IMPROVING LETTER-TO-SOUND CONVERSION PERFORMANCE WITH AUTOMATICALLY GENERATED NEW WORDS

Jia-Li You[*1], *Yi-Ning Chen*[2], *Frank K. Soong*[2], *Jin-Lin Wang*[1]
[1]Graduate School of the Chinese Academy of Sciences, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]Microsoft Research Asia, Beijing, China
[1]youjiali@mails.gucas.ac.cn , [2]{ynchen, frankkps}@microsoft.com, [1]wangjl@dsp.ac.cn

ABSTRACT

We propose a novel way to alleviate the data sparseness problem in training Letter-to-Sound (LTS) N-gram models by adding automatically generated new words to the training set. The proposed method consists of two procedures: (1) generating a large pool of new words automatically; (2) selecting good new word candidates from the new word pool via semi-supervised learning. The new words are created by replacing stressed syllables of an existing word with other stressed syllables under specified contextual constraints. The new word selection by semi-supervised learning is based upon consistent pronunciation predictions by different LTS models. After adding new words to the training set, the performance of LTS conversion is significantly improved. For the NetTalk dictionary, compared with the performance from the N-gram baseline model, 21.6% relative word error rate reduction is obtained. For the CMU dictionary, 9.1% and 5.6% relative word error rate reductions are obtained, respectively, with/without considering the stress.

Index Terms— Letter-to-Sound, data sparseness, artificial data, semi-supervised learning

1. INTRODUCTION

In many speech applications, a reasonably large pronunciation lexicon is needed for specifying the spelling and corresponding pronunciations of commonly used words. However, no matter how large the lexicon is, there are always some out-of-vocabulary (OOV) words which are not covered. To predict the pronunciations of these OOV words, a good LTS module is desirable.

Different methods have been investigated on LTS conversion. Both manually constructed rules and data-driven algorithms have been tried [1-9]. Manually constructed rules need expert knowledge of a linguist and they are hard to be extended from one language to the other. Data-driven techniques are the state-of-art methods, including: decision tree [2,3], hidden Markov model (HMM) [4], N-gram model [5-7], maximum entropy model [8], and transformation-based error-driven approach [9]. They can be automatically trained and language independent.

In data-driven techniques, statistical inference between graphemes (spellings) and phonemes (pronunciations) can be made from the training data and thus trained model can be used to predict the pronunciations of unseen words. In training such a model, more data tends to train a model with better predicting capability as shown Fig. 1.

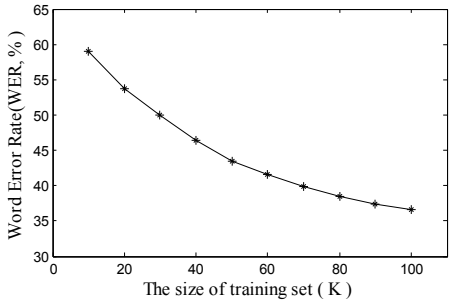

Fig 1. The LTS performances for different sizes of training set. (The results are from the CMU dictionary which will be introduced in Section 4.)

Therefore, if a large word list with corresponding pronunciations is available, most of the LTS conversion rules of a language can be learned. However, in most situations, it is impossible to obtain such a word list. While it is relatively easier to extract a large word list with spellings, e.g., from the web, it is still a grand effort to get the correct corresponding pronunciations. According to our experience, it takes about 8 hours for an expert to guess the pronunciations of 1,000 words.

In this paper, we propose two novel methods to generate new words with pronunciations. The first one is to generate new words from the available pronunciation dictionary, the other one is to generate data based upon semi-supervised learning [10-13] for a given list of word spellings. After the automatically generated words are used to augment the original training set, we hope to improve the performance of LTS model by alleviating the intrinsic data sparseness problem.

The rest of the paper is organized as follows. In Section 2, the baseline of LTS conversion is introduced. In Section 3, the methods of generating new data are proposed in details. The experiments and analyses are shown in Section 4. Finally, conclusions are drawn in Section 5.

