Out-of-vocabulary word detection and beyond *

Stefan Kombrink, Mirko Hannemann, Lukáš Burget

Speech@FIT, Brno University of Technology, Czech Republic
{kombrink,ihannema,burget}@fit.vutbr.cz

Abstract. In this work, we summarize our experiences in detection of unexpected words in automatic speech recognition (ASR). Two approaches based upon a paradigm of incongruence detection between generic and specific recognition systems are introduced. By arguing, that detection of incongruence is a necessity, but does not suffice when having in mind possible follow-up actions, we motivate the preference of one approach over the other. Nevertheless, we show, that a fusion outperforms both single systems. Finally, we propose possible actions after the detection of unexpected words, and conclude with general remarks about what we found to be important when dealing with unexpected words.

1 Unexpected events in speech recognition

Events in speech can be arbitrary sounds. One possible challenge is to decide whether a particular sound is actually speech or noise and is called speech/non-speech or voice activity detection (VAD). Another challenge is to find the most likely sequence of words given a recording of the speech and a speech/non-speech segmentation. This is commonly known as automatic speech recognition (ASR) where words are constructed as a sequence of speech sounds (usually phonemes).

Although the set of speech sounds is considered to be limited, the set of words is not[1]. Language models are commonly used in ASR to model prior knowledge about the contextual relationship of words within language. This prior probability distribution over words is conditioned on a history of preceding words and highly skewed. Usually, this distribution is discrete, i.e. only a limited set of most frequent words is known to the system. Unknown words constitute an unexpected event, and since most words occur rarely, enlarging the vocabulary does not alleviate this effect. In fact, the recognizer will replace each of these so-called out-of-vocabulary (OOV) words by a sequence of similar sounding in-vocabulary (IV) words, thus increasing the number of word errors and leading to loss of information.

Here, we investigate two different approaches to detect OOV words in speech sharing a similar strategy: finding incongruences between the output of a generic

* This work was partly supported by European project DIRAC (FP6-027787), by Grant Agency of Czech Republic project No. 102/08/0707, Czech Ministry of Education project No. MSM0021630528 and by BUT FIT grant No. FIT-10-S-2.

[1] There is no known limit for the length of a word.

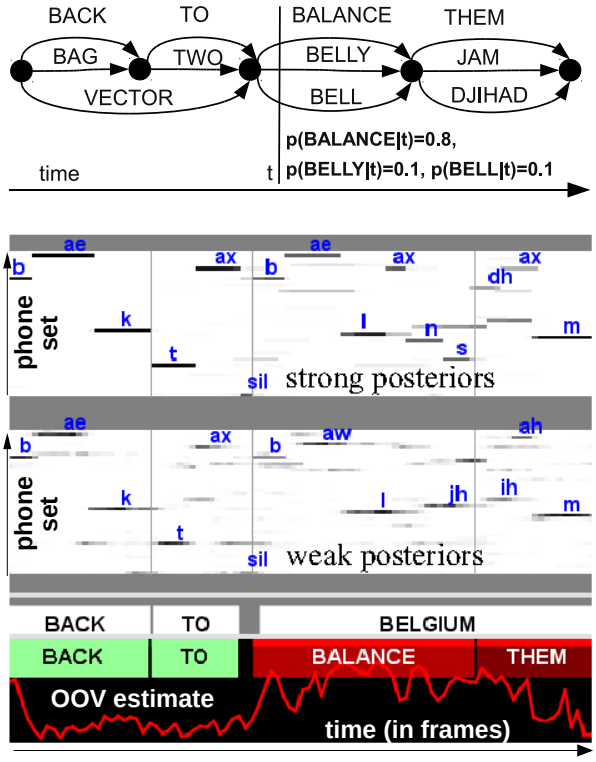

Fig. 1. NN-based system detecting the OOV word "BELGIUM": Processing of a word lattice (top) produced by the specific recognizer. The incongruence between the specific and the generic phone posteriors is detected by the neural net and identifies the corresponding words as OOVs (bottom).

