
- Bidirectional encoder representations from transformers (BERT)
- Multitask unified model (MUM)
- Natural language understanding (NLU)
- Masked language modeling (MLM)
- ASR confidence score

BACKGROUND

Speech recognition enables users to interact with multimodal or smart devices such as smartphones, tablets, virtual assistants, etc., via their voice, irrespective of whether the device has a screen or not, whether the device has a physical keyboard or virtual keyboard, etc. However, errors in speech recognition during a user's interaction with such devices can be problematic and lead to a less than satisfactory user experience.

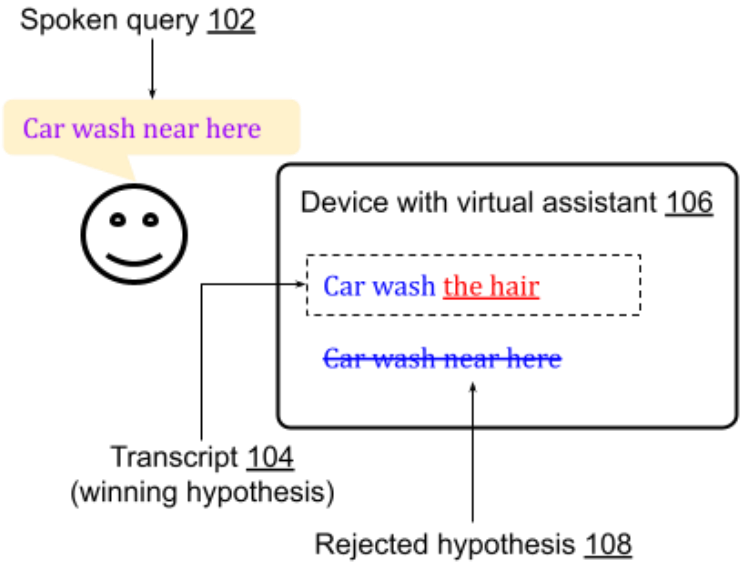

Fig. 1: Selection of the wrong hypothesis results in a speech misrecognition

As illustrated in Fig. 1, an automatic speech recognizer (ASR) is configured to transcribe the user's utterance as one of several hypotheses, each of which is a possible recognition of the issued query. The user issues a spoken query (102), 'car wash near here.' An ASR, e.g., implemented as part of a virtual assistant provided via a device (106), picks the hypothesis 'car wash the hair,' (104). However, although having the highest confidence score, this is a misrecognition. An alternative hypothesis produced by the ASR - 'car wash near here' (108), although correct, is rejected because of its lower score. As is often the case, the correct utterance

is one of the hypotheses; however, due to the interpretation of the captured acoustic signal, a wrong hypothesis (recognition) is picked as the winner.

DESCRIPTION

This disclosure describes a text-based signal that can help recover from ASR errors by identifying broken queries and by running a full natural language understanding (NLU) stack to achieve a coherent alternative recognition of the user's utterance. The alternative recognition (or query) runs in parallel to the original, misrecognized query. Actions triggered by the misrecognized and the NLU-augmented queries are compared to pick the better hypothesis.

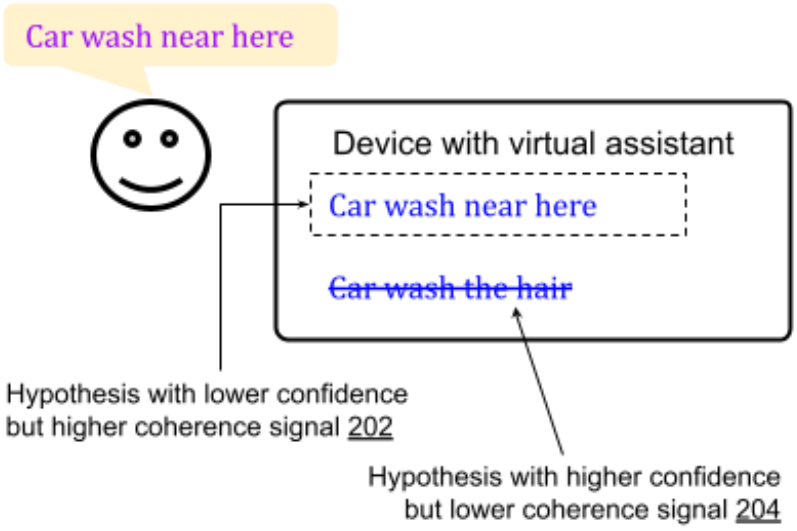

Fig. 2: Using NLU to pick a hypothesis of greater relevance to the user's query

Fig. 2 illustrates the use of NLU to pick a hypothesis that is likely more relevant to the user's utterance and has a greater chance of being correct. The hypothesis (204) with the higher raw confidence ('car wash the hair') is determined to be likely broken based on the lack of coherence, essentially the lack of meaning to the potential action triggered by the query. An alternative with lower raw confidence (202) has higher coherence based on the potential action

or answer to the query. Since the lower confidence hypothesis triggers a better feature when fetched with its answer, it is selected as the winning hypothesis and the corresponding response is provided, optionally with a transcript of the interpretation of the user's utterance.

As explained above, a coherence signal is associated with a query. The coherence signal is one of the factors, in addition to the confidence score generated by the ASR, that is used to pick the winning hypothesis. The hypothesis with the highest raw ASR confidence score can be replaced by a hypothesis with a lower raw ASR confidence score and a stronger coherence signal. This results in the query with a strong coherence signal being fulfilled for the user.