2. BASELINE: CHUNK BASED N-GRAM FOR LTS CONVERSION

N-gram statistical modeling techniques have been applied successfully to speech, language and other data of sequential

[*] This work is done when the first author visits Microsoft Research Asia as an intern

nature. In LTS conversion, N-gram has also shown its effectiveness in predicting word pronunciation from its letter spellings [5-7]. The relationship among grapheme-phoneme (Graphoneme) pairs is modeled as Eq. (1).

$$\begin{aligned}\tilde{S} &= \arg\max_{S}\{P(S \mid L)\} \\ &= \arg\max_{S}\{P(S, L)\} \\ &= \arg\max_{S}\{\prod_{i=1}^{n} P(g_i|g_{i-1},...,g_1)\}\end{aligned} \quad (1)$$

where $L = \{l_1, l_2, ..., l_n\}$ is the grapheme sequence of a word W; $S = \{s_1,s_2,...,s_n\}$ is the phoneme sequence; and $g_i = <l_i,s_i>$ is a graphoneme; l_i and s_i are aligned as one letter corresponding to one or more phonemes (including null) by the dynamic programming algorithm described in [2,3].

Limited by the amount of training data, the N-gram model cannot be well trained if context becomes too large. However, many linguistic phenomena need a long-distance dependency modeling. To solve this problem, some stable (more frequently observed) spelling-pronunciation chunks are extracted as independent units and corresponding N-gram models are trained. For generating chunks, mutual information (MI) between any two chunks is calculated to decide whether two chunks should be joined together to form one chunk. This process is shown in Fig.2 which is similar to the one used in [6].

Step 1: Initiate the chunk set with all of the graphonemes gotten after alignment;
Step 2: Calculate MI for all succeeding chunks in the training set;
Step 3: Add the chunks with MI higher than a preset threshold into the chunk set as a new letter chunk;
Step 4: Stop if the number of chunks in the set is above a certain threshold or no more new chunk is identified; otherwise go to step 2.

Fig.2. A description of MI algorithm for chunk extracting

In decoding, paths of all possible pronunciations that match the input word spellings are efficiently searched via the Viterbi algorithm and the pronunciation of the maximum likelihood path is retained as the final result.

3. METHODS FOR GENERATING NEW WORDS

3.1. Generating new words based upon replacing stressed syllables

In statistical machine translation, paraphrases are generated as an efficient way to enlarge the data set to alleviate the data sparseness problem [14]. Similar to the paraphrase generation, we generate new words (artificial words) to train LTS statistical models.

3.1.1. The new word generation process

Given a pronunciation dictionary, the process to generate new words is as follows:

a) If no syllable boundaries in a dictionary, mark them for all words at phoneme level based upon some syllabification rules.

b) Align graphemes with phonemes by dynamic programming. Then transfer syllable boundary marks from marked phonemes to the correspondingly aligned graphemes.

c) Make a list of primary stressed syllables from all words in the dictionary.

d) Generate artificial data. All words in the dictionary are the seed words. For each seed word, extract the primary stressed syllable and compare it with the replacement candidates in the prepared list of stressed syllables. If the replacement rule (details in 3.1.2) is satisfied, replace the primary stressed syllable. A new word is thus generated with its corresponding pronunciation. After all seed words are processed, a new word list with pronunciations is generated.

3.1.2. Replacement rules for generating new words

To generate words that are more plausible both in letter spelling and phonemic structure, two replacement rules are adopted:
(1) Replacement based upon similar phonemic structure.
(2) Replacement based upon similar graphonemic structure.

For a syllable, its structure is extracted based upon its phoneme sequence. All consonants are denoted by the symbol "C" in the structure. In the phonemic structure rule, vowels are represented in its original phonemic symbol; in the graphonemic structure rule, graphonemes of vowels (letter-phoneme symbol pair of the vowel) are used in the structure. All should conform with their positions in the original syllable. Fig. 3 shows an example of new words generation by the two replacement rules.

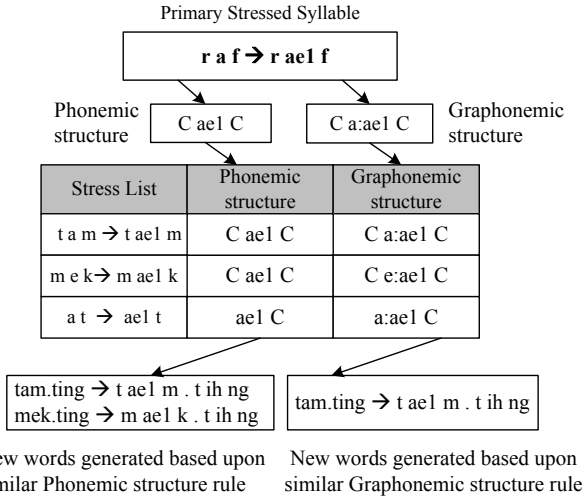

Fig.3. An example of new words generation by the two replacement rules

Each rule can generate its own new word list with corresponding pronunciations. The graphonemic structure rule, along with its spelling conformation requirement, is more restricted than the phonemic structure rule.

