APPLYING LOG LINEAR MODEL BASED CONTEXT DEPENDENT MACHINE TRANSLATION TECHNIQUES TO GRAPHEME-TO-PHONEME CONVERSION

Rong Zhang and Bowen Zhou
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{zhangr,zhou}@us.ibm.com

ABSTRACT
Grapheme-to-Phoneme conversion is a challenging task for speech recognition and text-to-speech systems for which the functionality of automatically predicting pronunciations for OOV words is highly desirable. In this paper, Grapheme-to-Phoneme conversion is viewed as a special case of sequence translation problem and we propose to tackle it with phrase based log-linear translation model. We improve standard machine translation method by utilizing context dependent units which lead to a better many-to-many alignment between chunks of graphemes and phonemes. Furthermore, hypotheses combination technique is applied to combine outputs generated by multiple translation models trained with different alignment units. Our proposed approach was evaluated on NetTalk and CMUdict datasets. Significant improvements on conversion accuracy are observed on both sets compared to conventional translation method: phoneme level error rates are reduced relatively by 18.4% and 22.5%, respectively. Our approach also performs better than or as good as previously published data driven methods examined on the same tasks.

Index Terms— Grapheme-to-Phone conversion

1. INTRODUCTION
Grapheme-to-Phoneme (G2P) conversion is the task of converting a word from its spelling form written as a sequence of letters to its pronunciation represented as a sequence of phonemes. G2P conversion is a crucial technique of handling OOV words in text-to-speech systems given the infeasibility of building a universal dictionary to cover all the words in one language, especially proper names, neologisms and morphological variations. Other important applications of G2P conversion include speech recognition and speech-to-speech translation systems, which often require the capability that new word along with its automatically generated phonetic baseform can be added to system on the fly.

Converting grapheme to phoneme is a difficult task due to the lack of regular correspondence between spelling and pronunciation as seen in many languages [1]. Various data driven and statistical machine learning algorithms have been proposed. With a few exceptions, these methods can be roughly classified into two categories, multi-class classification or sequence prediction, according to if the conversion is implemented to consider sequential dependency between phonemes. In the classification category, represented by Neural Nets [2] and Decision Tree [3], phonemes are predicted independently and locally from input grapheme chunks normally consisting of a letter and its context. By contrast, sequence prediction methods, such as Hidden Markov Model [4, 5] and joint N-gram model [6, 7, 8, 9] also model the structural dependency in output sequence in the way that previously predicted phonemes are used as features in the inference of current phoneme. Among the sequence prediction methods, joint N-gram model is the state-of-the-art approach that holds the best record of conversion accuracy in G2P research. In this approach, N-gram language modeling techniques are utilized to estimate the joint distribution of graphemes and phonemes on the basis of Graphoneme, a model unit composed of a pair of grapheme string and phoneme string with possibly different lengths. More recently, machine translation technique was proposed to tackle the G2P conversion problem [10]. However, the results reported are surprisingly poor compared to other data driven methods especially joint N-gram model.

In this paper, G2P conversion is also viewed as a special translation process where the source language is the sequence of graphemes and the target language is the sequence of phonemes. We improve standard machine translation method, which assumes context independency in word alignment, by utilizing context dependent units which help to achieve many-to-many alignment between chunks of graphemes and phonemes. Furthermore, ROVER is applied to combine hypotheses generated from translation systems trained with different alignment units. Experimental results on NetTalk [2] and CMUdict [11] show that our approach not only outperforms conventional machine translation model by a wide margin, but also manifests superiority to other data driven methods examined on the same tasks.

2. LOG LINEAR MODEL FOR PHRASE BASED MACHINE TRANSLATION

The objective of machine translation task is formulated as to find the most likely target language sentence for a questioned source language sentence. For G2P conversion, this is to find the best phoneme sequence $p = p'_1, p'_2, \ldots, p'_n$ given the input grapheme sequence $g = g'_1, g'_2, \ldots, g'_m$ that maximizes the conditional probability $\Pr(p \mid g)$. Alternatively to the classical noise channel approach, the modern phrase based machine translation techniques directly estimate $\Pr(p \mid g)$ using log linear model [12] as that

$$
p^* = \arg\max_p \Pr(p \mid g) = \arg\max_p \frac{\exp \left( \sum_k \lambda_k f_k(p, g) \right)}{\sum_p \exp \left( \sum_k \lambda_k f_k(p', g) \right)}
$$

where $f_k(.)$ is a feature function that maps pairs $(p, g)$ to a nonnegative value, and $\lambda_k$ is the feature weight.

