COPING WITH OUT-OF-VOCABULARY WORDS:
OPEN VERSUS HUGE VOCABULARY ASR

Matteo Gerosa and Marcello Federico

FBK-irst - Fondazione Bruno Kessler
Via Sommarive 18, 38100 Povo (TN), Italy
`http://hlt.fbk.eu - surname@fbk.eu`

ABSTRACT

This paper investigates methods for coping with out-of-vocabulary words in a large vocabulary speech recognition task, namely the automatic transcription of Italian broadcast news. Two alternative ways for augmenting a 64K(thousand)-word recognition vocabulary and language model are compared: introducing extra words with their phonetic transcription up to 1.2M (million) words, or extending the language model with so-called graphones, i.e. subword units made of phone-character sequences. Graphones and phonetic transcriptions of words are automatically generated by adapting an off-the-shelf statistical machine translation toolkit. We found that the word-based and graphone-based extensions allow both for better recognition performance, with the former performing significantly better than the latter. In addition, the word-based extension approach shows interesting potential even under conditions of little supervision. In fact, by training the grapheme to phoneme translation system with only 2K manually verified transcriptions, the final word error rate increases by just 3% relative, with respect to starting from a lexicon of 64K words.

Index Terms— Automatic Speech Recognition, Open-vocabulary speech recognition, OOV words

1. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) systems operate with a fixed and finite vocabulary, with typical vocabulary sizes in the order of 60K-100K word forms. In open vocabulary applications (e.g. broadcast news transcription) these systems often encounter words not included in the recognition vocabulary, also called out-of-vocabulary (OOV) words. OOV words are a significant source of errors in modern automatic speech recognition (ASR) systems. Not only the ASR system replaces an OOV word with a phonetically similar word/words, but often the error spreads to the nearest words [1]. In addition, OOV words are often named-entities and can be key-words for applications such as spoken document retrieval.

Recently, several works have proposed solutions to deal with OOV words. Some of these works propose approaches that use a combination of $weakly$ (i.e. phone) and $strongly$ (i.e. word) constrained recognizer to detect OOV and overcome vocabulary limitation [2, 3, 4]. Other works adopt an open-vocabulary recognition approach, using a hybrid language model (LM) containing words and subword units [5, 6, 7]. The most used subword units in OOV recognitions are graphones [8].

A graphone is a pair of a letter sequence and a phoneme sequence of possibly different lengths. Graphones are widely used in OOV recognition since they allow immediate conversion from phones to word. Several works [6, 7, 9] investigated the use of graphone units to extend a vocabulary with 20K-64K words. While this is a typical vocabulary size for LVCSR, current state-of-the art systems can handle much larger vocabularies.

In this work we first present a method to automatically generate word pronunciations (lexicons) starting from a manually created training lexicon. This method is based on a simple adaptation of an off-the-shelf statistical machine translation toolkit, to generate both a phone pronunciation and the most likely graphone sequence for each word. The proposed method achieves results similar to those obtained with the joint-multigram model presented in [8]. Hence, we compare the approach described in [6], namely to recognize OOV words by means of a hybrid word/graphone LM, with the simpler alternative of extending the recognition vocabulary up to 1.2M words by automatically extending the corresponding lexicon. Finally, we also considered the hypothesis of bootstrapping the grapheme-to-phoneme translation tool from a small supervised lexicon of 2k-5k words, showing in this way that the proposed method is robust even under limited training data conditions.

The paper is organized as follows. Section 2 describes our grapheme-to-phoneme translation system and presents accuracy results on our Italian lexicon. Section 3 describes the experimental framework and details about corpora used to train the FBK-irst speech recognition system. Section 4 presents results of speech recognition experiments employing automatically generated word pronunciations together with word- and graphone-based language models. Section 5 reports experiments carried out by bootstrapping grapheme-to-phoneme translation from small manually verified lexicons. Discussions and conclusions are reported in Section 6.

