MORPH-TO-WORD TRANSDUCTION FOR ACCURATE AND EFFICIENT AUTOMATIC SPEECH RECOGNITION AND KEYWORD SEARCH

A. Ragni, D. Saunders, P. Zahemszky, J. Vasilakes, M. J. F. Gales, K. M. Knill

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{ar527, ds636, pz251, jav39, mjfg, kate.knill}@eng.cam.ac.uk

ABSTRACT

Word units are a popular choice in statistical language modelling. For inflective and agglutinative languages this choice may result in a high out of vocabulary rate. Subword units, such as morphs, provide an interesting alternative to words. These units can be derived in an unsupervised fashion and empirically show lower out of vocabulary rates. This paper proposes a morph-to-word transduction to convert morph sequences into word sequences. This enables powerful word language models to be applied. In addition, it is expected that techniques such as pruning, confusion network decoding, keyword search and many others may benefit from word rather than morph level decision making. However, word or morph systems alone may not achieve optimal performance in tasks such as keyword search so a combination is typically employed. This paper proposes a single index approach that enables word, morph and phone searches to be performed over a single morph index. Experiments are conducted on IARPA Babel program languages including the surprise languages of the OpenKWS 2015 and 2016 competitions.

Index Terms— morph-to-word transduction, speech recognition, keyword search, single index

1. INTRODUCTION

Accurate and efficient automatic speech recognition (ASR) and keyword search (KWS) have been a subject of extensive research for many years. Initiatives, such as the IARPA Babel program [1], have also looked at generalisation of those approaches beyond English [2], Mandarin [3] and Arabic [4] languages. Unlike English, many languages are highly inflective and agglutinative so a typical 60 hours speech corpus would yield a high out-of-vocabulary rate (OOV). This causes significant issues to downstream tasks such as KWS where missing a word occurrence is penalised orders of magnitude higher than predicting one if it does not in fact occur [5]. A number of approaches have been proposed to deal with the OOV problem such as web data and subword modelling. The use of web data [6, 7] relies heavily on the presence of a given language on the internet. For instance, there are about 1 million words of Dholuo texts available at the time of writing this paper. Subword modelling addresses the OOV problem by decomposing words into sequences of subword units such as morphs, syllables and phones. These units represent a different tradeoff between the number of units and the scope of modelling. Whereas syllable modelling requires language specific knowledge, morph modelling can be conducted in a completely unsupervised fashion [8, 9]. One standard issue with subword modelling is a rather poor word level performance which is believed to originate from the limited scope of language modelling [10] and, more generally, decision making. A number of approaches have been proposed to address this issue. These include the use of high order n-gram morph language models [10] and syllable transductions [11] for converting syllable sequences into word sequences upon which powerful word language models can be applied.

Although successful, subword systems are often used to complement word systems rather than being a stand alone approach in tasks such as KWS [12]. This significantly increases the cost of deployment. A number of approaches have been proposed to address this issue. One example is a parallel index combining indices generated by a word and different subword systems [13]. This approach requires multiple ASR runs and indices. Another example is the use of mixed word and subword units [14, 12, 15]. This approach is more advantageous, requiring single decoding, index and search only. The drawback of this approach is the large vocabulary, especially in the presence of web data, which would make decoding very slow.

This paper proposes a different approach to achieve a diverse single index approach. The main idea consists of picking a morph index and using various transductions to perform different unit searches on it. The key transduction in this work is between morph and word units. The previous work with syllable transduction proposed a two stage approach requiring a second alignment stage to propagate timing information [11]. This paper proposes a single pass approach utilising a lexicographic semiring [16] to propagate both acoustic scores and time information. In order to ensure that different unit searches can be applied on the morph index it should enable changing posterior probabilities depending on the unit to incorporate new information such as word language model probability. This paper shows that this is possible to achieve by keeping information such as acoustic and language model scores separate and pushing a modified form of index to compute posteriors.

