Similarity Scoring for Recognizing Repeated Out-of-Vocabulary Words

Mirko Hannemann, Stefan Kombrink, Martin Karafiát, Lukáš Burget

Brno University of Technology, Speech@FIT, Brno, Czech Republic

{ ihannema, kombrink, karafiat, burget } @fit.vutbr.cz

Abstract

We develop a similarity measure to detect repeatedly occurring Out-of-Vocabulary words (OOV), since these carry important information. Sub-word sequences in the recognition output from a hybrid word/sub-word recognizer are taken as detected OOVs and are aligned to each other with the help of an alignment error model. This model is able to deal with partial OOV detections and tries to reveal more complex word relations such as compound words. We apply the model to a selection of conversational phone calls to retrieve other examples of the same OOV, and to obtain a higher-level description of it such as being a derivation of a known word.

Index Terms: out-of-vocabulary, OOV, hybrid word/sub-word recognizer, similarity measure, alignment error model

1. Introduction

Certain OOVs tend to occur several times in some documents, while they do not appear in the majority of other documents, and thus are not included in the dictionary of the speech recognizer (ASR). Typically, those are topic-specific words - e.g. while "Mycelium" is a rare word, in a lecture about mushrooms, we observed it more than twenty times. Another example is the name of a new person, which appears in the news only for a certain time period. A conventional ASR will replace those words by similar sounding in-vocabulary words (IVs) and since rare words have a low impact on the word error rate, this is often neglected. However, from an information retrieval perspective, it is desirable to correctly recognize those words. Topic-specific terms contain important information and are also suited for document indexing. After recognizing OOVs, we want to detect whether some of them are reoccurring. For that, it is necessary to develop a similarity measure for recognized OOVs. Ultimately, we would like to compose new word models from reoccurring OOVs and to add them to the recognizer.

Several approaches to OOVs detection exist, either based on computing confidence scores on the ASR output [1, 2], or using backoff-, filler- or generic-word models, which model portions of speech that do not match the pronunciations of words in the vocabulary. Approaches based on confidences mark a recognized word (or just a part of it) as 'wrong due to the presence of an OOV', if the confidence score is low. A conventional recognizer is sufficient for such approaches, but in the presence of OOVs, the word boundaries of the wrongly recognized IVs often do not match the reference words, so that the exact start and end points of the OOV are difficult to obtain.

[0]This work was partly supported by European project DIRAC (FP6-027787), 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. We thank Josef Žižka for fast development of our OOV demo and Igor Szöke for help with the hybrid ASR.

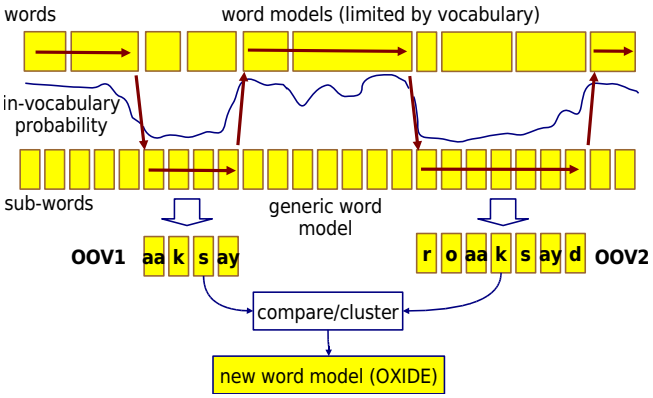

Figure 1: *OOV recognition with hybrid word/sub-word model and comparison of sub-word sequences, extracted from one-best recognition output.*

We, however, are not only interested in detecting whether an OOV was present, but also want to retrieve a description for it, covering the OOV region as precisely as possible. Therefore, we detect OOVs with a hybrid word/sub-word recognizer [3, 4, 5, 6], which contains a generic word model based on sub-words (e.g. phones, syllables, multi-phone units). In the search for the most likely word sequence, the recognizer chooses the word models that maximize the likelihood of the overall path. Everywhere, it has the freedom to choose either an in-vocabulary word or the generic word. Thus, portions of speech, which cannot be modeled well by any word model, will be recognized as a sequence of sub-words. The resulting word boundaries in OOV regions are potentially more accurate than in the conventional word-based ASR.