Log linear model essentially implements a powerful statistical combination framework that various knowledge sources, each of which is expressed by a feature function, are integrated together in making joint decision approaching global optimum. These features can take on any form and even overlap, representing redundant information on each other. Particularly, the knowledge sources employed by a machine translation system typically include target
language model, source distortion model, phrase translation models and lexical translation models in two directions.

The training of a machine translation system amounts to building each component feature models and optimizing their associated weights. Starting from a collection of parallel sentences, the first and crucial step is to perform word alignment in two directions, from source to target and from target to source, deriving two sets of Viterbi alignments. The two sets are then merged into a single word alignment matrix serving as the boundary points for phrase extraction. Phrase pairs with respect to the boundary constraint are then identified using certain heuristic rule and pooled together to form phrase translation tables. Translation probabilities, on both phrase and lexical levels, are then learned using the Maximum Likelihood criterion.

As seen in many languages, the irregular correspondence between graphemes and phonemes and the multiple pronunciations of a grapheme varying with its context [1] require a many-to-many context dependent grapheme-phoneme alignment. The objective, however, is only partially fulfilled by standard machine translation model training. The classical IBM alignment model [13], which is the basis of whole training, is of strong independence assumptions due to its pursuit of efficiency in processing large parallel corpus. Specifically, IBM word alignment model is performed as a generation process from target language to source language: each target language word is first duplicated several times, from zero or many, according to a fertility function, and then each copy is mapped to a source language word based on word-to-word translation probabilities. In the process, neither target nor source context is considered. Distortion model, which implicitly captures the non-consecutive dependency often observed in irregular languages e.g. English. For example, when the letter e is added to mat to make word mate, the pronunciation of letter a, rather than the preceding letter t, is affected.

\[ e \rightarrow \text{mate} \]

Hence the pronunciation of letter e, \( \epsilon \), is affected.

\[ \epsilon \rightarrow \text{unigram at current position} \]

\[ c_0 \rightarrow c_1 \]

right context bigram starting from current position

\[ c_1 \rightarrow c_0 \]

left context bigram ending at current position

\[ c_1 \rightarrow c_0 \rightarrow c_1 \]

trigram centred at current position

\[ c_0 \rightarrow c_1 \rightarrow c_2 \]

trigram starting from current position

\[ c_0 \rightarrow c_2 \]

bigram with one position skipping

Table 1 Context Dependent Alignment Units

As illustrated in the example above, we apply context dependent alignment units to both grapheme and phoneme sequences. This is different to some previous attempt [14] which only works on source language side. Table 2 lists all the context dependent grapheme-phoneme combinations used in our experiments.

### 3.2. Hypotheses Combination

One concern about using context dependent unit in MT model training is that it may lead to data sparseness problem. Although severe problem wasn’t observed in our experiments, a systematic solution to address potential data sparseness is still desirable. ROVER, an ensemble based hypothesis combination approach is adopted in our experiments as the solution. Specifically, the final conversion hypothesis is the combination result of the outputs from multiple machine translation systems which are trained separately using different context dependent units given in Table 2. Hence the degradation in a particular system caused by data sparseness can be prevented from spreading to other systems.

4635
4. EVALUATION

4.1. Datasets and Experimental Set-up

Our experiments were conducted on two dictionary datasets: NetTalk and CMUDict, both of which have been examined by many G2P researchers using various techniques. The NetTalk dictionary was originally developed for the research of neural network. It has 19,802 words and 20,008 pronunciations due to the existence of words with multiple pronunciations. The phone set has 52 phonemes, including the null phoneme and five double phonemes. Three stress markers are used, and two other markers provide syllabic information. CMUDict was designed for continuous speech recognition, which contains many abbreviations, proper names, loan words, and uncommon pronunciations. The dictionary consists of more than 119,000 words, with a total of over 127,000 pronunciations. The phone set is composed of 39 phones. Lexical stress markers are also available. Our research focuses on the prediction of baseform, so stress and syllabification markers are removed from both datasets. Duplicate entries, caused by the removal of stress and syllabification markers, are also removed. For each datasets we randomly selected 80% of the words, along with their pronunciations which can be multiple, as the training set, 10% as dev set, and the rest 10% for testing.

