Thomas Schaaf
3M | Health Information System / M*Modal
7514 Thomas Blvd.
Pittsburgh, PA 15208
tschaaf@mmm.com

Pittsburgh, PA 2022/11/27

Review of the doctoral thesis "OUT-OF-VOCABULARY WORDS DETECTION AND RECOVERY" by Ekaterina Egorova

Dear Committee Members:

Ms. Egorova's research addresses the challenge that Out-Of-Vocabulary (OOV) words pose to Automatic Speech Recognition (ASR) systems. With the introduction of neural network-based end-to-end systems primarily using sub-word units it might appear that the OOV issue no longer exists. However, such ASR systems significantly underperform on words that have not been seen during training, can output non-sensical words, and in real-world applications, often have the disadvantage that the vocabulary cannot be easily extended or customized if a domain shift occurs. Therefore, research in this area is essential and relevant for the speech community.

The thesis will be summarized and discussed by individual sections, followed by highlighting Ms. Egorova contributions to the research field, some remarks, my recommendation, and a few questions for the candidate.

Chapter 1 introduces the challenge of dealing with OOV words and defines an OOV word as a word not seen during training. It further explains the different metrics used in the thesis, which are based on Word Error Rate (WER) with and without mapping the reference words to OOV symbols, and Precision and Recall based on the overlap of the predicted OOV time segment with a reference OOV time range. It defines that the goal of the thesis is not only to detect but also to recover OOV words.

Chapter 2 provides an overview of related work and the models that the research is built upon.

Chapter 3 introduces LibriSpeech as data set used for experimentation. It motivates how the words for reoccurring OOV detection and recovery are selected using a creative idea based on the property that LibriSpeech contains books from many epochs by identifying "archaic" words that have fallen out of usage over time. About 1000 words have been identified this way, some of which can be quite frequent in the corpus. This elegant experiment design simulates the appearance of new words in a language. It might be helpful to note that this results in OOV words that are more diverse concerning their semantic categories, which are often names of some sort in many domains.

Chapter 4 explains Weighted Final State Transducers (WFST) operations and introduces the semirings used in the hybrid decoding approach and for the novel method of extracting OOV candidates from lattices. The chapter includes a detailed description of the baseline ASR system and how it is extended by integrating a phoneme loop that preserves the language model context on the word level. The extended system generates hybrid lattices that contain words and phonemes with special tokens indicating where an OOV word might start and end. A novel method of extracting OOV candidates from lattices using a WFST approach is introduced, and different methods to cluster OOV candidates to identify reoccurring OOVs and to derive better phoneme descriptions by combining the probabilities over the distribution of pronunciations from OOVs in the same cluster. The final OOV recovery was made using a phoneme-to-grapheme model on the best pronunciation of clusters containing at least two predicted OOV occurrences. Experiments explore how newly introduced parameters controlling the phoneme loop affect the detection of OOV words and compare the detection and recovery of OOV words from first-best hypothesis and lattices, leading to the conclusion that the new approach of lattices derived OOV candidates provide a significant advantage over using only first-best results.

Chapter 5 changes the underlying ASR model to a word-based Listen-Attend-Spell (LAS) approach and provides a detailed description of the model architecture of the baseline system and its training. This approach uses the first-best hypothesis. When an OOV token is detected, it predicts the position in the input and a phoneme recognizer is used to generate 50 pronunciation candidates, which are then clustered together with other detected OOV words. As for the hybrid ASR system, a spelling is generated for each cluster with more than one OOV candidate. A significant finding is that contrary to a previously reported approach, the center of attention is not a reliable indicator of the OOV location. On average, it is shifted, however, even after compensating for a shift, the center of attention is not a good indicator. Introducing Connectionist Temporal Classification (CTC) objective function during training and using CTC during inference was essential to locate OOV word locations. In addition, the CTC loss also encouraged the generation of more OOV word predictions, which positively affected OOV word detection.

Chapter 6 extends the previous LAS architecture with a novel speller module that allows generating spellings for detected OOV words. This represents a more principled approach to modeling different granularities jointly and simplifies the recovery approach since no phoneme recognition or phoneme to grapheme conversion models are needed. It also introduces a tied word embedding module that the speller module can use as input. In addition, the speller can get information from other sub-modules of the LAS model concerned with the acoustic context and decoding state. A thorough investigation finds that for the speller, all sources are useful; however, an interesting finding is that the most critical input is derived from acoustic context and decoding state, while information derived from the OOV embedding is not vital, even if multiple OOV embeddings are used or combined. However, utilizing multiple OOV embeddings indicates that the system clusters them largely based on grammatical function.

