Handling OOV Words In Arabic ASR Via Flexible Morphological Constraints

Nguyen Bach, Mohamed Noamany, Ian Lane, and Tanja Schultz

InterACT, Language Technologies Institute
School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nbach, mfn, ian.lane, tanja}@cs.cmu.edu

Abstract

We propose a novel framework to detect and recognize out-of-vocabulary (OOV) words in automated speech recognition (ASR). In the proposed framework a hybrid language model combining words and sub-word units is incorporated during ASR decoding then three different OOV words recognition methods are applied to generate OOV word hypotheses. Specifically, dictionary lookup, morphological composition, and direct phoneme-to-grapheme. The proposed approach successfully reduced WER by 1.9% and 1.6% for ASR systems with recognition vocabularies of 30K and 219K. Moreover, the proposed approach correctly recognized 5% of OOV words.

Index Terms: speech recognition, out-of-vocabulary words recognition, subword language modeling

1. Introduction

One key problem in Automatic Speech Recognition (ASR) is the detection and recognition of out-of-vocabulary (OOV) words that do not occur in the ASR vocabulary. During system development words that do not occur frequently in the training corpora are typically removed from the ASR vocabulary in order to reduce model size and complexity. However, these words are often critical for spoken language understanding as they contain key information. Such information is vital to realize effective information extraction and retrieval from multimedia data [5]. Current state-of-the-art ASR systems, however, cannot recognize words which are not contained within their recognition vocabulary. Furthermore, recognition errors due to OOV words typically also induce errors in neighboring words.

Performing recognition with increasingly large recognition vocabularies, however, is not an appropriate solution. Regardles of the size of the recognition vocabulary it cannot provide coverage over all possible words, some of which may be truely novel. Second, increasing vocabulary size significantly increases the complexity of the ASR system and thus cannot be applied systems are targeted for portable devices such as pocket PC, PDA, smart phone, and laptops. To realize effective natural language applications and close to real-time performance, handling out-of-vocabulary words is crucial.

To overcome these problems previous works have focused on explicitly modeling OOV words during ASR as a sequence of subword models. One approach is to model with an all-phone (generic word) such as described in [2], [11], and [8]. However, these approaches heavily depend on the accuracy of the phone recognizer. One extension of this approach is to model OOV words as graphone-based models [6], [1].

Related works including [4] and [5] attempt to correct ASR errors by using offline monolingual corpora and information retrieval techniques. These approaches successfully reduced WER and CER but the whole system runs completely out of the ASR system. These approaches however are not robust as they require both accurate ASR confident scores, and minimal errors in 1-best ASR hypothesis.

To improve the robustness of OOV recognition we propose an OOV recognition framework for Arabic. First, we transform conventional language models to hybrid-language models where words are seen in the language model training data but not in the ASR vocabulary are expressed as a sequence of subword units (phoneme or syllable). Next, OOV words are recognized via an iterative back-off scheme where external knowledge sources are used to apply an increasingly weaker set of constraints. Specifically, three methods are applied dictionary lookup, morphological composition, and direct phoneme-to-grapheme (P2G) conversion. The proposed framework obtained significant improvements over a strong baseline for large vocabulary Arabic broadcast news transcription system.

2. Proposed Framework

We first perform ASR decoding using a hybrid language model (LM) to get the 1-best hybrid hypothesis. If phone sequences are detected in the hypothesis we use three different methods to recognize OOV words. Each method applies progressively weaker constraints on the phoneme sequence as shown in Figure 1.

2.1. Generating hybrid hypothesis

To generate graphone-level hypotheses of OOV words we perform ASR using a hybrid LM similar to the approach described in [1]. One significant difference compared to previous works however is to explicitly map all OOV words which appear in the LM training data but do not appear in the ASR vocabulary to subword units. This contrasts to the approach applied in [1] which modelled OOV words by subword units of the low frequency words in the ASR vocabulary. This scheme generates a hybrid LM containing both of word and subword units. We train the LM, include subword units the vocabulary, and add their pronunciations to pronunciation dictionary. Finally the system performs decoding using the tradition decoding method. Table 1 shows an example of the baseline system and the proposed hybrid LM system associated with the reference transcription.