<table>
<thead>
<tr>
<th>Model</th>
<th>Graph. Unit</th>
<th>Phone. Unit</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>c₀</td>
<td>c₀</td>
<td>8.7%</td>
<td>37.7%</td>
</tr>
<tr>
<td>M2</td>
<td>c₀</td>
<td>c₀</td>
<td>10.5%</td>
<td>45.2%</td>
</tr>
<tr>
<td>M3</td>
<td>c₀</td>
<td>c₀</td>
<td>8.2%</td>
<td>37.1%</td>
</tr>
<tr>
<td>M4</td>
<td>c₀</td>
<td>c₀</td>
<td>7.7%</td>
<td>36.1%</td>
</tr>
<tr>
<td>M5</td>
<td>c₀</td>
<td>c₀</td>
<td>10.4%</td>
<td>45.2%</td>
</tr>
<tr>
<td>M6</td>
<td>c₁</td>
<td>c₀</td>
<td>8.4%</td>
<td>37.9%</td>
</tr>
<tr>
<td>M7</td>
<td>c₁</td>
<td>c₀</td>
<td>13.1%</td>
<td>52.0%</td>
</tr>
<tr>
<td>M8</td>
<td>c₁</td>
<td>c₀</td>
<td>13.8%</td>
<td>53.6%</td>
</tr>
<tr>
<td>M9</td>
<td>c₁</td>
<td>c₀</td>
<td>9.0%</td>
<td>41.8%</td>
</tr>
<tr>
<td>M10</td>
<td>c₁</td>
<td>c₀</td>
<td>11.1%</td>
<td>46.5%</td>
</tr>
<tr>
<td>M11</td>
<td>c₁</td>
<td>c₀</td>
<td>9.7%</td>
<td>42.5%</td>
</tr>
<tr>
<td>M12</td>
<td>c₁</td>
<td>c₀</td>
<td>9.0%</td>
<td>41.6%</td>
</tr>
<tr>
<td>M13</td>
<td>c₀</td>
<td>c₀</td>
<td>8.8%</td>
<td>38.8%</td>
</tr>
<tr>
<td>M14</td>
<td>c₀</td>
<td>c₀</td>
<td>9.1%</td>
<td>42.7%</td>
</tr>
<tr>
<td>ROVER Combination</td>
<td>7.1%</td>
<td>33.0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Performance of Log Linear Model Based MT Methods on NetTalk

Conversion performance is measured by Phoneme Error Rate (PER), which is the edit distance between predicted phonemes and desired phonemes. For a word with multiple pronunciations, the closest one is used in calculation. We also report Word Error Rate (WER), for which the whole predicted phone sequence is judged as correct if and only if it exactly matches to one of given pronunciations, otherwise incorrect.

Folsom toolkit [15], one of the best phrase based MT trainers and decoders, was utilized in our experiments. Separate MT models were trained for each of the context dependent grapheme/phoneme settings given in Table 2, resulting in 14 systems. Each system includes phrase and lexicon level translation models and a phoneme language model. 4-gram is used as default if not specified. Word reordering model is not employed since G2P conversion is essentially a monotone decoding process. Model weights are optimized on dev set using discriminative training to maximize BLEU score. For ROVER combination, hypothesis voting weight is tuned on dev set using hill-climbing algorithm.

4.2. Experiments on NetTalk

4.2.1. Results of Machine Translation Model

Table 2 presents the performance of machine translation techniques on NetTalk. The result shows that conversion error is reduced significantly by using context dependent alignment units. The PER of the baseline system M1, the one that Grapheme= c₀ and Phoneme= c₀, is 8.7%. When applying context dependent bigrapheme and bi-phoneme to training (Model M4), PER is down to 7.7%. When ROVER is utilized to generate final integrated hypothesis, PER is further reduced to 7.1% which represents a relative reduction of 18.4% from the baseline system. Accordingly, WER is reduced by relatively 12.5%. This is a very encouraging result that not only evidences the effectiveness of context dependent modeling in G2P conversion, but suggests that the proposed alignment units are complementary to each other in generating hypotheses with less-correlated error patterns.