(unconstrained) and a specific (constrained by prior knowledge) system. We combine both approaches, show results of the fusion, and interpret those. The final part is dedicated to possible follow-up actions after OOVs have been detected.

2 Neural network based OOV word detection system

If the outcome of an unbiased observation contradicts the expectations raised by higher level knowledge, we refer to this as an incongruent event. The incongruence can be detected by comparing the output of a generic and a specific recognizer. In our case, the specific recognizer uses prior knowledge in form of a language model, vocabulary and pronunciation dictionary, and searches for the best sequence of words with the highest overall likelihood. The generic recognizer uses only a limited temporal context, and is thus less constrained. A neural net with the output classes $C_{NN} = \{silence, ivcorrect, ivincorrect, oov\}$ is used to determine

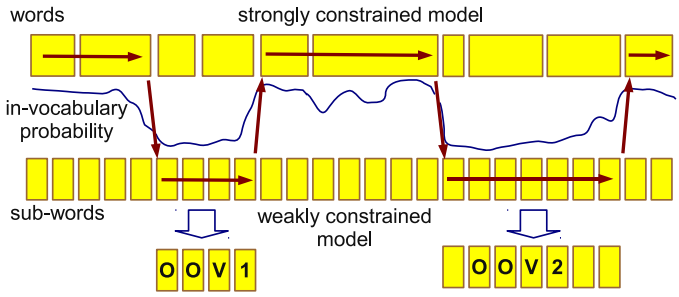

Fig. 2. OOV word detection using hybrid recognition: The best path (arrows) contains words and sub-word sequences, which can be regarded as potential OOVs. In addition, a frame-wise confidence measure is extracted from the combined word/sub-word lattice output of the recognizer shown as in-vocabulary probability.

- whether a recognized word is overlapped with OOV speech or not
- whether a word is mis-recognized or not
- whether a word is OOV or not, given the word was mis-recognized
- what is the most probable class $c \in C_{NN}$ of the word

In [Kom09] we applied this approach to noisy telephone speech, reported improvements, and found it to generalize reasonably well across different data sets.

3 OOV word detection based on a hybrid word/sub-word recognizer

However, our NN-based OOV detection approach does not retrieve a description of the underlying OOV, and, in cases where the recognized word boundaries do not match the reference, it cannot indicate accurately where, within a word, an OOV starts or ends. That is why we recently used a hybrid recognizer which consists of specific word models and a generic word model [Del95] for OOV word detection. The generic model is able to detect OOV words as sequences of sub-words. The search for the most likely word sequence can choose either an in-vocabulary word or the generic word as shown in figure 2. We compare the real output of an existing word-only recognizer and the best possible output of a hybrid word/sub-word recognizer, respectively:

```
 reference: SORT OF A BLUEISH(OOV) MEDIUM
  word rec: SORT OF EVOLUTION        YOU
hybrid rec: SORT OF A bl.uw.ih.sh  MEDIUM
```

It can be seen, that the hybrid recognizer carries potential to simplify and improve the detection and localization of OOV words over our NN-based system. This is mainly due to the following reasons:

- The resulting word boundaries in OOV regions are more flexible, thus potentially more accurate. Context words are less often mis-recognized.

– The decision of the recognizer to prefer sub-word sequences over word sequences provides good evidence for an OOV word.
– Often, two or three words in the word recognition are overlapped with a single OOV word. When using the hybrid recognition output, however, in many cases one sub-word sequence aligns to just one reference OOV word.

Using this setup, we have two possible choices for evaluation: Either we treat each sub-word sequence in the recognition output as potential OOV. This yields high precision, but many OOV words are missed. Alternatively, all words and sub-word sequences in the recognition output can be potential OOVs, which corresponds to the task performed previously using the neural-net based OOV detection system. In that case the recall in OOV detection improves, but the number of false alarms increases and the regions of OOV words tend to be less accurate. In case the detected OOV word was decoded as a sub-word sequence, we implicitly obtained a phonetic description of the OOV. Unlike before, we now just performed OOV detection using a hybrid confidence measure estimating the posterior probability for $C_{hybrid} = \{iv, oov\}$.