Figure 1 shows the approach: The fact, that the generic model was chosen for a portion of speech, indicates the presence of an OOV and also retrieves its starting and ending times. The sequence of recognized sub-words serves as a description of the detected OOV. Given several of these detected sub-word sequences (called 'OOV candidates'), we want to use a similarity measure to decide, whether some can be clustered as being the same OOV. For the comparison of sub-word sequences, we introduce an alignment error model. The focus of this paper is not to make any claims about optimality or superiority of the used techniques, but to introduce the task of recognizing repeatedly occurring OOVs and the challenges arising from it.

2. Alignment of recognized OOVs

The output of the OOV detection using the hybrid word/sub-word recognizer is a set of OOV candidates, i.e. detected sequences of sub-words. When operating on multiple output hy-

OOV	recognized sub-word sequence	closest in-vocabulary
abnormalities	ae b n ao r m ae l ax t iy z	abnormally, abnormal
bioluminescence	b ay ax l uw m ax n eh s en s	bio, luminous
counterilluminate	k aw n t axr ax l uw m ax n ey t	counter, illuminated
monochromatic	m aa n ax k r ax m ae t ih k	mono, chromatography
polychromatic	p aa l ih k r ax m ae t ih k	poly, chromatography

Figure 2: *Examples of well-recognized OOVs (36k words in vocabulary) from the TED database*[1].

potheses as e.g. in lattices, each candidate is associated with a posterior probability. So far, we only used the one-best output and made a hard decision on what is considered as OOV.

Our task is to identify for each OOV candidate those other OOV candidates, which are likely to correspond to the same reference word. We do this by aligning the sub-word sequences of two OOV candidates in comparison - the example shows recognized sub-word sequences for the OOV 'illumination':

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

- The alignment requires deletions, substitutions and insertions.
- The second detected sub-word sequence does not cover the whole OOV region (recall of region),
- Only a part of it is correctly overlapping with an OOV (precision of region).

Looking at examples of recognized OOVs (figure 2), we observe, that except e.g. proper names in foreign languages like 'Eyjafjallajökull', the majority of OOVs can be (morphologically) related to other known words or to other OOVs:

- Derivational suffixes: 'abnormalities' → 'abnormal(ly)'
- Compound words: 'counterilluminate' → 'counter' + 'illuminate'
- Semantic prefixes: The OOVs 'poly-' and 'monochromatic' both introduce the OOV 'chromatic'.

It is desirable to identify 'chromatic' as the root (here also OOV), since it can occur on its own (e.g. 'chromatic dispersion') and both prefixes are in-vocabulary. Those examples motivated us to introduce a more general form of comparison by alignment, where we allow to strip pre- and suffixes, and not only compare the OOV candidate to all other OOV candidates, but also to all IVs, and to all combinations of several OOVs or IVs. We implemented this as a search, where we decode a particular detected sub-word sequence to a sequence of other OOV candidates or IVs.

The search space is represented as a lexical finite state transducer (FST) L, which contains all recognized OOV candidates and all IVs in a word-loop[2] (figure 3). Given a particular sub-word sequence (represented as input FST I), we retrieve sequences of OOVs/IVs that exactly match the input by using finite state composition:

$$I \circ L \tag{1}$$

[1] with permissions "http://www.ted.com"

[2] We did not apply a word language model in L, since we assume, that the OOVs are detected in places, where such prior knowledge had already failed.

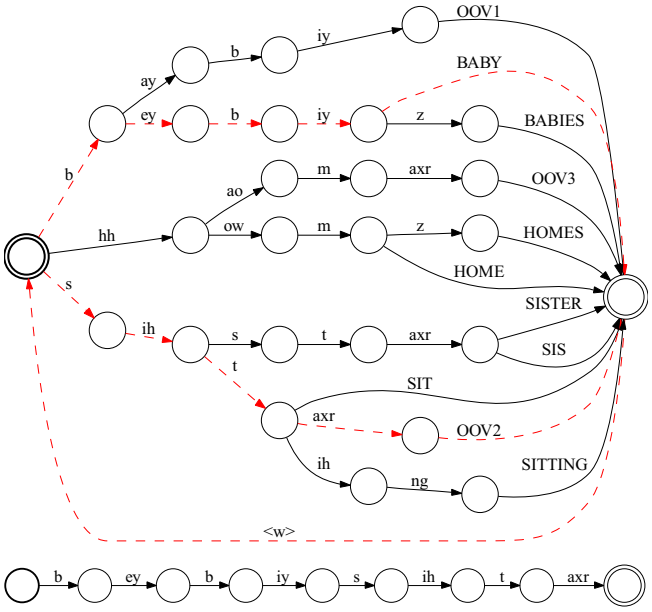

Figure 3: *Toy example of lexical FST L containing all recognized OOVs and IVs and a given sub-word sequence, represented below as input FST I. The composition I ∘ L is dashed.*