4.2.2. Comparison to Neural Nets, Conditional Random Fields and Joint N-gram Model

Neural Nets is a representative multi-class classification approach used in G2P conversion. We implemented a back-propagation Neural Nets similar to the one in [2]. The input layer encodes a window of 5 input graphemes including the current one and its two left and right context graphemes. The output layer contains 52 nodes each of which corresponds to a phoneme to be predicted. The hidden layer has 50 nodes. To make a fair comparison with Neural Nets which doesn’t utilize the sequence structure of phonemes, we disabled the language model module in each translation system. Hence the generation of phonemes only relies on translation models. Furthermore, the translation systems, that use contextual phoneme or that contextual grapheme is out of the scope of 5-grapheme window, are not considered. Experimental results are presented in Table 3. The translation model using bi-grapheme outperforms Neural Nets by a big margin. The results also stress the importance of modeling structural dependence of phonemes in G2P conversion. Without the help of language model, conversion performance deteriorates significantly.

<table>
<thead>
<tr>
<th>MT Models</th>
<th>Graph. Unit</th>
<th>Phone. Unit</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₀</td>
<td>c₀</td>
<td>13.2%</td>
<td>58.1%</td>
<td></td>
</tr>
<tr>
<td>c₀</td>
<td>c₁</td>
<td>10.7%</td>
<td>49.6%</td>
<td></td>
</tr>
<tr>
<td>c₁</td>
<td>c₀</td>
<td>14.9%</td>
<td>64.4%</td>
<td></td>
</tr>
<tr>
<td>c₁</td>
<td>c₀</td>
<td>15.1%</td>
<td>60.9%</td>
<td></td>
</tr>
<tr>
<td>c₀</td>
<td>c₁</td>
<td>12.5%</td>
<td>53.1%</td>
<td></td>
</tr>
<tr>
<td>c₀</td>
<td>c₂</td>
<td>11.9%</td>
<td>52.3%</td>
<td></td>
</tr>
<tr>
<td>ROVER Combination</td>
<td>9.3%</td>
<td>44.5%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Comparison between MT Methods and Neural Nets

Conditional Random Fields (CRF) are a type of statistical sequence modeling approach often used for the labeling or parsing of sequential data. CRF++, a well-known CRF toolkit [16], is used in our experiment. CRF++ supports the modeling of long-range features on source side, but for target side, only bigram is currently available. The features used in our CRF experiment include
unigram, bigram and trigram of graphemes, and bigram of phonemes. Accordingly, we downgrade our translation systems by limiting the language model to bigram from 4-gram, and removing the systems which contextual grapheme are out of the scope of trigram. Comparative results are given in Table 4, from which we can see that by using context dependent alignment units, single MT model (Grapheme=c0−c1 & Phoneme=c0−c1) achieves equivalent performance as compared to CRF. Furthermore, the conversion result after ROVER combination is much better than that of CRF especially when measured by WER. Both MT approach and CRF are based on log linear model. The difference between them is that, MT approach uses a divide-and-conquer strategy that the training of a complex system is broken down into the trainings of a set of sub-systems with different optimization criteria. In contrast, CRF pools all features together under a single training criterion. Therefore, MT approach is likely to be more capable and practical than CRF in handing large complicated task.

Joint N-gram model is the state-of-the-art approach that holds the best record of conversion accuracy in G2P research. Table 5 presents the results of joint N-gram models which were obtained by using the toolkit described in [8]. The best performance is reached by using 7-gram. We also reported the performance of joint 4-gram which is comparable to the language model used in our machine translation approach. Comparison between Table 2 and Table 5 shows that single context dependent translation model, M4, demonstrates even slightly better performance than the joint N-gram models on this dataset. When conducting hypotheses combination, the superiority of machine translation approaches is more obvious: the PER and WER are 9.0% and 9.1% relatively better than the best joint 7-gram model.

In this paper, we proposed to use log linear model based machine translation techniques to address G2P conversion problem. We improved conventional MT method by utilizing context dependent modeling and hypotheses combination technique, which lead to significant improvement of conversion accuracy. Besides G2P, there are problems having a similar nature, such as word segmentation and POS tagging, which can also be addressed by using statistical MT techniques. These will become the topics of our future work.

### 6. REFERENCES