4 Fusion of both methods

We combined the scores of our both OOV detection methods by using linear logistic regression. 2.5 hours of Fisher data (telephone speech) were used for training and 7.5 hours for evaluation. The OOV rate was around 6.1%, and the neural-net based OOV detection system was trained using a disjunctive set of OOV words. All scores for the fusion were created initially on frame-level[2] and represented posterior probabilities:

$$\sum_{c \in C} p(c|frame) = 1,\ C \in \{C_{NN}, C_{hybrid}\} \tag{1}$$

A hybrid confidence measure (Hybrid CM) estimating a probability of being in OOV was extracted from the lattice output of the hybrid recognizer and a binary score (REC) based on the recognition output of the hybrid recognizer (1 for frames covering sub-word sequences, 0 otherwise) were included in the fusion experiment. Our NN-based system estimated posterior probabilities of four classes using two neural nets using different type of context in the input [Kom09]. We converted the posterior probabilitities into log-likelihood ratios[3] and averaged them over the word boundaries provided by the hybrid recognition output to obtain word-level scores.

Figure 4 shows OOV word detection performance of scores of both systems (bold lines) and their fusion (dashed lines). The left plot shows the zoomed view of the operational range reasonable for almost all tasks. The performance of all scores across a wide range is shown in the right plot. The best performance is achieved using different fusions for different ranges of false alarms (FA):

[2] 10 ms length.
[3] $LLR(p) = \ln \frac{p}{1-p}$

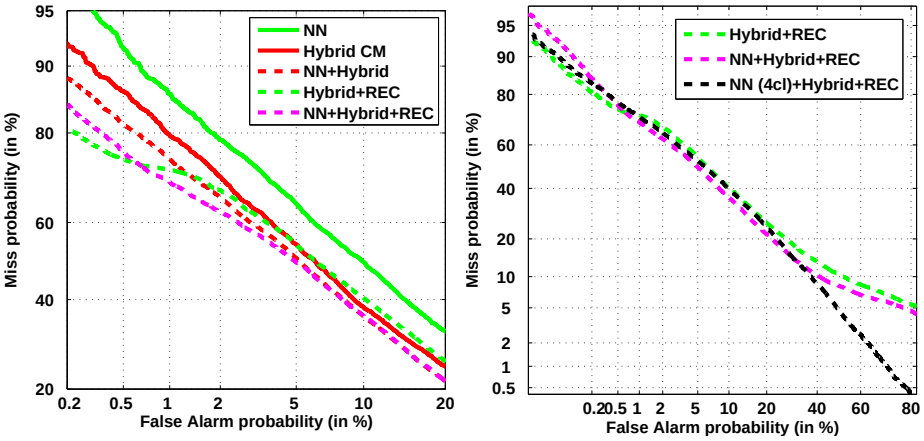

Fig. 3. Combined OOV word detection performance: Detection error tradeoff across a wide range (right) and a range suitable for real application (left).

1. Up to 0.57% FA - hybrid system only
2. From 0.57% up to 20% FA - hybrid and NN system
3. From 20% FA - hybrid and NN(4cl) system

In the first range, we obtain a high precision in OOV detection. The best fusion intersects with the operation point determined by the binary score obtained from the word/sub-word recognition output. This is around 0.57% FA, where the fusion during the *second range* slowly starts to gain from the NN-based scores. Here, we retrieve already more OOV targets as opposed to the smaller amount of targets contained in the sub-word sequences in the word/sub-word recognition output of the hybrid system. *The third range* benefits from using the scores of all four classes of the neural net. Some OOV words gets detected better by the NN-based system, but at the cost of retrieving many false alarms - far too many to be of practical use.