2. GRAPHEME-TO-PHONEME CONVERSION

Phonetic transcription of words can be addressed as a grapheme-to-phoneme machine translation task: a ”source” string of characters is translated into a corresponding ”target” string of phonemes. Thus, we employed a publicly available phrase-based statistical machine translation toolkit, called Moses [10]. Hence, phrase-pairs of grapheme-phoneme strings were automatically learned from a training lexicon by applying the procedure reported in the documentation of Moses [1]. Besides the phrase-table, a target phoneme-based n-gram language model was estimated, as well as feature functions that penalize overlong or too short translations. The combination of

This work was partially financed by the Province of Trento, under the project RIVA.

[1] See `http://www.statmt.org/moses/`

all such components was finally optimized with a minimum error training step. Finally, translations were computed with a monotonic decoding setting to inhibit word re-ordering between source and target positions. Graphones are derived from the best translation hypothesis by recovering the corresponding phrase-pairs used by the Moses decoder.

Grapheme-to-phoneme conversion was tuned and tested on a baseline 64K-word lexicon that was developed semi-automatically for the FBK-irst Italian broadcast news system [11]. As a preprocessing we removed words with letters other than A-Z and apostrophes. We retained multiple pronunciations of words. The lexicon was randomly split into a training, development and test set (see Table 1 for details).

Total words	63,786
Training words	40,658
Development words	10,279
Testing words	12,849
Transcriptions per word	1.007

Table 1. Statistics and partitions of the Italian lexicon.

Phrase table and target LM of the translation system were estimated on the training portion of the baseline lexicon. We tested performance using phone-based n-gram LMs with order ranging from 5 to 7. Optimal phrase (or graphone) maximum length was found within the range 3 to 6. We evaluated results using the phone-error-rate (PER), computed with respect to the closest pronunciation variant available in the baseline lexicon. (Notice that only 458 words contain multiple phonetic transcriptions.) Tables 2 reports PER results achieved on the test portion of the lexicon.

LM order	Maximum graphone length			
	3	4	5	6
5-grams	1.78%	1.73%	1.72%	1.72%
6-grams	1.76%	1.73%	1.73%	1.73%
7-grams	1.77%	1.73%	1.73%	1.72%

Table 2. Phone-error-rate (PER) of grapheme-to-phoneme conversion of Italian words through statistical machine translation. PERs are reported for different target language model orders and maximum phrase lengths used by the phrase-based decoder.

It seems that for the Italian language n-gram order and maximum graphone length have almost no influence on translation performance. This is probably due to the fact that for Italian word pronunciation is very close to its spelling, with only few exceptions depending on the context in which letters appears. For the sake of comparison, we trained and tested on the same data the grapheme-to-phoneme tool based on the joint-multigram model described in [8], obtaining comparable results (1.40% PER). From this preliminary analysis, we decided to use in the following experiments a 5-gram LM and graphones/phrases of maximum length 4.

3. EXPERIMENTAL FRAMEWORK

The test bed for our investigation is the automatic transcription of Italian broadcast news. Acoustic models (AMs) were trained on about 130 hours of manually and automatically transcribed speech [11]. LMs were estimated on 600M word corpus including newswire and newspaper articles. The LM corpus contains about 1.2M unique words. Finally, test data consists in 21 TV news shows recorded during March-April 2008. Table 3 reports detailed statistics about these speech corpora.

	Train	Dev	Test
# hours	129h:30m	3h:46m	3h:29m
# utterances	115,024	544	470
# words	1,228,000	40,701	33,181
Dictionary	-	7,652	6,321

Table 3. Statistics of speech corpora used for ASR experiments.

Experiments were carried out using the FBK-irst speech recognition system [12]. In particular, feature extraction embeds cepstral mean subtraction, variance normalization, and projection of acoustic features, based on Heteroscedastic Linear Discriminant Analysis. Finally, acoustic data are normalized using Constrained MLLR-based Speaker Normalization [13]. For acoustic models we employ state-tied, cross-word triphone HMMs. Output distributions associated with HMM states are modeled with mixtures of up to 16 diagonal covariance Gaussian densities.

