Department of Computer Graphics and Multimedia
Faculty of Information Technology, Brno University of Technology

Review of doctoral thesis
Out-of-Vocabulary Words Detection and Recovery
s*ubmitted by* Ekaterina Egorova

In this thesis, techniques to deal with words outside the recognition vocabulary (OOVs) in automatic speech recognition (ASR) systems are investigated. There is a particular focus on OOVs that repeat several times in the test set. OOVs are important since the number of possible words is practically unlimited and rare words are usually content words like proper names. Without these techniques, the recognition is incorrect, and all subsequent processing modules - like semantic analysis or information retrieval - cannot easily recover from the error. The following problems are addressed in the thesis: detecting the presence of OOVs, finding the start and end times, extracting the pronunciation, clustering detected OOVs to identify repeating words, recovering the spelling and integrating the newly discovered words into the recognition vocabulary.

The structure of the thesis reflects the advances of ASR systems in the last years: moving from word based hybrid acoustic / language models (LM) to End-to-End (E2E) ASR systems based on words or character sub-words without using a pronunciation lexicon. This has big implications on the role of OOVs. Detection and recovery techniques are developed for both type of systems and are evaluated on a publicly available data set.
The thesis has 7 chapters and 87 pages, (67 excluding the appendices, abstract, table of contents, etc.) This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents the overall conclusion and recommendation to the committee.

Technical content of the thesis and remarks to chapters

Chapter 1 introduces the OOV problematic, defines the tasks and metrics for ASR, OOV detection and recovery and states the goals and major claims of the thesis.
One of the goals of the thesis is to contribute to a standardization of OOV tasks and metrics used in OOV research to reach replicability of results. The careful design of the OOV metrics (F-score, WER1, WER2 and WERr) that allows for systematic and intuitive evaluation of the systems is one of the big plus in this thesis, and I hope more research will adopt this methodology.

The different systems are systematically evaluated with the metrics from Chapter 1, but due to the historical development of the research and the usual time/space constraints, not all the systems are trained on the same data, and are evaluated with all the (same) metrics and on the exact same data, which makes the comparison sometimes (unnecessarily) difficult. The use of the terms recall and precision could be more precise: some parts measure just on OOVs from the selected list and sometimes all OOVs are considered. Recall is usually based on tokens, but precision for recovery is based on OOV types (entries in the OOV list). Results would be easier to compare when always giving recall and precision for OOV detection - some parts report it just after spelling recovery, sometimes recall/precision is avoided an just the WERr (recovered WER) is reported.

Concerning the task OOV detection, I would suggest to go beyond current custom and define it more precisely. There are different kinds of training data and thus different kinds of OOVs (the dimensions combine): words not in the lexicon (i.e. either pronunciation or spelling unknown), words not seen in the acoustic training data, words not seen in the textual training data and words removed from the recognition vocabulary (due to e.g. model size limits). Those distinctions can help to better understand some of the results in the following chapters.

In Chapter 2, a literature review is conducted, explaining the OOV problem for past and current systems, systematizing historical and state-of-the art approaches for OOV detection and recovery. The reader gets a good overview of the field and the literature review is done thoroughly. Just for the OOVs in the E2E framework I would appreciate a broader review including the accuracy of E2E ASR (also sub-word based) on rare words.

I think one should avoid the short-cut: when using sub-words (e.g. BPE) there are no OOVs. E2E models require a lot of audio training data and therefore, words that are not frequently seen in the training data (rare words) or OOVs are often mis-recognized, even if the sub-words can in theory construct those OOVs.
The selection of different kinds sub-word / word units and size of vocabulary is a compromise between the amount of training data vs type of parameter sharing (or amount of trainable parameters) and a compromise against the size of the model and the runtime performance. OOVs are a natural consequence of those trade-offs. The choice of modeling units dictates what kind of different OOVs result from it. There is some kind of unobservedness (OOV) in all systems - e.g. an unobserved character or BPE unit (syllable, morpheme), an unseen sequence of sub-word units, an unseen word or compound word (inflection, ...) etc.