Table 1: *An example for OOV words detection.*

Reference	... bMnh **syHAkm** *gyAbyA* ...
Baseline	... bMnh **syHAsb** *gyAbyp* ...
Hybrid Hyp.	... bMnh **[s] [y] [H] [A] [k] [m]** *gyAbyA* ...

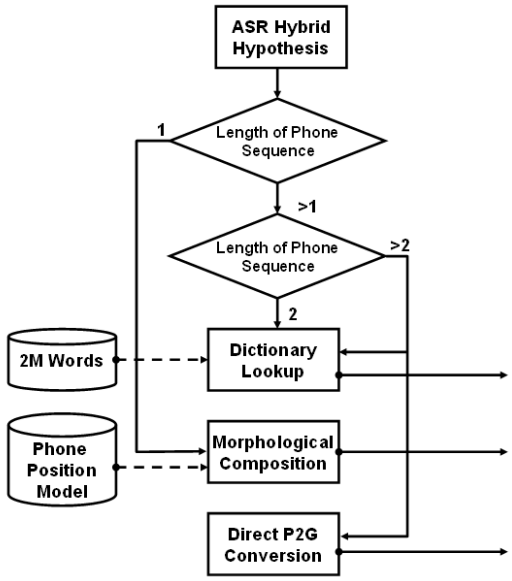

Figure 1: *OOV words recognition framework*

2.2. OOV words & recognition

In this paper we explicitly focus on OOV word recognition for Arabic. Arabic is a high inflectional morphology language [9] and this causes several issues for ASR such as more OOV words found in unseen data, more low frequency word in language model training data, and lower token-type ratios. Even with a 219K vocabulary ASR system, we identified that 5.5% ASR vocabulary appears less than 6 times in 600 million words language model training data. Moreover, given a Arabic stem there are theoritically about 490K ways to form a new word by combining prefixes and suffixes. All of these clues suggest that 1) mapping all words with frequency less than 6 may degrade ASR accuracy and 2) to effectively recognize OOV words morphology analysis should be applied.

We propose three methods to recognize OOV words from phoneme sequence for Arabic. Methods can be individually applied to recognition OOV words, or can be jointly applied in a framework showed in Figure 1.

2.2.1. Dictionary lookup

The first method is dictionary lookup. Given a phoneme sequence with length larger than 1 we convert the sequence into a word. If the new word appears in a very large dictionary with 2 million entries we keep the word otherwise we discard it. The dictionary is generated from Arabic Gigaword corpus [1] without any cutoff frequency.

[1]LDC2006T02: Arabic Gigaword Second Edition

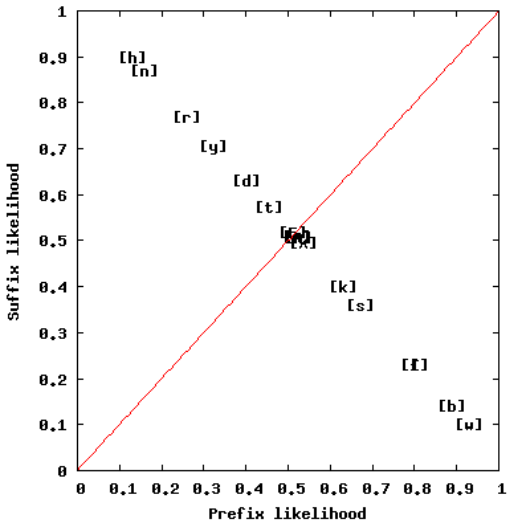

Figure 2: *Phone Position Likelihood*

2.2.2. Morphological composition

In the case that an OOV hypothesis consists of only a single phone, we would like to know whether we should discard, or combine it with the following or previous word. This situation is handled by morphological composition.