To conclude, the NN-based score improves the OOV detection performance across a wide range when fused with the hybrid CM. However, the better decision is to use the one-best binary score in the fusion, unless recall is more important than precision. In that case, the neural net is still able to retrieve some OOV words which otherwise would have been missed in the mid-range of the detection error trade-off curve.

5 Beyond OOV detection

Upon detecting an unexpected event, the system should react. As a default strategy, even ignoring words detected as OOVs prevents mis-recognitions. However, unexpected events potentially carry a high amount of information - i.e. OOVs are most often content words. Thus, it is desirable to localize and analyze the

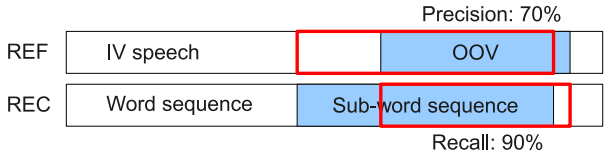

Fig. 4. The quality of a detected OOV word is determined by precision and recall.

event, which is a prerequisite for further processing stages to deal with the event in a more sophisticated way. The following actions could be taken upon detection of an OOV word:

- *Analysis:* obtain a phonetic description.
- *Recovery:* obtain the spelling and insert it into the recognizer output.
- *Judge importance:* some classes of OOVs might be particularly interesting, e.g. the class of OOVs that suddenly occur several times, such as the name of a formerly unknown politician in broadcast news.
- *Query-by-example:* find other examples of the same word.
- *Similarity scoring:* group re-occurring (or similar) unknown words.
- *Higher level description:* relate the new word to known words and to other detected OOVs.
- *Model update:* estimate a new word model and integrate it into the system.

The usefulness of particular OOV detections may vary from task to task. If it is just to detect mis-recognized words in the recognition output (due to the presence of an OOV), it is sufficient to find a single phone or frame in the word that has a low confidence score. However, if the task is to describe the OOV or to retrieve other examples of it, detecting a single phone of the OOV is not any helpful - we need to get the OOV region as exactly as possible. Therefore, we analyze detections by measuring recall of the OOV region and precision of the detected region (the sub-word sequence), as shown in figure 4.

5.1 Spelling recovery of OOVs

Using grapheme-to-phoneme (G2P) conversion [Bis08], we retrieve the spelling of a word from the phonetic description. By substituting the sub-word sequences with the estimated spellings, we are able to correct a significant portion of recognition errors due to OOVs [Kom10] and can also identify false alarms, in case the sub-word sequences convert back to known words. The retrieved spelling is a human readable representation of the OOV (e.g. EXTINCTION, PANDEMIC, GRAVITATIONAL) , which is interpretable within context also in case of slight errors (e.g. COURTICAL, EMBALLISHMENT).

5.2 Finding re-occurring OOVs

Due to the higher level structure of audio/texts (into documents, broadcast shows, telephone calls), several OOVs do not only occur once, but repeat several

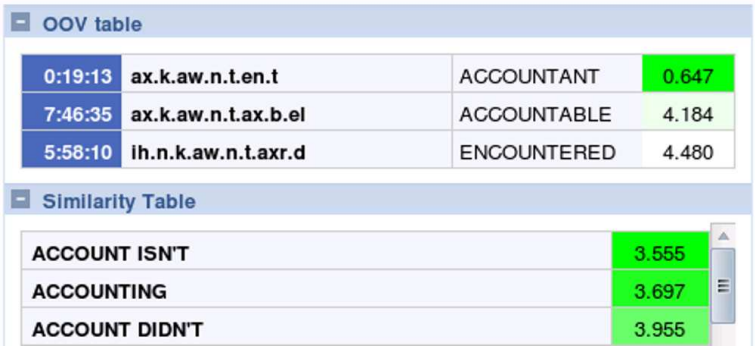

Fig. 5. OOV demo on the selected OOV detection 'ax.k.aw.n.t.en.t': the top table shows time stamps where similar detections are found, and their recovered spelling, respectively. The output is ranked by a similarity score, with the selected detection ranking at top. The bottom table shows similar IV/OOV compounds.