Chapter 3 explains the methodology for creating a realistic experimental setup to measure performance of OOV detection and recovery. It's a difficult task in an academic setting, where data and computing resources are more limited. I appreciate that the thesis avoids to simply use an artificially small recognition vocabulary as done often in the literature, but uses a realistic vocabulary size of 200k words (for the hybrid system). The experimental design based on the publicly available data set (LibriSpeech) can serve the community as a benchmarking set.

The difficulty is in having enough repeating OOVs for the clustering while at the same time having rare enough words to have to some extent realistic and meaningful OOVs to measure the OOV detection performance. The thesis choses usage statistics from the Google Books corpus to select out-of-fashion words as simulated OOVs (together with proper names). This criterion achieves a reasonable balance between "rareness" of the word and number of repetitions. However, inspecting the OOV list in detail, I think a good portion of the simulated OOVs are still valid words of today's English or can be easily derived from in-vocabulary words, and my assumption is that realistic OOVs can be even more difficult. To avoid this, I think it would be acceptable to separately measure OOV detection only on the sub-set of really rare OOVs, and OOV recovery on the sub-set of frequent OOVs.

I think the presentation of statistics on OOVs in this chapter could be more clear and better structured - collecting all OOV rates and OOV token counts into a single table for all vocabulary sizes and all relevant data sets. When specifying the OOV rates, one could clarify the two kinds of OOV tokens that contribute: OOVs due to reduced vocabulary (200k, 10k, 5k) and OOVs due to the OOV list that were removed additionally.
If I understand correctly, the OOVs are defined only with respect to the LM training text (thus recognition vocabulary). Concerning the more precise definition of OOVs, it would be good to show the statistics of the mean occurrences of OOVs from the list on the sets used for acoustic training. Given "mean of 51 occurrences" (p.21) for the 360h set, my back-of-the-envelope calculation is that they occur on average 85 times in the 100h+500h acoustic training data, is that correct? That plays a role in Chapter 5.

Even if the author comes from a linguist background, the thesis doesn't motivate why only English is investigated. Question: How are the approach and the results expected to generalize to other languages? What kind of OOVs would you expect and how should we modify the techniques presented here?

Chapter 4 is the biggest chapter and presents a fully integrated procedure for OOV detection, clustering, recovery and addition of new words to the vocabulary for a hybrid weighted FST (WFST) based ASR. Even if hybrid ASR is no longer state-of-the-art, this part of the thesis I appreciate most, since a realistic vocabulary is used (200k) and everything is done in a fully integrated, lattice based and probabilistic approach working completely within the

WFST framework as opposed to many previous works that only considered one-best paths. Also a greedy, hierarchical clustering algorithm for detected OOV lattices based of WFST composition is presented. Adding the discovered OOVs to the recognizer graph lowered the overall WER. All the code was made available on GitHub and the approach was even taken on by another lab (JHU) and further extended by [Zhang 2020].

In this chapter, the usage of "WER" is not specified consistent with Chapter 1 (WER1/WER2/WERr). In my opinion, for OOV detection, the most meaningful metric would be WER2 and not WER1. And for tuning the parameters for recovery, we should tune on WERr and not on WER1. I would welcome a more detailed error analysis (e.g. breaking WER down into substitutions, insertions and deletions) to understand why it was not possible to get better than the baseline (11.7%)? Does this mean that recognizing the OOV regions did never help by e.g. at least fixing surrounding words? Did you consider to include the recovery of in-vocabulary words from detected sub-word sequences (OOV false alarm) into the results?

The thesis mentions as disadvantage the problems with the size of the hybrid decoding graph and the resulting lattices - I think those could be addressed by optimizing the structure of the graph, e.g. by using techniques like on-the-fly FST replacement with the sub-word graph (using OpenFST lazy implementation) and by using a higher-order LM in decoding, but limiting the context for storing the lattice or keeping the hypotheses.