4. OPEN VS HUGE VOCABULARY ASR

ASR experiments carried out with the aim of reducing the influence of OOV words. Grapheme-to-phoneme translation was trained on the 64K-word baseline lexicon, described in Section 2, and used to expand the baseline vocabulary in various ways.

1. Following a procedure similar to the one reported in [6], we can replace all OOV words in the LM training corpus with their most likely graphone sequence. The recognition vocabulary is then augmented with all graphones inferred by this procedure and the modified text is used to train a hybrid word-graphone LM.
2. A second approach is to extract the list of OOV words from the LM training corpus and generate pronunciations for each word. All OOV words are then added to the recognition lexicon and then a conventional word-based LM is trained using the extended vocabulary.
3. A mixed strategy for representing OOV words is also investigated. We select a subset of the most frequent OOV words for which we generate an automatic pronunciation, while the rest is replaced in the LM training corpus with their most likely graphone sequence. OOV words and graphones are added to the recognition vocabulary and a hybrid word-graphone LM is trained like in the first strategy.

Two 4-gram word-based LMs were trained: one using the 64K-word baseline lexicon (“64K”) and one using an extended lexicon using all words in the training corpus, about 1.2M words, 1,136K of which requiring automatic phone transcriptions. We also trained three 4-gram hybrid word-graphone LMs, with a vocabulary of 64K words, 128K words and 256K words, each corresponding to the most frequent words in the LM training corpus. The total number of different graphones was 8,799, 8789 and 8,499 for the 64K, 128K and 256K word vocabulary, respectively. All 4-gram LMs were trained using improved Kneser-Ney smoothing.

Table 4 reports the number of OOV words and OOV rates on the development and test sets for the different vocabulary sizes. We can see that very low OOV rates are achieved with the 1.2M-word vocabulary.

LM	Dev		Test	
	# OOV	OOV%	# OOV	OOV%
64K	794	1.95%	572	1.72%
128K	399	0.98%	310	0.93%
256K	211	0.51%	134	0.40%
1.2M	103	0.25%	69	0.21%

Table 4. OOV statistics of language models with increasing vocabulary sizes.

Table 5 reports recognition results, in terms of word error rate (WER) achieved on the development and test sets using the five aforementioned LMs. Notice that prior to the evaluation process we converted graphones to their constituent letters and merged adjacent graphones to form word hypotheses .

LM	Dev	Test
64K	21.52%	18.30%
64K+graphones	20.97%	17.58%
128K+graphones	20.65%	17.31%
256K+graphones	20.55%	17.25%
1.2M	19.74%	16.64%

Table 5. Word-error-rate results of the speech recognizer when using language models of increasing vocabulary size, possibly including graphones.

The test corpus is quite challenging, given the presence of background music, disfluent speech and noise, which led to relatively high WERs. It can be noted that there is a significant difference between performance achieved with the 64K word LM and with the 1.2M word LM. Relative reduction in WER between the two LMs is about 8-9%, on both data sets. Hybrid word/graphone LMs perform slightly better than the baseline 64K word LM. In particular, the "64K+graphones" LM allows a 2.6% and 4% relative reduction in WER with respect to the "64K' LM, on dev and test sets, respectively.

We can conclude that generating automatic transcriptions of OOV words and adding them to the recognition vocabulary produces better performance than using a hybrid word-graphone LM. In fact, WER decreases in both data sets as the dictionary size increases. However, we have to remember that the hybrid word-graphone LMs are able to recognize previously unseen words, while the 1.2M word LM is in fact a closed-vocabulary LM. For the sake of comparison, we also tested the 1.2M-word LM with a lexicon generated with a joint-multigrams model also trained on the baseline lexicon. Recognition performance on the dev and test set were only slightly better: 19.67% and 16.56%, respectively.