2.1. Alignment error model

Since the recognized sub-word sequences can contain errors, we apply a mediating alignment error FST E, which is widening up the hypothesis space by possible errors:

$$F = I \circ E \circ L \tag{2}$$

The result F is a lattice of possible alignments of the input sequence (figure 4). While in the example of figure 3 only one sequence ('BABY OOV2') matches the input, the hypothesized alignment errors allow to retrieve more OOVs/IVs (figure 4).

An error model fulfills two purposes: it compensates for the thin representation of the output - i.e. recovers from just using the one-best hypothesis, and it also adapts to repeating error patterns which were not observed during training of the ASR system (or result from insufficient modeling). For example, if an ASR is applied on foreign accented speech, an error model could be trained on typical vowel confusions.

In our case, the error model E is implemented as a weighted finite state transducer (WFST) [7][3]. When using the log semiring, the weights are negative log-probabilities, which we train on observed alignment patterns in development data. Also the lattice F (Eq. 2) is a WFST - each path through F is a possible alignment of the input sequence to a sequence of OOVs/IVs and has a score/probability attached, used as cost of the alignment.

[3] We use the OpenFST toolkit "http://www.openfst.org/".

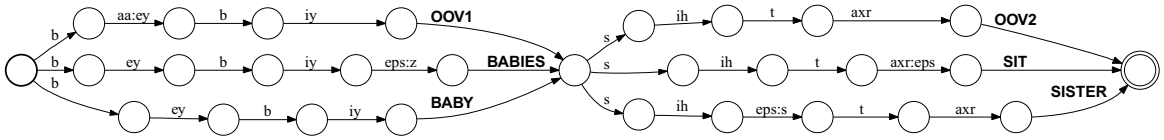

Figure 4: *A resulting alignment lattice $F = I \circ E \circ L$ using the toy example from figure 3.*

In its simplest case, E has just one state [8] and implements three basic alignment operations[4] as self-loops:

- substitutions: `xx:yy`, $-\log(p(\texttt{yy}|\texttt{xx}, TrainText))$
- insertions: `xx:eps`, $-\log(p(\texttt{xx}|Ins, TrainText))$
- deletions: `eps:xx`, $-\log(p(Del|\texttt{xx}, TrainText))$

To better deal with imprecise regions and partly retrieved OOVs and to cope with added/deleted pre- or suffixes, we assume, that at the beginning or at the end of every recognized sub-word sequence there is a region, which can be stripped off, replaced or added at a lower cost. Thus, edit costs depend on, whether belonging to a consecutively edited sequence at the begining/end or to a normal operation within a word. We extended the error model to a layered structure:

- The core layer contains the error model introduced above (possibly conditioning edit operations on phone context).
- An outer layer models consecutive edit operations at the sequence boundaries.
- A word layer models the cost of compounding words.

3. Similarity scoring and clustering

As a similarity score (distance metric) of recognized sub-word sequences, we use the scores from the alignment error model, i.e. the cost of aligning two sequences to each other. Given the similarity score, we could use soft cluster assignments (like in fuzzy clustering), but so far we cluster simply by thresholding the obtained scores.

For a particular OOV candidate, we obtain the lattice F containing possible alignments, with scores. We prune the alignment lattices by only keeping paths in a beam (multiplicative factor) around the score of the alignment of the OOV to itself. The self-alignment score should normalize for the length and the prior probability of the sub-word sequence. We perform a Viterbi forward search and then obtain the n-best paths using the A*-algorithm backwards from the final state. The word labels in the pruned alignment lattice (see fig. 4) serve as other OOV candidates and IVs, which are considered to be similar to the particular OOV candidate. They are used for two tasks[5]:

1. **Query-by-example**: Given an OOV candidate (sub-word sequence), retrieve all other occurrences in the database. We just extract all OOVs from the pruned alignment lattice and report their positions in the data.
2. **Higher-level OOV description**: We retrieve the best full alignment paths. An OOV can be described as being a compound word or as being derived from IVs. E.g. `aa f ax s m ae k s` ('Office Max') → `OFFICE OOV1572`

[4] A state transition is represented as `input:output,weight`.