The rest of this paper is organised as follows. Section 2 introduces the single pass morph-to-word transduction. A single index approach is discussed in Section 3. Experiments are presented in Section 4. Finally, conclusions are given in Section 5.

2. MORPH-TO-WORD TRANSDUCTION

In inflective and agglutinative languages morphological decomposition of words into constituent subword units called morphemes play...
an important role. Consider the following morphological decomposition of two English words

\[
\text{looking} \quad \text{look}^+ \quad \text{+ing} \\
\text{talked} \quad \text{talk}^+ \quad \text{+ed}
\]

where plus signs (+) are used to indicate how morphemes can and cannot be joined. A swap of the first morphemes creates two new valid English words. This simple example illustrates the power of morphological modelling. It is often the case that no morphological decomposition is available for a particular vocabulary or language. Unsupervised morphology induction approaches could be used to address this issue [8, 9]. These normally employ an automatic procedure embedding some general considerations such as the size of morph set and the cost of word decompositions.

It has been observed, however, that reducing entropy among morph units helps to improve both ASR and KWS performance. The simplest approach was illustrated earlier where a simple addition of plus signs (+) would cause a significant impact on the nature of decoded morph sequences. Another example would be to use an equivalent of syllable-to-word transduction to collapse morph lattices into word lattices where more constrained word language models could be used. In weighted finite state transducer [17] terminology, such morph-to-word transduction can be expressed as

\[
\mathcal{W} = \mathcal{M} \circ \mathcal{M}_2 \mathcal{W}
\]

where \(\mathcal{M}\) and \(\mathcal{W}\) are morph and word lattices, \(\mathcal{M}_2 \mathcal{W}\) is a morph-to-word transducer which maps morph sequences in \(\mathcal{M}\) into word sequences encoded by \(\mathcal{W}\), epsilon removal and projection on the output label are not shown for simplicity. Figure 1 shows a toy example of these transducers. Not shown on the figure are acoustic log-likelihoods, language model log-probabilities and time information. Note that morph language model probabilities are a placeholder for word language model probabilities to be applied. In order to ensure these quantities are propagated from \(\mathcal{M}\) to \(\mathcal{W}\) all operations must be performed in a form of a lexicographic semiring [16] where multiplication (path continuation) is given by

\[
\langle a_1, l_1, t_1 \rangle \odot \langle a_2, l_2, t_2 \rangle = \langle a_1 + a_2, l_1 + l_2, t_1 + t_2 \rangle
\]

where \(a_1, a_2\) and \(l_1, l_2\) and \(t_1, t_2\) are acoustic log-likelihoods, language model log-probabilities and durations of the first and second transition respectively. The example in Figure 2 shows morph-to-word transduction in the lexicographic semiring for a single path of the morph lattice in Figure 1 (a). Given a word lattice, it is possible to apply word language models to re-rank paths. An alternative approach [11] requires determining word lattices prior to constrained decoding followed by the application of word language models.

### 3. SINGLE INDEX

The approach presented in the previous section enables a single morph decoding to yield two sets of lattices: a morph and word. If used directly two indices would be created. An additional phone index may also be needed to search for keywords not found in either of those indices. This significantly increases the footprint of a keyword search system. An interesting theoretical question is whether one index could be used. By manipulating the index accordingly morph, word and phone searches then could have been performed.

Consider an index [18] in Figure 3 constructed from the morph lattice shown in Figure 1 (a). Note that here weight tuples encode posterior probabilities (multiplied along paths), start time and end time (summed along paths). The output label of the final transitions encodes utterance identity. Such an index cannot be used to perform word search. One issue is morph posterior probabilities that cannot be split into individual acoustic model log-likelihoods and language model log-probabilities. An option would be to encode them separately. However, pushing such an index will not yield valid morph and word posterior probabilities since it has lost the original temporal arrangement of the morphs due to transitions related to the retrieval of individual morphs (1 \(\rightarrow\) 9, 0 \(\rightarrow\) 3, 0 \(\rightarrow\) 4 and 0 \(\rightarrow\) 5).