In addition to WER results, we are also interested in the ability to recover OOV words. Table 6 reports the precision, recall and F-score of the hybrid LMs and of the 1.2M word LM when recognizing OOV words on the test set. Results are computed considering OOV words with respect to the 64K word vocabulary.

All LMs show very high precision values, meaning that there are almost no insertions or substitutions of OOV words. Recall values achieved by the hybrid LMs is quite low, with about a third of the OOV words correctly recognized. Remarkably, the 1.2M-word LM shows the best recall, i.e. 0.43.

LM	Precision	Recall	F-score
64K+graphones	0.99	0.32	0.48
128K+graphones	0.99	0.36	0.53
256K+graphones	0.99	0.36	0.53
1.2M	0.98	0.43	0.60

Table 6. Precision, Recall and F-score of OOV words on the test set, with respect to the 64K words vocabulary.

5. EXPERIMENTS WITH LIMITED RESOURCES

When training a recognition system on a new language with scarce language resources, a common bottleneck is usually the size of the pronunciation lexicon. Thus, we investigated the performance of the grapheme-to-phoneme translation tool with respect to the amount of available training data.

The graphene-to-phoneme translation tool was trained on a subsample of the baseline lexicon described in Section 2. We considered two sample sizes, 2K and 5K words. Two sampling methods were also considered:

- **Frequency-based sampling**: the words with the highest frequency in the LM training corpus are selected. This approach guarantees to have the highest number of words with correct (manual) transcriptions. On the contrary it does not guarantee any good phonetic coverage.
- **Random-based sampling**: words are selected randomly. This approach shows to provide better phonetic coverage, but it can possibly produce wrong transcriptions for very frequent words.

We generated one lexicon for each size and sampling method, leading to 4 different lists. A grapheme-to-phoneme translation tools were trained on each vocabulary. Table 7 reports performance, in term of PER, achieved with the 4 systems on a 10K words subset of the baseline lexicon, which is disjoint from the 4 training sets, of course. PER was again computed with respect to the closest pronunciation variant.

Sample size	Sampling criterion	
	Random	Frequency
2K words	3.26%	4.16%
5K words	2.96%	3.74%

Table 7. Phone-error-rates by grapheme-to-phoneme decoders trained on different sub-samples of the baseline lexicon.

Random-based sampling shows to provide better transcription accuracy that than frequency-based sampling. Notice however that the used test set is a list of words that does not reflect frequency properties of words.

As expected, results obtained with a 5K word training vocabulary are better than the ones achieved with a 2K word training vocabulary. PERs are however significantly worse than those achieved with the system described in Section 2.

These four systems were then used to transcribe all the words contained in the LM training corpus and not included in their respective training vocabulary, generating an automatic pronunciation for each word. The four 1.2M-word vocabularies thus obtained were used as recognition vocabularies for ASR experiments. The IBN corpus (see Table 3) was also transcribed with each vocabulary, and then used to train AMs with the same procedure described in Section 3. Table 8 reports recognition results, in terms of WER, achieved with the four trained ASR systems. As a reference, results achieved with the 1.2M vocabulary, generated with the system trained on the 64Kword baseline lexicon, are also reported (see Table 5, last row).

Bootstrap vocabulary	Dev	Test
2K-random	20.54%	17.62%
2K-frequency	20.26%	17.13%
5K-random	20.36%	17.70%
5K-frequency	20.01%	17.18%
64K	19.74%	16.64%

Table 8. Word-error-rates achieved with a 1.2M word LM using automatic phonetic transcriptions learned from small sub-samples of the baseline lexicon.

Results are to our view interesting: with only 2K manually checked word pronunciations, we can develop a system that performs better than one using a 64K-word LM relying on a manually verified lexicon (see Table 5, first row). With respect to the sampling procedure, we see opposite results with respect to the phone-error-rates. As remarked, this is mainly due to the fact that the frequency-based sampling guarantees to use correct transcriptions of very frequent words. Results obtained with 5K lexicon are similar to those obtained with a 2K lexicon on the test set, and just slightly better on the dev set. Remarkably, these WERs are just slightly worse, 1.5% to 3% relative, than those achieved after training the grapheme-to-phoneme system on the 64K-word baseline lexicon, reported in Table 5.