The proposed lattice-based approach represents a big win in recall against the one-best approach, but even after that the recall is still not satisfactory, which deserves deeper discussion. As mentioned in the thesis, in part this due to the big vocabulary since there are many possible confusions with in-vocabulary words which causes to miss the OOV region. However, I think the main reason is not necessarily that hybrid ASR is not state-of-the-art anymore and that E2E ASR is better for OOV detection, since also the E2E ASR has a problem in recognizing rare long-tail words and since its vocabulary was reduced (to e.g. 10k).
For further insights why recall is low (also for E2E) one could inspect the LM scores on the correct path (word and phonotactic sub-word LM) or run oracle experiments: running the recognition with added OOVs or constructing an oracle sub-word LM from phoneme sequences from the OOV list to see whether OOV words are still confused with in-vocabulary words. Due to possible phoneme errors in the cluster pronunciation, it would be good to introduce some fuzziness into the composition score to increase the recall - e.g. one could use a phoneme confusion matrix in the composition.

However, the state-of-the-art has moved on, and in **Chapter 5**, the author shows how to detect OOVs in one type of E2E ASR systems (LAS with CTC). The challenge is that those systems no longer use a frame-based alignment are neither guaranteed nor trained to output the (word) labels inside the acoustic region of a word. Thus the main problem is the location of the OOV start and end times. The thesis presents an approach to locate word boundaries for LAS and CTC using their inner representations (i.e. attention weights or CTC alignments): The relation of the OOV region and the attention weights is investigated. I also appreciate the development of the more probabilistic, less greedy clustering algorithm for lattices based on Dirichlet Processes (Chinese Restaurant Process - CRP).

The E2E systems used here use word labels, since that naturally allows the definition and detection of OOVs and leads to OOV regions in between word boundaries. The thesis shows that the maximum of the attention weights is not necessarily in the OOV region, but that a region with 90% of the attention mass is likely to catch the OOV (higher recall), however it's a relative large region, definitely longer than the OOV duration, which results in bad precision for recovery. Therefore, a parallel CTC is added, which allows to recover the OOV region from its alignments (word labels and blanks). A phoneme recognizer (similar to the hybrid ASR) is needed to generate phoneme lattices, and those are cut based on the timings of the OOV region from the best path from E2E ASR. Based on that, using the new clustering algorithm, the new E2E approach reaches very good OOV detection recall (81.5%) with reasonable precision.

The main insight is that a second ASR system is needed for either finding the timing or the sub-word lattice extraction. Therefore, I would consider splitting the OOV detection task: the detection whether the OOV is present (finding the word string with OOV) can be done by the powerful E2E ASR and finding the exact timing can be done by alignment with a conventional hybrid ASR.

The presentation of results could be more clear by distinguishing detected OOVs that are from the OOV list (the desired ones) from those that result from the reduced 10k vocabulary. It seems that both enter into the precision numbers and are used in the clustering and thus might lower the percentage of correct clusters. However, my guess is that the reason why the OOVs from vocabulary 10k-200k don't influence the clustering is that they are actually much less frequent than the OOVs from the list. That might be due to the slightly artificial "creation" of the OOVs that need to be recurrent and this might not be the case in realistic data.