Consider now an index in Figure 4 which encodes the same information as the index in Figure 3. In addition, it preserves the temporal arrangement of the morphs. This is accomplished by means...
of skip $\epsilon$-transitions, shown dashed. Such a modification would allow computation of valid morph and word posterior probabilities by pushing the combined weight of acoustic log-likelihoods and language model log-probabilities (weighted appropriately) to the initial state if the skip transitions are omitted from propagation.

Morph search in the skip index $I$ can be performed by

$$\mathcal{R} = (\mathcal{M}_2 W \circ \mathcal{Q}) \circ \text{push}(I)$$

where $\mathcal{Q}$ is a word query and $\mathcal{R}$ is the result. In order to perform word search the index needs to be converted first. This is accomplished by composing it first with the inverse of $\mathcal{M}_2 W$ and then with a language model transducer $L$. Note that $\mathcal{M}_2 W$ must encode $\epsilon$-transitions and handle them correctly in composition. Search then can be performed by composing with the outcome of the word query

$$\mathcal{R} = \mathcal{Q} \circ \text{push}(L \circ \mathcal{M}_2 W^{-1} \circ I)$$

There are several options how phone search can be performed. One option is to search the word index for keywords similar to the one requested and known as proxy keywords [19]. Similarity is usually option is to search the index needs to be converted first. This is accomplished by composing it first with the inverse of $\mathcal{P}_2 P$ and then with a language model transducer $L$. Note that $\mathcal{P}_2 P$ must encode $\epsilon$-transitions and handle them correctly in composition. Search then can be performed by composing with the outcome of the word query

$$\mathcal{R} = \mathcal{Q} \circ \text{push}(L \circ \mathcal{M}_2 W^{-1} \circ I)$$

where $\mathcal{P}_2 P$ is a phone-to-word transducer which maps phone sequences into words. Another option is to create a phone index using either the original morph index or created word index [20]. The word index is expected to give more accurate scores hence

$$\mathcal{R} = (\mathcal{P}_2 W \circ \mathcal{Q}) \circ \mathcal{P}_2 P \circ \mathcal{P}_2 W^{-1} \circ \text{push}(L \circ \mathcal{M}_2 W^{-1} \circ I) \quad (5)$$

For a dense $\mathcal{P}_2 P$ matrix it may be important to restrict the search space of the phone search. For instance, only $n$ highest scoring phone sequences could be selected in equation (6) prior to composing with the created phone index [20].

## 4. EXPERIMENTS

Experiments in this paper were conducted on 4 IARPA Babel program languages. As shown in Table 1, these languages come with varying amounts of web data. A full language pack (FLP) comprises 60 hours of conversational telephone speech (CTS) data was used for training plus additional 10 hours is available for development. All systems described in this paper are graphemic and built using approaches described in [21]. Tandem and Hybrid acoustic models are used for each language [22]. These were built using features comprising perceptual linear prediction coefficients [23], pitch [24], probability of voicing [24] and bottleneck (BN) features. For Swahili the BN features were extracted from a feed-forward neural network (NN) trained on the FLP data. For the remaining languages these were extracted from multi-task feed-forward NNs trained on 24 Babel languages plus English, Arabic, Mandarin and Spanish CTS data provided by LDC. The multi-language NNs were trained by IBM and RWTH Aachen [25]. 4 acoustic models were built for these languages and 2 acoustic models for Swahili. For efficiency these multiple acoustic models were used in a single pass of joint decoding [26]. Language models (LM) are simple n-grams estimated on acoustic transcripts and web data where appropriate [6]. Unsupervised morphological decomposition was performed using the Morfessor toolkit [8]. This was estimated on the FLP data and then applied to the web data where appropriate. Morph LMs were then built in the same fashion as word based LMs. Table 2 summarises morph