[5] See "http://www.prednasky.com/_ted/" for a demonstration on Fisher data.

3.1. Cluster evaluation

We decided to exclude short detected sub-word sequences (1-3 phones) from scoring and from our lexicon FSA, since we observed, that they are likely to be false alarms and that the detected region mostly covers only a small fraction of the underlying OOV. Therefore, it can not serve as a discriminative representation of the OOV to be recognized elsewhere. The same applies if an OOV candidate is only partially overlapping with an OOV from the manual reference. We consider a reference word to be 'signicantly overlapping' with the OOV candidate, if the word is either overlapping almost completely, or the overlap is large enough to distinguish the word (e.g. at least 4 phones).

If the same sub-word sequence is recognized several times, most often the same reference words are present (80% of cases), but this is not always true. For example, the OOV candidate `k ao r t eh k s` was recognized for the reference OOVs 'cortex', 'neocortex' and 'Gore-Tex'. Thus, using a particular sub-word sequence as OOV candidate may be ambiguous.

Also, an OOV candidate might overlap with more than one reference word (some of them may be in-vocabulary). For example, for the OOV candidate 'Frederick's photograph'

`f r eh d r ih k s f ow t ax g r ae f`

Several distinct OOVs (and IVs) can be correctly clustered:

`f ow t ax g r ae f ih k` 'photographic'

`f r eh d r ih k` 'Frederick'

We consider two OOV candidates to be correctly clustered together, if there is any reference word (in any of its occurrences), that overlaps significantly with both of them.

When retrieving other examples of the OOV candidate, we consider all words with a common word stem to be correct - e.g. the whole family 'convulse', 'convulsing', 'convulsions', ... should be retrieved and not only the exactly matching word. To be able to evaluate that, we produced a reference stemming dictionary for the IVs and the OOVs in the reference. In a first step, we applied the Porter stemmer[6] to our dictionary and then hand-corrected it by also adding some splittings not considered by the algorithm so far.

4. Setup

Our LVCSR uses 2-pass decoding with speaker adaptations (CMLLR, VTLN) and was derived from the AMIDA 2005 CTS recognizer as previously used in [1]. As acoustic features, we used posterior features using long temporal context. The acoustic models were trained speaker independently on 250 hours of Switchboard data. As recognition network, we used a hybrid word/sub-word language model (LM). The sub-word LM consists of 3977 phone and multiphone units trained on the RT06 dictionary ([5], 47k words). The word LM (bigram open-set Katz-backoff) was trained on $\approx$ 2250 hours of Switchboard (1+2) and Fisher (except test set).

We used 6.2 hours from Hub5 Eval01 as development data to train the error model (negative log-probabilities in a phoneme

[6] Martin Porter, "http://tartarus.org/∼martin/PorterStemmer/"

ASR word accuracy	OOV precision	OOV recall	OOV candidates ($\geq$ 4 phones)	overlap with reoccurring OOV
67.78%	85.3%	23.3%	2323	928

Table 1: *Results of OOV detection with hybrid word/sub-word recognizer on a selection of Fisher data.*

beamwidth	4.0	5.0	6.0	7.0
recall reoccurring OOVs	51.7%	60.0%	66.1%	71.4%
avg. cluster accuracy	70.0%	58.6%	40.7%	25.5%
avg. cluster completeness	42.6%	49.5%	57.2%	62.5%
accuracy OOV description	55.0%	47.8%	38.5%	29.1%

Table 2: *Clustering of OOV candidates using scores from similarity measure at different beamwidths.*

score	OOV candidate	overlaps
0.525	1x `aa k s ay d`	oxide
0.936	3x `d aa k s ay d`	oxide
2.936	1x `d aa k s`	Dachshund
missed	1x `n ay t r ih k aa k s ay d`	nitric oxide