3.2. Generating new words based upon semi-supervised learning

It is easy to extract new words from the Internet or other text databases, however, thus new words are usually without pronunciations. For LTS training, we like to generate correct or at least probabilistically correct pronunciations to enrich our training samples.

Semi-supervised learning [10-13] can use unlabeled data to improve the model training efficiency. The basic idea is to

automatically annotate (label) the unlabeled samples using a classifier trained on a small labeled set. The model is then retrained or refined with additional auto-labeled data.

Agreement learning is one type of semi-supervised learning which has been tested in automatic accent annotation [13]. It needs several classifiers to classify the unlabeled data separately. The labeling results which are in agreement among different classifiers are deemed as reliable and they will be used for retraining. In chunk N-gram based LTS training, we found that different chunks may have different capabilities to characterize the training set. For example, in the NetTalk dictionary, the decoded pronunciation paths from 3 different chunk N-grams (the number of chunks are 500, 1,000 and 3,000, respectively) are quite different. Only 52.4% of the paths are the same. However, the word error rate of the agreed part is 16.6% which is much lower than the error rate (~35%) of any individual model. Therefore, selecting the results in agreement from different chunk models to enrich the training dataset is quite feasible for improving LTS performance. Although a large percentage of the results do not agree among multiple models, the new word list is large enough to generate ample good new word candidates for retraining the LTS model. The framework for predicting pronunciations of new words in semi-supervised learning is shown Fig. 4.

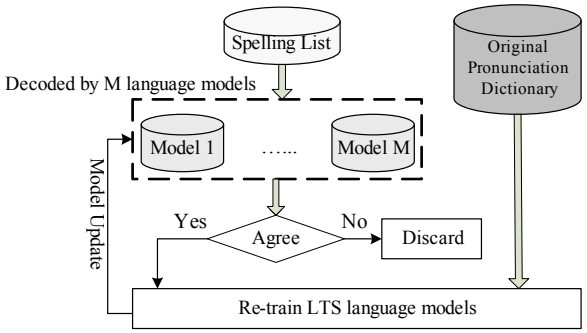

Fig.4. Semi-supervised learning for predicting pronunciations of new words.

In this framework, the words decoded by all models are added to the training set; and all LTS models are re-trained after new words are added; then updated models will perform a new iteration, until the agreement part is the same as the data gotten from the last iteration.

4. EXPERIMENTS AND ANALYSES

4.1. Data

All of the experiments are carried out on two dictionary databases: NetTalk [15] and CMU dictionaries [16]. NetTalk is a dictionary with manually labeled alignments and syllable boundaries at both grapheme and phoneme levels. It consists of 20,008 words with corresponding pronunciations. The CMU dictionary contains about 120,000 words without letter-phoneme alignments and syllable boundaries. To compare our results with other research works [6, 8], we use 90% of the entries for training LTS conversion models, and the rest 10% for testing without considering stress marks in NetTalk. In addition, a word list (spelling only) extracted from Web text is used to generate new words by semi-supervised learning. It has about five-million word entries.

4.2. Performances after generated words are added to the training set

New words are generated by methods presented in previous sections and they are used to augment the original training set to re-train the LTS models. We use the abbreviations listed in Table 1 to denote different methods.

Table 1. Description of the abbreviations of different methods

Abbreviation	Description
Baseline	Graphoneme chunk based N-gram model (No extra data added)
PS	Artificial words generated with similar phonemic structure.
GPS	Artificial words generated with similar graphonemic structure.
Semi	Pronunciations generated by semi-supervised learning
Legitimate words	Words extracted from Web text (spellings only, no pronunciations)

The LTS performances of NetTalk and CMU databases are given in Tables 2 and 3. When artificially generated words with pronunciations are used in augmenting the training set, the LTS performance of NetTalk is significantly improved. Further, if a word list (legitimate words or spellings of artificial words) is labeled with predicted pronunciations by semi-supervised learning, it also brings some benefit for LTS models, especially for the CMU dictionary.

