What's next

R&D

Improving named entity speech recognition accuracy

The latest generation of Nuance's deep learning technology for speech recognition features a novel algorithm which nearly halves the baseline error rate on named entities

Jesus Andres Ferrer
Posted September 1, 2021

Speech is the most natural way for us to interact with each other and machines. Automatic Speech Recognition (ASR) technology transforms spoken words into text and is a critical component powering many conversational AI solutions. Examples include Dragon Ambient eXperience (DAX), which has redefined the way doctors interact with patients allowing them to focus on the patient and doctor conversation, and the Nuance Mix platform which empowers

organizations to create advanced enterprise-grade conversational experiences.

The latest generation of speech recognizers is based on deep learning which concurrently learns to extract relevant features from data and utilize them for prediction via the composition of relatively simple building blocks.

Figure 1: Dragon Ambient eXperience (DAX) demo. Patient names are specific to an office encounter.

Conventional ASR systems are engineered with independent specialized deep learning components. Specifically they consist of an acoustic model which recognizes spoken units (such as phonemes), a language model (LM) that approximates the target distribution of written words and a pronunciation model which provides the mapping between written and spoken units. More recently, end-2-end (E2E) ASR systems, which directly map input audio to text, have come to the fore. E2E systems implicitly integrate the acoustic, pronunciation and language models into a single deep neural network (DNN) allowing for direct, joint optimization of the ASR objective and yielding improvements in accuracy, latency and complexity. Though recent techniques aim to reduce labeling costs, these data hungry DNNs require a large amount of transcribed audio to learn an accurate and robust system. Infrequent words and phrases will not be well represented in the training data and present a particular difficulty for these systems.

An important subset of those infrequent phrases are ***named entities*** such as doctor and

patient names, account names and locations. Individual named entities are not typically well represented in our training data. Moreover, we cannot even wish to accurately model these named entities solely from the training distribution since in many cases the deployment distribution for these phrases dynamically shifts. Consequently, our systems need to process external information regarding the expected entity names so as to gracefully adapt to the new distributions, for instance, specific appointment patient names or an individual user's payee list.

Conventional systems used class-based LMs to dynamically inject entities into a weighted finite state transducer allowing the system to recognize a given list of named entities, regardless of the entity value distribution observed in training. This is conceptually equivalent to having a placeholder for the named entities and then adding the list of possible values that can be recognized via this placeholder. On the other hand, the integrated elegance of E2E systems concurrently presents a challenge: how best to dynamically adapt to new word sequence distributions despite the language model reflecting the training data being internal to the single network? Recently, contextual shallow fusion was proposed as a way to bias the named entity distribution for E2E systems but without dealing with their implicitly embedded language model distribution. There have also been several methods investigated to leverage text-only data via external language models (LMs) including while compensating for the E2E system's internal LM. These approaches are however focused on biasing the language model distribution as a whole to new domains or to exploit abundant text data without corresponding audio.

In the paper titled "***Contextual Density Ratio for Language Model Biasing of Sequence to Sequence ASR Systems*** we demonstrate improved recognition accuracy of named entities by a newly proposed algorithm called **contextual density ratio.** We start from an E2E system trained on 1500 hours of labeled doctor and patient conversations. During training, we enclose all doctor and patient names, which we use as named entities for the study, between begin-name and end-name tags.

Figure 2: Contextual Density Ratio (and contextual shallow fusion) enclose named entities, in this case Mozart, between begin-entity and end-entity tags during training.

The E2E system learns via these tags when to expect these named entities based on the surrounding contextual information. A simple example is given by the prefixes *Mr., Ms, Mrs.* which typically precede a surname; however, the DNN also learns from more subtle contexts like ‘how is your knee doing lately, **<name/>** ?’

Figure 2: Scoring process for a given hypothesis for the baseline E2E and Contextual Shallow Fusion systems and our newly proposed Contextual Density Ratio. Note that the static contextual internal Language Model (depicted with a small purple circle embedded in the E2E scores) is not compensated for by baselines.

After training, the system recognizes the entities by adding two additional scores to each hypothesized transcription when named entity tags are predicted. The first component is a biasing LM trained on the doctor and patient name list (available from a scheduling feed), so as to dynamically adapt to the specific named entity distribution of the conversation, similar to contextual shallow fusion and in contrast to the E2E baseline. The second component, which is subtracted, is an internal language model (iLM) estimate of the implicit language model distribution the E2E system has learned from the entities appearing in the training data.

We compared our proposed approach to standard E2E and contextual shallow fusion baselines. In the lab experiment, our proposed Contextual Density Ratio obtained large relative name recognition accuracy improvements of 46.5 % with respect to the E2E system and of 22.1 % with respect to a strong contextual shallow fusion baseline. Moreover, the proposed approach does not degrade general recognition performance.

Finally, Contextual Density Ratio was found to be robust to simulated noise that might be encountered in a deployment. We tested the robustness by augmenting conversation name lists with (a) random names and (b) names similar to those in the conversation.

Figure 3: Word error rate computed within the conversation names. The lower the better.

In conclusion, our results push forward named entity recognition accuracy for E2E systems

showing large improvements without degrading global recognition accuracy and being robust to noise.

Dario Albesano contributed equally to this research. Paul Vozila and Puming Zhan contributed to the paper and this blog post. The paper will be presented in September at the Interspeech 2021 conference.

Tags: ASR, Contextual Language Model, Contextual Shallow fusion, DAX, deep-learning, Density Ratio, end-to-end, Shallow fusion, Speech Recognition

About Jesus Andres Ferrer

Jesus Andres Ferrer is a Senior Principal Research Scientist at Nuance.

View all posts by Jesus Andres Ferrer