The chapter concludes that the E2E approach is better for OOV detection and recovery than the hybrid FST approach. I think however, that while the system in Chapter 5 is performing better than the system in Chapter 4, the experiments are not suited for a general comparison between E2E and hybrid ASR:
different clustering algorithms were used, the E2E only uses one-best outputs (hybrid uses lattices), the E2E was trained on significantly more data (600h vs. 100h), but more importantly, the E2E system uses a reduced vocabulary (it cannot be trained well for a huge vocabulary with limited amount of acoustic data).
10k words seems artificially small. The E2E system can simply label all words from 10k-200k as OOV and thus when "confusing" them with the "real" OOVs this will be treated as correct, while the hybrid system has to distinguish all those words among each other and against the real OOVs. Words from 10k-200k constitute a small percentage of tokens (small impact on precision) - if the E2E imply labels all words beyond 10k as OOV, it will have a much higher recall, but the impact on precision (false alarms) is limited. Also from Table 5.3 we see that aiming for best WER does not necessarily mean aiming for best OOV detection performance - maybe a smaller vocabulary could be better suited for OOV detection? However, I want to stress that the systems proposed in Chapter 5+6 perform excellently on the OOV detection task.

Unless I missed it, the thesis doesn't seem to mention which of the OOVs in the evaluation were seen in the acoustic model training data or whether the data was filtered for those OOVs. That is probably because in the research field, OOVs are only defined with respect to the LM texts and recognition vocabulary. However, given the speculation above (average OOV occurs 85 times in 600h training data) - the OOVs were seen sufficiently often in acoustic training to be learned. The hybrid ASR is less powerful (less use of context etc.) and will not much benefit from that, but I think the E2E ASR has an "unfair" advantage from the way how it was trained. When it uses the OOV/UNK label as one output, the system is actually discriminatively trained for the OOV detection task. While the aim is to generalize to new OOVs, the system is still explicitly trained on certain OOVs from the training set. The phoneme sequences are not labeled, but the system can remember acoustic sequences that are mapped to UNK in a similar way as it learns to map acoustic sequences from regular words to their corresponding output symbol.

Therefore, it is important which of the OOVs were seen in acoustic training - on those the OOV detection performance could be "artificially" high since they can be explicitly learned. Otherwise for a fair comparison, also the hybrid ASR should use the acoustic sequences of training OOVs (e.g. discriminatively train an OOV detection module or use the phoneme sequences of OOVs explicitly in the sub-word LM to recognize them as UNK). However, I think the goal of thesis is not to compare hybrid and E2E, I would rather argue that those OOVs seen in the acoustic training data are actually not that interesting, since we already know them - we could explicitly recognize them with a corresponding sub-word graph. Maybe what we really should aim for is to detect completely unseen OOVs - i.e. achieve generalization. For the evaluation, it might be interesting to split the sets of OOVs into those two sets and evaluate the performance on both separately.

Question: Do you have any insights how much the new CRP clustering algorithm improves recall/precision over the hierarchical one?

In Chapter 6 the author proposes a new system that addresses the main shortcoming of the E2E approach - the integration of the two granularities - words and sub-words. The speller architecture (i.e. character predicting recurrent network) presents a principled way of jointly training (and recognizing) word and character based outputs (organized in a hierarchical way) and this thesis is its first application in word-based E2E ASR. The speller recovers words directly during E2E decoding. The thesis analyzes the effect of input information from different parts of the LAS system into the speller. Three types of information are important: the hidden state representing the context words, the attention vector summarizing the relevant acoustic information and the embedding that is trained to represent both word context and spelling information (character sequence).

An interesting results is that using spelling information in word embeddings gives impressive WER improvements - an important contribution to word based E2E systems for the community.

The thesis also explores how to correctly use the OOV word label in the word predicting E2E ASR and its (word) embeddings in the speller. The problem is that the "OOV" label is not very informative, since the embedding should be predictive of the spelling of the actual word. Therefore, allowing multiple OOV embeddings for the one OOV label is explored with unsupervised clustering of those during training. The resulting OOV embedding clusters and the corresponding types of recovered words are also analyzed and show e.g. relations to grammatical categories.