Table 3: *Scoring output for the candidate* `aa k s ay d` *using beamwidth 6.0. The first line contains the self-alignment. The cluster accuracy is* $3/4$ *(three correct 'oxide', one wrong 'Dachshund') and the cluster completeness is also* $3/4$ *('nitric oxide' missed).*

confusion matrix). As test set, we selected 57 entire telephone calls (10 hours) from the Fisher database (noisy, conversational speech), so that each call is mainly centered around a particular topic and is thus likely to contain reoccurring topic-related rare words. We substantially reduced the vocabulary size to the 2860 most frequent words (unigram probabilities from LM training), which resulted in 6.1% OOV rate on the test set. This is artificially high (also due to call selection), but since our primary focus was not OOV detection, it provided us with sufficient OOVs for the development of an OOV similarity metric. Table 1 shows the OOV detection. We preferred operating points with high OOV detection precision, since we assume, that this results in a higher clustering accuracy.

5. Results

Table 2 evaluates the similarity measure for our two tasks. An OOV candidate is targeted, if it is overlapping with an OOV, and if that OOV reoccurs in the overlaps of other OOV candidates. The *recall of reoccurring OOVs* is the percentage of targeted candidates, for which we retrieve at least one correct example of the reoccurring OOV.

When retrieving other OOV candidates in query-by-example mode (task 1), we compute two measures for each OOV candidate (see example in table 3):

- *Cluster accuracy* is the percentage of clustered words, that are correct (i.e. have a significant overlap in a word with a common stem).
- *Cluster completeness* is the percentage of OOV candidates with the same reference word that were retrieved (is not applied to false alarms).

Thus, according to table 2, on average, 40-60% of the resulting cluster members are correct and the resulting clusters cover 50-60% of the desired other examples.

For task 2, using the alignment scoring as a higher-level OOV description (to find e.g. derivations and compounds), we compute the *OOV description accuracy*, which is the percentage of suggested alignments, that are correct, considering common word stems. According to table 2, on average 40-50% of the suggested alignments are correct.

6. Discussion and conclusions

Finding reoccurring OOVs is a new task, which to our knowledge has not been addressed so far. It is related to query-by-example techniques [10], which try to spot other occurrences of a keyword given one or more examples of it (audio snippets or phonetic representations), but especially deals with partial detections.

We presented a similarity measure for detected OOVs based on alignment error. Through alignment, we also try to reveal more complex relations like composed and derived words. We think, that systems dealing with an unlimited vocabulary should be able to make use of such word relations and e.g. not fail, simply because one inflected form was not seen during training. The sub-word (multiphone) based approach to OOV detection seems to be compatible with that idea, since it also composes words from common sub-words observed during training.

The introduced methods are rather simple and represent 'a first shot'. We want to apply the techniques to the TED database, using a vocabulary that was not artificially limited. Instead of just using the one-best output of the ASR, we want to switch to using ASR lattices and a more sophisticated (probabilistic or fuzzy) clustering algorithm. Building on that, we could either train the error model in maximum likelihood sense on training lattices, or apply query-by-example or keyword spotting techniques on lattices.

7. References

[1] Burget, L. et al, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs", ICASSP, 2008.

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

[3] Bazzi, I. and Glass, J. R., "A Multi-Class Approach for Modelling Out-of-Vocabulary Words", Proc. ICSLP, 2002.

[4] Bisani, M. and Ney, H., "Open vocabulary speech recognition with flat hybrid models", Interspeech, 2005.

[5] Szöke, I., Fapšo, M., Burget L., Černocky, J., "Hybrid word-subword decoding for spoken term detection", Proc. SSCS 2008: Speech search workshop at SIGIR, 2008.

[6] Rastrow, A. et al, "Towards Using Hybrid Word and Fragment Units for Vocabulary Independent LVCSR Systems", Interspeech, 2009.

[7] Mohri, M., Pereira, F. C. N. and Riley, M., "Speech recognition with weighted finite-state transducers", Heidelberg, Germany: Springer-Verlag, 2008.

[8] White, C. et al, "Confidence Estimation, OOV Detection and Language ID using Phone-to-Word Transduction and Phone-Level Alignments." ICASSP, 2008.

[9] Bisani, M. and Ney, H., "Joint-Sequence Models for Grapheme-to-Phoneme Conversion", Speech Communication, vol. 50, no. 5, pp. 434-451, 2008.

[10] Shen, W., White, C. and Hazen, T., "A comparison of query-by-example methods for spoken term detection", Interspeech 2009.