6. CONCLUSIONS

In this paper, we investigated different approaches to improve OOV word recognition in a large vocabulary speech recognition task. First, we presented a simple method to automatically generate word pronunciations and graphone sequences starting from a manually checked 64K-word lexicon. We used this method to extend the 64K-word baseline recognition vocabulary up to 1.2M words (i.e. the entire LM training corpus vocabulary), with automatically generated phonetic transcriptions. We then used the new augmented vocabulary to train a conventional word-based 4-gram LM. We compared this approach with an approach similar to the one presented in [6], that is to create an open-vocabulary recognition by replacing all OOV words in the LM training corpus with their most probable graphone sequence and training with the modified text a hybrid word-graphone LM.

Recognition experiments were carried out on an Italian broadcast news task. Results showed that the augmented vocabulary produces lower error rates than representing OOV words with graphones. In addition, although the 1.2M word recognition vocabulary is finite, it resulted in very low OOV rates of about 0.2% and better OOV word recognition than then hybrid word-graphone LM. Experiments carried out assuming little supervised lexicons to train grapheme-to-phoneme translation systems, suggested that very competitive performance can be achieved with only 2K words.

While these result is very promising, there are still open issues that we want to investigate. First, to overcome the limitations of closed-vocabulary systems we will consider LMs combining a huge vocabulary with graphone units. Moreover, we will evaluated the proposed approaches on the English language, which has quite different characteristics than the Italian language.

7. REFERENCES

[1] I. Bazzi, "Modelling Out-of-Vocabulary Words for Robust Speech Recognition," *Ph. D. Thesis, Department of Eelectrical Engeneering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA*, 2002.

[2] H. Ketadbar, M. Hannemann, and H. Hermansky, "Detection of out-of-vocabulary words in posterior based asr," in *Proc. of INTERSPEECH*. 2007, ISCA, 1757-1760.

[3] L. Burget et al., "Combination of strongly and weakly constrained recognizers for reliable detection of oovs," in *Proc. of ICASSP*. 2008, IEEE.

[4] N. Bertoldi, M. Federico, D. Falavigna, and M. Gerosa, "Fast speech decoding through phone confusion networks," in *Interspeech 2008*. 2008, ISCA, To appear.

[5] A. Yazgan and M. Saraclar, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition," in *Proc. of ICASSP*, Montreal, Canada, May 2004, vol. 1, pp. 745–748.

[6] M. Bisani and H. Ney, "Open Vocabulary Speech Recognition with Flat Hybrid Models," in *Proc. of INTERSPEECH*, Lisboa, Portugal, Sept. 2005, pp. 725–728.

[7] K. Vertanen, "Combining open vocabulary recognition and word confusion networks," in *Proc. of ICASSP*. 2008, pp. 4325–4328, IEEE.

[8] M. Bisani and H. Ney, "Investigations on Joint-multigram Models for Grapheme-to-Phoneme Conversion," in *Proc. of ICSLP*, Denver, CO, Sep. 2002, pp. 105–108.

[9] M. Akbacak, D. Vergyri, and A. Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition system," in *Proc. of ICASSP*. 2008, IEEE.

[10] P. Koehn et al., "Moses: Open source toolkit for statistical machine translation," in *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proc. of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180.

[11] F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani, "Advances in automatic transcription of italian broadcast news," in *Proc. of ICSLP*, Beijing, China, Oct. 2000, pp. 660–663.

[12] Fabio Brugnara et al., "The itc-irst transcription systems for the tc-star-06 evaluation campaign," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 117–122.

[13] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization.," *Computer Speech and Language*, vol. 20, no. 1, pp. 107–123, Jan. 2006.