times within different contexts. Those words often belong to topic-related vocabulary and are particularly important. Given one example of the word, we want to find other examples (query-by-example) and we want to cluster all detected OOVs to judge, whether some of them are re-occurring, and thus, important. For both tasks, we need a similarity measure of detected OOVs. The phonetic description of the detected OOVs, however, will not match precisely, as shown in this example detections for the OOV "ILLUMINATION":

```
ax l uw m ax n ey sh en
   l ih m ax n ey sh en z
```

In [Han10], we described a similarity measure based on the alignment of recognized sub-word sequences. With the help of an alignment error model, which is able to deal with recognition errors and boundary mismatches (varying recall and precision of OOV region), we could retrieve roughly 60% of the re-occurring OOVs in telephone calls.

5.3 Relating OOVs to other words

Looking at examples of OOVs [Han10], we observe that unknown words most often are not entirely unknown. Except e.g. proper names in foreign languages, the majority of OOVs can be - morphologically or semantically - related to other known words or to other OOVs (derivational suffixes, semantic prefixes, compound words). Such a higher-level description of the unknown word can identify word families and identify the parts of the word, that are not modeled yet. We achieved this analysis by aligning a detected OOV to sequences of IVs and other detected OOVs. This is essentially a second stage of decoding, where we decode the detected sub-word sequences using a vocabulary consisting of all IV words and all other detected OOVs.

Figure 5 shows a screen shot of our OOV word detection and recovery demo available at `http://www.lectures.cz/oov-fisher`. It demonstrates the follow-up tasks such as spelling recovery, finding of similar OOV detections using similarity scoring and related compounds created out of known and unknown words.

6 Conclusions

In this work, we investigated into two approaches for OOV word detection. We compare both systems in a fusion experiment, and describe how to actually make use of the detected incongruence. We successfully implemented some out of the proposed follow-up actions (spelling recovery, similarity scoring and higher level description). Our approach relates parts which are well-known (sub-word units) to whole words which are not modelled yet (OOV words), which corresponds to the part-membership relationship postulated in the theoretical DIRAC framework.

Speech recognition is a sequential problem: prevention of damage in the context, and identification of the region affected by an unexpected event is particularly important to us. When aiming to go beyond OOV word detection, it became clear, that designing a system just primarily for detecting unexpected events might not be desirable. This became clear, especially when specific and generic systems were combined for the purpose of incongruence detection, but the localization was difficult and valuable information necessary for the follow-up process was lost. After extending our first approach by a hybrid recognition, we improve detection, and sustain higher accuracy in localization.

Another conclusion is, that a standard task definition for OOV word detection does not exist, and neither does it seem reasonable to define it. The usefulness of a particular OOV detection depends highly on the intended follow-up tasks, which again commends to first examine *how to react* on an unexpected event, in order to gain insights about how to improve its detection.

References

[Del95] Deligne et al: Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams. ICASSP, 169-172, Detroit, MI, 1995

[Jia05] Jiang, H.: Confidence measures for speech recognition: A survey. Speech communication, vol 45, no 4, 455–470, 2005

[Bis08] Bisani, M., Ney H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, vol. 50, no. 5, 434–451, 2008

[Kom09] Kombrink S., Burget L., Matějka, P., Karafiát M., Heřmansky H.: Posterior-based Out-of-Vocabulary Word Detection in Telephone Speech. Proc. Interspeech 2009, Brighton, UK.

[Han10] Hannemann M., Kombrink S., Burget L.: Similarity Scoring for Recognizing Repeated Out-of-Vocabulary Words. Submitted to Interspeech, Tokyo, JP, 2010

[Kom10] Kombrink S., Hannemann M., Burget L., Heřmansky, H.: Recovery of rare words in lecture speech. Accepted for Text, Speech and Dialogue (TSD), Brno, CZ, 2010