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocab (#)</th>
<th>Char OOV</th>
<th>ASR</th>
<th>KWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swahili</td>
<td>7.3</td>
<td>5.8</td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Dholuo</td>
<td>7.5</td>
<td>5.5</td>
<td>0.25</td>
<td>0.0</td>
</tr>
<tr>
<td>Amharic</td>
<td>25.6</td>
<td>4.3</td>
<td>0.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Georgian</td>
<td>9.6</td>
<td>6.2</td>
<td>2.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 2. Summary of morph grapheme statistics

LM statistics for each language. Note that the number of characters is shown per vocabulary rather than per corpus which was set to 3. The resulting number of morphs extracted for each language varies a lot. The large number of Amharic morphs originates from the large number of graphs, 247, representing different consonant vowel sequences. These were split into constituent consonant and vowel graphemes for acoustic modelling. As can be seen from Table 2 morphological decompositions yield very low OOV rates for both ASR and KWS. Morph-to-word transduction experiments were performed using an internal version of the OpenFST toolkit [27]. Keyword search experiments were performed using proprietary IBM keyword search software [18]. About 2,000 keywords are available for each language [28]. These are split into in-vocabulary (IV) and out-of-vocabulary (OOV) keywords. The IV keywords are searched in word, morph and word indices for word, morph and morph-to-word systems respectively. The OOV as well as IV keywords not yielding any hits may be searched at the grapheme level by converting queries and lattices into their graphemic form and applying a grapheme-to-grapheme confusion matrix to yield top-$n$ confusable graphemes. Language models (LM) are simple n-grams estimated on acoustic transcripts and web data where appropriate [6]. Unsupervised morphological decomposition was performed using the Morfessor toolkit [8]. This was estimated on the FLP data and then applied to the web data where appropriate. Morph LMs were then built in the same fashion as word based LMs. Table 2 summarises morph

<table>
<thead>
<tr>
<th>Language</th>
<th>Vocab (#)</th>
<th>Char OOV</th>
<th>ASR</th>
<th>KWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swahili</td>
<td>7.3</td>
<td>5.8</td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Dholuo</td>
<td>7.5</td>
<td>5.5</td>
<td>0.25</td>
<td>0.0</td>
</tr>
<tr>
<td>Amharic</td>
<td>25.6</td>
<td>4.3</td>
<td>0.0</td>
<td>0.8</td>
</tr>
<tr>
<td>Georgian</td>
<td>9.6</td>
<td>6.2</td>
<td>2.5</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 2. Summary of morph grapheme statistics

LM statistics for each language. Note that the number of characters is shown per vocabulary rather than per corpus which was set to 3. The resulting number of morphs extracted for each language varies a lot. The large number of Amharic morphs originates from the large number of graphs, 247, representing different consonant vowel sequences. These were split into constituent consonant and vowel graphemes for acoustic modelling. As can be seen from Table 2 morphological decompositions yield very low OOV rates for both ASR and KWS. Morph-to-word transduction experiments were performed using an internal version of the OpenFST toolkit [27]. Keyword search experiments were performed using proprietary IBM keyword search software [18]. About 2,000 keywords are available for each language [28]. These are split into in-vocabulary (IV) and out-of-vocabulary (OOV) keywords. The IV keywords are searched in word, morph and word indices for word, morph and morph-to-word systems respectively. The OOV as well as IV keywords not yielding any hits may be searched at the grapheme level by converting queries and lattices into their graphemic form and applying a grapheme-to-grapheme confusion matrix to yield top-$n$ confusable graphemes. Language models (LM) are simple n-grams estimated on acoustic transcripts and web data where appropriate [6]. Unsupervised morphological decomposition was performed using the Morfessor toolkit [8]. This was estimated on the FLP data and then applied to the web data where appropriate. Morph LMs were then built in the same fashion as word based LMs. Table 2 summarises morph