Choosing which phoneme should be combined with the following/previous word or discard it based on the probability of that phoneme when it appears as the prefix/suffix phoneme of an OOV word. A phone position model is learned from the LM training data where we compute the likelihood of each individual phone appears as affixes of an OOV words. Figure 2 shows that the model learnt phones such as [w] and [b] are likely to appear as the prefix of an OOV word but not as suffix, while [h] and [n] are more likely to appear as suffix. Phones like [t] and [k] do not appear as either prefix or suffix. Arabic linguistic knowledge also supports these results.

The phone position model with a threshold leads to morphological composition rules. For example if we see individual phone [w] or [h] in the hypothesis and we find that the combination of [w] with the next words or [h] with the previous word result in a truly novel word, then we should keep the new word in the hypothesis.

2.2.3. Direct phoneme-to-grapheme conversion (DirectP2G)

The third method is direct phoneme to grapheme conversion. Our analysis of OOV words showed that 1 or 2 phoneme long words are rare. Therefore, given a phoneme sequence with length larger than 2 we convert the sequence into a word and keep the word in the hypothesis, otherwise we discard. These length constraints match with the Arabic morphological constraints suggested in [10].

3. Experimental Evaluation

In this section, we evaluate the performance of our proposed framework in term of the word error rate. We evaluated the performances of the baseline system, a system using the approach described in [1], and the proposed OOV recognition framework.

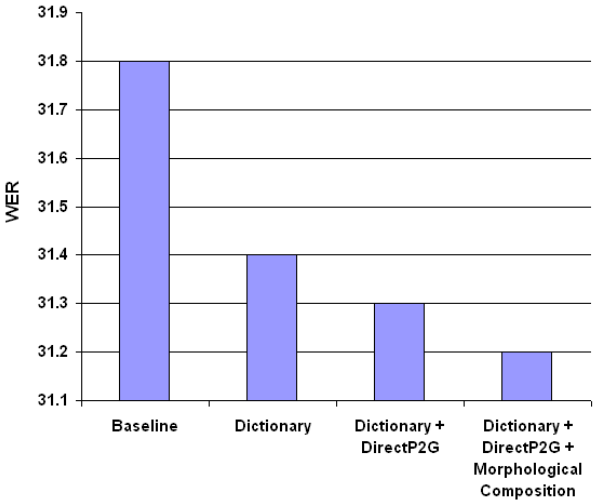

Figure 3: *WER in Small Vocabulary System*

3.1. Evaluation system setup

Speech recognition was performed using an Arabic unvowelized ASR system developed using JANUS toolkit with the IBIS decoder [12]. For the baseline system, a system similar to that described in [3] was applied. Both the acoustic and language models were trained on broadcast conversation, broadcast news, and newspaper domain corpora. We use 90 million words to train a 4-gram language model with modified Kneser-Ney smoothing by using SRI LM Toolkit [7]. Romanization of Arabic words was performed using the BAMA toolkit [2]. The RT04 Arabic evaluation with 3,724 word types and 46,634 word tokens is used as the test set.

3.2. Effectiveness for small vocabulary ASR

First, we evaluated the performance of the proposed framework when applied to a small vocabulary (30K) Arabic ASR system. For the baseline system the most frequent 30K words in the training corpora were selected as the ASR vocabulary. The OOV rate is 10.9%.

In Figure 3 we compare the performance of four systems namely 1) the baseline, 2) combination of baseline with dictionary lookup, 3) combination of baseline with dictionary lookup and direct P2G, 4) combination of baseline with dictionary lookup, direct P2G, and morphological composition. These systems obtain 31.8%, 31.4%, 31.3% and 31.2% WER respectively. For the best system we achieve 1.9% relative WER reduction. The results also show that using the combination of baseline with morphological composition, dictionary lookup and direct P2G the system reaches the best performance.