Introducing of the speller substantially improves the amount of correctly recovered OOVs, even if they occur just once. Unfortunately, the accuracy on OOVs with the speller architecture doesn't reach the recovery rates of the BPE baseline system - which is not able to detect OOVs, but can correctly recognize the spelling of many OOVs. As the thesis already stated, there are probably different kinds of words for which either system works better. I suppose e.g. the frequency of the word plays a big role. The BPE system was not restricted in vocabulary and it has seen many of those spelling sequences in training. This is an indication that one should use the information from all words in the training of the system - not just 10k words. The good results for BPE suggest that a speller with BPE outputs might be worth trying. To be able to better compare the two systems, OOV recall and precision numbers for the detection and the recovery task would be helpful. "rOOVs" doesn't distinguish wrong spelling (recovery precision) or missed OOV (recall) or fuzzy OOV region (detection precision). Since different test sets and training sets were used, we cannot directly compare to the results in earlier chapters.

The text is not fully clear on how the OOV spellings were used in the speller training. I assume the spellings of all words up to 200k vocabulary were used. It's slightly confusing since in the other chapters, those words are treated as "completely OOV" and neither their spelling nor pronunciation was used in the training (and decoding). As argued above, since we actually know up to 200k words, we want a word-sub-word recognizer that can back-off to a spelling model to still be able to recognize those words, even if we cannot fit them into the 10k word vocabulary. So if we train the speller on those words, we will hopefully recognize them correctly.
Thus, the proposed combined speller architecture nicely concludes the research "journey" by mixing the best of both worlds from Chapter 4 and 5 - the power of E2E for high recall and precision in detection and the integrated and fully aligned sub-word system for the recovery of spellings.

Question: Throughout the thesis, you chose phonemes as sub-word units - what's your expectation how the approaches would work when using larger units like BPE or multi-phoneme/character units?

Question: Since your speller is inspired by [Mielke and Eisner, 2018] could you discuss on what was changed to the original approach? (e.g. use of word begin and end tags (BOW/EOW), feedback of character embedding, word embedding needed to be projected to a lower-dimensional space etc.)

Chapter 7 summarizes the finding of the thesis and provides an outlook to future research by stressing the importance of recovered OOV for subsequent tasks.
I provided a list with minor corrections and technical remarks to the author.

Summary on the technical content of the thesis

The thesis clearly demonstrates the qualities of the candidate – capability to study non-trivial literature, suggest own novel solutions to difficult problems, implement them, use advanced machine learning toolkits, carefully test and discuss the results of experiments.

I highly appreciate the results published in high-profile conferences, the care that has been put into developing and standardizing the methodology for OOV metrics and the quantity and quality of experiments done on the publicly available data-set. I find it very valuable that all publications are accompanied by releases of code. The contribution of the candidate to the international research community is clear.
The suggested combined word-prediction and speller system using internal inputs from word prediction is a promising approach for the coming E2E era.

Comments on the formal aspects

The thesis has a very logical structure and it is easily readable. It is written in well readable, almost flawless English (except a few typos or missing determiners which do not impact the understandability). The quality of presenting the results is excellent, the tables and figures are carefully designed and well readable, well annotated, and provide straightforward information on how the approaches work and what was achieved. All formulas seem to be correct and the mathematical notation is consistent (just alpha is used for several contexts).

Summary and recommendation

I have carefully examined the doctoral thesis of Ekaterina Egorova. Despite some criticism raised above (many points are rather recommendations or discussion on deeper analysis than critique), in my opinion, it is a solid work that contributes to substantial progress in the ASR research field. I also examined her publication track and find it meeting the standards for a PhD candidate at a respected University. I appreciate the first authored papers at standard speech conferences such as Interspeech and ICASSP, the journal contribution in the IEEE, as well as the many co-authored papers.

To conclude, I am pleased to accept this thesis without reservations and I consider this thesis does meet the generally accepted requirements for the doctoral degree and I recommend conferment of this degree by Brno University of Technology.

In Prague, December 1st, 2022

Dipl.-Ing. Mirko Hannemann, Ph.D.
Research Staff Member
Apple Siri
Václavské náměstí 47
110 00 Praha 1
Czech Republic