Chapter 7 provides the conclusion and future directions.

In her research, Ms. Egorova made several significant contributions to the field of automatic speech recognition. Some of the contributions are listed below:

- An experimental design to identify suitable (archaic) OOV words for reoccurring OOV word detection based on the idea that language changes over time.
- An advanced Hybrid ASR system that integrates phoneme loops for OOV detection in a principled way in the decoding graph keeping the different granularities separate, with a novel method to extract OOV candidates from generated lattices that are not necessarily on the first-best path using WFST methods.
- A novel WFST-based method of clustering OOV candidates and correctly combining the probabilities of extracted phoneme sub-lattices represented as WFST from multiple occurrences results in an efficent and elegant way to improve the pronunciation of detected reoccurring OOV words.
- Experimental results prove that her lattice-based approach of recovering reoccurring OOV words outperforms the first-best approach.
- A thorough investigation of a word-based LAS end-to-end model on identifying the location of predicted OOV words. It led to the conclusion that attention has significant shortcomings for this task and that adding the CTC objective function during training and then using CTC segmentation leads to substantial improvements in locating OOV boundaries.
- A novel model extension for the word-based LAS end-to-end model was proposed introducing a speller component and thoroughly investigated. The approach allows the joint training of

different granularities (words and characters) in a principled way. Experiments find that information representing "acoustic" and "language" model context is critical to generate the correct spellings. In contrast, the introduction of multiple OOV embeddings mainly allows for capturing the grammatical properties of OOV word usage.

As a non-native speaker of English, I find the thesis is very readable and have no major comments. Minor remarks and a few typos will be passed separately to Ms. Egorova. My only suggestion is to extend the related work section (Chapter 2) to provide readers with more context into what has been done previously regarding dealing with OOV words and how this work differs. For example, Hetherington 1996, Kemp & Jusek 1996, Florian Gallwitz 1996, Schaaf 2001, Rastrow 2009, investigated collateral damage OOV words cause, a diverse set of acoustic and language model extensions, and the usage of lattice and extended dictionary to recover OOV words.

Ms. Egorova's research findings have been published in leading conferences in the field of Automatic Speech Recognition, i.e., the International Conference on Acoustic Speech and Signal Processing (ICASSP), the ISCA Interspeech conference, and related IEEE publications of the Signal Processing Society. She has significantly contributed to the Automatic Speech Recognition community, as indicated by other groups taking up her approach [cf. pp. 38] and building on her research. Her thesis and experimental design lay the foundation for other researchers to address challenges in handling OOV words and dealing with rare words in the rapidly changing technology environment of end-to-end models, which is also supported by making code from her experiments publicly available.

To the best of my knowledge, this is the first thorough analysis of dealing with reoccurring OOV words in hybrid ASR models and also addressing the challenges of end-to-end trained LAS models, which led to a novel extension of the LAS model that combined multiple granularities (words and letters) in a principled way.

In conclusion, it is without any reserve that this doctoral thesis meets the requirements of the proceedings leading to a Ph.D. title conferment.

Specific questions to the candidate:

- OOV words often cause collateral damage. The rule of thumb is that in English on average 1.3 to 1.5 errors are introduced per OOV word. Did you observe a similar effect?
- How did you tread the chosen 1000 OOV words during training?
- When the detected reoccurring OOV words were added to the Hybrid ASR system for a second decoding pass, they were added as unigrams. Why did you choose unigrams and not the OOV class? In this context, did you analyze how many additional correct OOV occurrences were detected on the first-best compared to the first pass, and how many collateral errors were removed?
- For end-to-end models your results indicate that WER improves with growing vocabulary sizes. Have you tried larger vocabularies than the one reported, or do you have an intuition of how large a word-based vocabulary would be to approach WER performance similar to BPE encoding?
- What is your intuition why the introduction of the speller model improved WER1 performance?

Sincerely,

Dr. Thomas Schaaf
Principal Research Scientist
3M | M*Modal