Figure 4 shows an example of the reference, the output of the baseline system, the output of the hybrid language model, and the final hypothesis when we apply dictionary lookup and the direct P2G method. The baseline hypothesis mis-recognized word **"syHAkm"** with **"syHAsb"**, however our system can detect and correctly recognize this word by correcting phoneme sequence **"sb"** to **"km"**. Moreover, the system also correct the

[2]LDC2004L02: Buckwalter Arabic Morphological Analyzer version 2.0

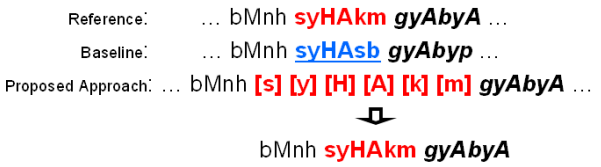

Figure 4: *Dictionary lookup and Direct P2G*

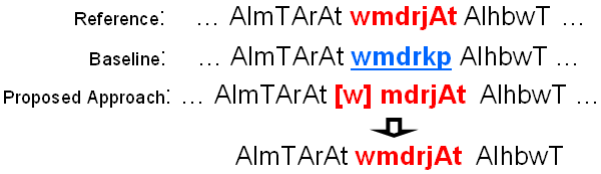

Figure 5: *Morphological composition*

errors surround the OOV word. Here, **"gyAbyp"** in baseline turns into **"gyAbyA"** in our proposed system. In these methods we only consider phoneme sequences with length larger than 2.

Figure 5 shows an example of the reference, the output of the baseline system, the output of the hybrid language model, and the final hypothesis when we apply the morphological composition. This example shows an interesting situation when **"wmdrjAt"** is an OOV word but **"mdrjAt"** is a known word. Due to the inflectional morphology **"wmdrjAt"** becomes an OOV word then the baseline system mis-recognized it as **"wmdrjkp"**. However, our hybrid language model hypothesis output **"[w] mdrjAt"** and by applying the morphological composition our system can correctly recognize the new word **"wmdrjAt"**.

3.3. Effectiveness for large vocabulary ASR

The large system is a 219K vocabulary Arabic ASR with 2.7% OOV rate. In Figure 6 we compare the performance of four systems with large vocabulary. The baseline has 28.3% WER while the system with algorithm described in [1] increase WER to 28.8%. However, the combination of baseline with morphological composition, dictionary lookup, and direct P2G obtains 28.1% WER. For the best system we gain 0.7% relative WER reduction. To address the robustness of the proposed framework we increase the size of language model from 90M to 600M Arabic words. The baseline system obtains 26% WER, while the combination methods reduces WER to 25.6%. We achieve 1.6% relative WER reduction.

4. Discussions & Analysis

We investigated why our approach works. Figure 7 indicates that when we build an ASR system with 30k vocabulary, the vocabulary coverage is 89% and we reach 31.8% WER. Furthermore, if we keep using this system but perform some post-processing applying a dictionary with 2 million entries, the vocabulary coverage increases to 98.9% but the WER only reduces to 31.4%. This phenomenon shows that even when we try to use a very large word list, the ASR system still can not handle truly novel words. Now, if we perform the combination approach the WER reduces to 31.2%. Our