<table>
<thead>
<tr>
<th># Unit</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>W Word</td>
<td>47.6</td>
<td>0.5684</td>
</tr>
<tr>
<td>M Morph</td>
<td>49.0</td>
<td>0.5145</td>
</tr>
<tr>
<td>M2W</td>
<td>47.9</td>
<td>0.5448</td>
</tr>
</tbody>
</table>

Table 3. Swahili word, morph and morph-to-word transduction
generated all word sequences present in the word system. The KWS results show a large gain for IV keywords. Note that the IV/OOV split for the morph system is based on the word vocabulary. Not shown in Table 3 is the performance of morph system where plusses were removed from decompositions. This relaces constraints on possible morph sequences but results in much poorer KWS performance totalling 0.4797. Table 4 shows Dholuo results. Here, the morph-to-word system shows better ASR results than the morph system. This indicates the usefulness of word level constraints imposed by the word language model for this language. Note that the use of a higher order morph LM does not show gains over the trigram LM. Amharic results in Table 5 show a mixture of the trends observed so far. Here, the morph-to-word system shows better ASR but worse KWS performance. Finally, Table 6 shows OpenKWS 2016 surprise, Georgian, language results which show improvements both in ASR and KWS. These results suggest that morph-to-word transduction can be an effective way to improve morph language modelling for ASR. However, the improvement largely depends on how effective the word language models are. Since TER and MTWV are not strongly correlated it is harder to predict the impact on MTWV. The results, however, confirm that large IV MTWV gains can be obtained. More research is needed to investigate why word level IV performance cannot be obtained with the current approach or why degradation is seen on Amharic. A summary of individual systems is shown in Figure 5. For 3 out of 4 languages combining single decoding runs and 4 searches over 4 different indices.

Table 4. Dholuo word, morph and morph-to-word results

<table>
<thead>
<tr>
<th># Unit</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BG</td>
<td>TG</td>
</tr>
<tr>
<td>W Word</td>
<td>39.6</td>
<td>39.3</td>
</tr>
<tr>
<td>M Morph</td>
<td>41.5</td>
<td>39.8</td>
</tr>
<tr>
<td>M2W</td>
<td>39.6</td>
<td>39.4</td>
</tr>
</tbody>
</table>

Table 5. Amharic word, morph and word-to-morph results

<table>
<thead>
<tr>
<th># Unit</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BG</td>
<td>TG</td>
</tr>
<tr>
<td>W Word</td>
<td>41.5</td>
<td>41.2</td>
</tr>
<tr>
<td>M Morph</td>
<td>44.0</td>
<td>43.5</td>
</tr>
<tr>
<td>M2W</td>
<td>42.7</td>
<td>42.4</td>
</tr>
</tbody>
</table>

Table 6. Georgian word, morph and morph-to-word results

<table>
<thead>
<tr>
<th># Unit</th>
<th>TER (%)</th>
<th>MTWV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BG</td>
<td>TG</td>
</tr>
<tr>
<td>W Word</td>
<td>39.9</td>
<td>39.3</td>
</tr>
<tr>
<td>M Morph</td>
<td>44.5</td>
<td>41.1</td>
</tr>
<tr>
<td>M2W</td>
<td>40.5</td>
<td>40.4</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

Finding an appropriate unit of language modelling is an open question for many applications and for many languages. Generally there are many options to choose from: words, morphs, syllables, phones, etc. This paper has looked at the morph units which provide a sufficiently wide scope of modelling yet empirically yield very low out-of-vocabulary rates. In order to enlarge the scope of modelling even further this paper has proposed a morph-to-word transduction that enables to convert morph sequence into word sequences. This has multiple benefits including possibility to apply powerful word language models as well as making decisions during pruning, confusion network decoding, keyword search at the word level. This paper has also looked at how word, morph and phone keyword searches can be efficiently performed on a single morph index. A modification to the standard index has been proposed to enable multiple searches. Experiments examining morph-to-word transduction were performed on 4 Babel program languages. These showed that it is possible to obtain speech recognition gains over morph systems as well as keyword search gains over individual word and morph systems.
6. REFERENCES