The coherence signal can be quantified using suitable language models such as bidirectional encoder representations from transformers (BERT), multitask unified model (MUM), etc., fine-tuned to classify queries as either coherent (valid) or incoherent (invalid). The language model accepts a query as input and outputs a coherence (validity) score. A query is valid when it is coherent, e.g., meaningful, and is something that a user would typically ask or say to a virtual assistant. A coherent query is different from a popular query because not all coherent queries have to be popular. Invalid queries are incoherent, e.g., broken, meaningless, unintelligible, or don't make sense in the context of a virtual assistant.

Query	Valid?	Reason
Turn on ceiling lights	Yes	Clear, self-contained query, typical of commands issued to virtual assistants. The user wants to turn something on, in this case, ceiling lights.
How to use Greece in my car	No	The term *Greece* doesn't make sense in the context of being used in a car. The user is more likely looking to use *grease* in a car.
Play rain sounds	Yes	Clear, self-contained query, typical of commands issued to virtual assistants. The user wants to listen to something, in this case, rain sounds.
99 corolla spare time	No	Difficult to say what *spare time* means in the context of the phrase '99 corolla,' which can be a reference to a car. More likely, the user is searching for a *spare tire*.
Next song	Yes	Clear, self-contained query, typical of commands issued to virtual assistants.
What is beedrills first fruit	No	The term *fruit* doesn't make sense in the context of Beedrills, a cartoon bee. More likely, the user is asking for the *first form*, not first fruit.

Table 1: Examples of valid and invalid user queries

Table 1 illustrates examples of valid and invalid user queries, with invalid ones attributable to misrecognition by the ASR. The language model can be fine-tuned with or without the MLM (masked language modeling) objective. The model can also be pre-trained with MLM on positive examples and later fine-tuned on a large number of labeled queries.

Fine-tuning or training the language model

Some example techniques of obtaining labels for the purposes of fine-tuning or training the language model include:

- *Human-rated ASR logs*: Assemble speech hypotheses in a template with instructions on how to rate them. The template includes at least the query in question and can also include contextual information and/or search results to enable accurate rating.

Are the following queries valid?

How do I wash virus germs off of fresh fruit?	[ ] Yes [ ] No [ ] Unsure
How do I watch virus germs off of fresh fruit?	[ ] Yes [ ] No [ ] Unsure

Fig. 3: An example of a ratings template

Fig. 3 illustrates an example of a ratings template, e.g., a set of queries provided to human raters who rate the validity of the query. The ratings by the human raters are used to train the language model to produce a coherence (or validity) signal.

- *Bootstrapping a coherence signal from the ASR*: The ASR logs, obtained with user permission, are sampled, and well-recognized queries (associated with high confidence) are labeled as examples of coherent queries. Corresponding candidates associated with lowest confidence are labeled as examples of incoherent queries. Labeled data thus obtained can be used to train the language model and to fix ASR errors. This approach relies on a highly accurate ASR being in place.
- *Obtain labeled examples from other logs or corpora*: Working (valid) queries are obtained from logs or other corpora permitted for such use and labeled as such. Invalid queries are obtained from valid queries by breaking them by randomly substituting words with alternatives, e.g., by masking and filling blanks with the second-highest scoring word from a query-BERT (or other) language model. Labeled examples of valid and invalid queries are used to train the model to produce a coherence (validity) signal.

Picking a hypothesis based on raw ASR confidence score and the coherence signal

Queries scored by coherence signal and raw ASR confidence scores can be ranked to pick the winning hypothesis using various techniques.

- *Re-ranking n-best hypotheses from the ASR based on the coherence signal*: The top-scoring hypothesis (which is based on coherence signal) is compared with the initial top recognition (which is based on raw ASR confidence score). If a hypothesis with a lower ASR confidence score nevertheless has a coherence signal greater than that of the initial top recognition (plus some threshold), the top initial recognition is replaced with the hypothesis with the greater coherence signal. The threshold can be fine-tuned based on the use case. Alternatively, the raw ASR confidence score can be combined with the coherence signal to arrive at a new, composite score for each hypothesis.
- *Applying policy filters*: Some example policy filters that can be applied (after the re-ranking) to constrain the picking of a winning hypothesis include:
    - Requiring the winning hypothesis to be of a minimum query and candidate length.
    - Using historical statistics to avoid rewriting (rejecting) queries that work.
    - Using contextual signals to avoid fixing queries that are semantically closer to the context than the considered alternative candidate, e.g., by comparing semantic distances.
    - Disabling particular feature transitions to boost quality by avoiding risky rewrites. For example, a calculator query may not be rewritten to another calculator query; more generally, queries within a vertical (finance, automobiles, etc.) may not be rewritten to another query within the same vertical.

The answer from the alternative recognition, e.g., the winning hypothesis based on raw ASR confidence score and the coherence signal, can be surfaced to the user using a separate, standalone ranker such as a conversation engine capable of assessing the quality of answers.

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 voice queries, 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 the use of language modeling to recover from automatic speech recognition (ASR) errors by identifying broken queries. The full natural language understanding (NLU) stack is executed to obtain a coherent, alternative, speech recognition. The alternative recognition (or query) runs in parallel to the original, misrecognized query. The potential actions triggered by the misrecognized and the NLU-augmented queries are compared to pick the query interpretation that is more likely to be correct.

REFERENCES

[1] Ramaswamy, Swaroop; Breiner, Theresa; Pisarev, Igor; Zivkovic, Dan; Chen, Mingqing; Mathews, Rajiv; and McConnaughey, Lara, "Personalizing Speech Recognition Based on User-entered Text", Technical Disclosure Commons, (February 08, 2022)

https://www.tdcommons.org/dpubs_series/4887