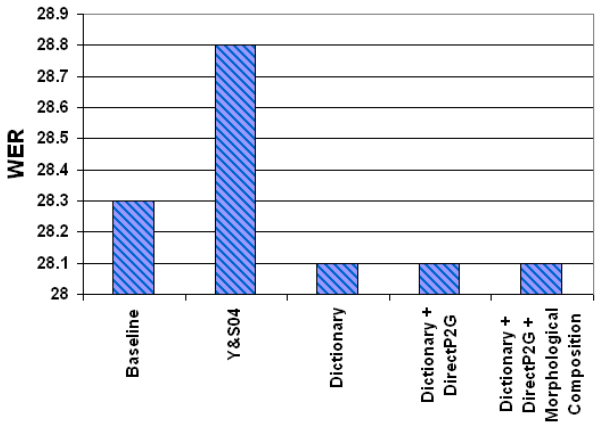

Figure 6: *WER in Large Vocabulary System*

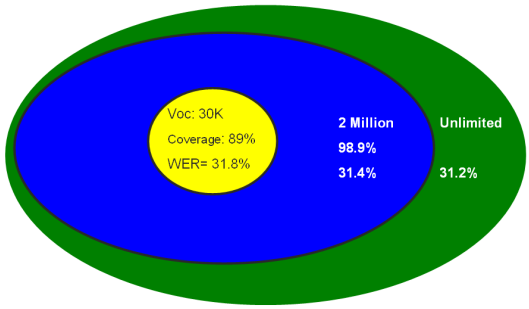

Figure 7: *Correlation of Vocabulary size, Vocabulary Coverage, and WER*

system successfully recognized 5% truly novel OOV words. It means that the proposed approach can create a new word and correctly recognize OOV words. In addition, given a path r in a confusion network N, our framework also can be used to estimate of the appearance of OOV words in path r of a confusion network N.

5. Conclusions

In this paper we proposed a novel approach to handle OOV words in Arabic ASR. To improve the performance we extended the hybrid language model to estimate OOV words via subword units, and incorporating three methods to recognize OOV words and correct errors made by misrecognizing OOV words. We also showed that our approach can give improvements in ASR performance for both small and large vocabulary ASR systems.

6. Acknowledgements

This work is in part supported by the US DARPA under the GALE (Global Autonomous Language Exploitation) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. References

[1] Yazgan, A. and Saraclar, M., "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition", Proc. Int. Conf. of Acoustics, Speech, and Signal Processing, Canada, May 2004, vol. 1, pp. 745-748.

[2] Bazzi, I. and Glass, J., "Modeling out of vocabulary words for robust speech recognition", Proc. Int. Conf. on Sposken Language Processing, Bejing, China, 2000.

[3] Noamany, M., Schaaf, T., and Schultz, T., "Advances in the CMU/InterACT Arabic GALE transcription system", to appear in Proc. of Human Language Technology, Rochester, New York, USA, 2007.

[4] Huang, F., Vogel, S., and Waibel, A., "Towards named entity extraction and translation in spoken language translation", Proc. of the Human Language Technology, Boston, USA, 2004.

[5] Favre, B., Bechet, F., and Nocere, P., "Robust named entity extraction from large spoken archives", Proc. of the Emperical Methods in Natural Language Processing, Vancouver, Canada, 2005.

[6] Bisani, M. and Ney, H., "Open Vocabulary Speech Recognition with Flat Hybrid Models", Proc. of the European Conf. on Speech Communication and Technology, pp. 725-728, Lisbon, Portugal, September, 2005.

[7] Stolcke, A., "SRILM – An Extensible Language Modeling Toolkit", Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901-904, Denver, USA, 2002.

[8] Schaaf, T., "Detection Of OOV Words Using Generalized Word Models And A Semantic Class Language Model", Proc. of the European Conf. on Speech Communication and Technology, Aalborg, Denmark, 2001.

[9] Zollmann, A., Venugopal, A., and Vogel, S., "Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation", Proc. of the Human Language Technology, New York, USA, 2006.

[10] Bing, X., Kham, N., Long, N., Richard, S., and John, M., "Morphological Decomposition for Arabic Broadcast News Transcription", Proc. Int. Conf. of Acoustics, Speech, and Signal Processing, Toulouse, France, 2006.

[11] Bazzi, I. and Glass, J., "A Multi-Class Approach for Modelling out-of-Vocabulary Words", Proc. Int. Conf. on Sposken Language Processing, pp. 1613-1616, Denver, USA, 2002.

[12] Soltau H., Metze F., Fugen C., and Waibel A., "A one pass-decoder based on polymorphic linguistic context assignment", Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.