Table 2. LTS performances of NetTalk

Rule	Number of Words	WER (%)
Baseline	18K	34.3
PS	+1,270K	27.4
GPS	+1,000K	26.9
PS + Semi	+173K	30.7
GPS + Semi	+148K	31.4
Legitimate Words + Semi	+1,463K	33.9

Table 3. LTS performances of CMU (*use stressed vowels)

Rule	Number of Words	WER (%)	
		CMU*	CMU
Baseline	113K	34.0	26.7
PS	+3,370K	42.3	36.0
GPS	+2,440K	39.6	33.0
PS + Semi	+866K	31.7	26.2
GPS + Semi	+664K	31.7	26.0
Legitimate Words + Semi	+860K	30.9	25.2

Table 4. Relative error reductions in the two dictionaries with different word generations (%) (* use stressed vowels)

	NetTalk	CMU*	CMU
PS	20.1	*No help*	*No help*
GPS	21.6	*No help*	*No help*
PS + Semi	10.5	6.8	1.9
GPS +Semi	8.5	6.8	2.6
Legitimate Words + Semi	1.2	9.1	5.6

Table 4 lists the relative error reductions when new words are added to the training set. From this table, we can see that generated words are quite useful in training LTS models for both NetTalk and CMU dictionaries.

4.3. Analysis and discussion

Results show that the improvement of LTS performance is rather different across the two dictionaries. NetTalk is a compact and clean dictionary where the syllable boundaries are manually labeled by expert at both grapheme and phoneme level. Therefore, based upon the reliable syllable boundaries, the artificially generated new words are of good quality. On the other hand, CMU is a large, complex dictionary where the word entries are from inhomogeneous sources and no syllable boundaries are marked. Boundary errors induced by automatic alignment are therefore unavoidable. For this reason, not all artificially generated words can be trusted and used for training LTS.

In addition, in evaluation of CMU, because stress is considered, for the new words generated by semi-supervised learning, only the words with one primary stressed syllable are kept which can filter a lot of noisy. Therefore, this is an important reason that the result of legitimate words for CMU from semi-supervised learning is better than it for NetTalk.

To verify whether the generated new words are as useful as the manually labeled data, we did some experiments. First, two models with different sizes of training set are trained. One is gotten from half size of the original training set; the other is from the whole original training set. When comparing the performances of these two models, we found that although half size of training data is increased which needs a high cost work for manually labeling, just around 17% errors can be removed both for NetTalk and CMU. In addition, this labeling work will be harder and improvement will be smaller when the training set is larger. If adding our new words into training set, we can get about relative 9.1%-20% improvement. Therefore, automatically generating methods are efficient in LTS task.

5. CONCLUSIONS

Two approaches of generating new words are proposed, and they are both tested on the two dictionaries: NetTalk and CMU. The results show that if the dictionary is of a high quality with manually labeled syllable boundaries at the grapheme level, the new words generated based on replacing stressed syllables can result in significant LTS performance improvement. For example, in NetTalk, a 21.6% relative error reduction is obtained. Without syllable boundary information, only legitimate words and the spellings of artificially generated words can be used. After generating pronunciations by semi-supervised learning, both legitimate words and artificial word spellings can bring some benefit for model training.

However, those generated new words may still be noisy, e.g., non-existing, atypical English words, or labeling errors from the semi-supervised learning. All these errors may affect the performance of LTS conversion. We did not investigate the issue of how to detect and prune those noisy words. But it is highly plausible that by purifying the training data, we will be able to further improve the LTS performance.

6. ACKNOWLEDGEMENTS

The authors would like to thank Min Chu for useful discussion and Lei Ji for supporting legitimate word list.

7. REFERENCES

[1] Damper, R. I., Marchand, Y., Adamson, M. J. and Gustafson, K., "A Comparison of Letter-to-Sound Conversion Techniques for English Text-to-Speech Synthesis", *in Proc. of the Institute of Acoustics,* 20(6), pp. 245-254, 1999.

[2] Black, A. W., Lenzo, K. and Pagel, V., "Issues in Building General Letter to Sound Rules", *in Proc. of the 3rd ESCA Workshop on Speech Synthesis*, pp. 77-80 1998.

[3] Jiang, L., Hon, H., and Huang, X., "Improvements on a Trainable Letter-to-Sound Converter", *in Proc. of Eurospeech*, pp. 605-608, 1997.

[4] Taylor, P., "Hidden Markov Models for Grapheme to Phoneme Conversion", *in Proc. of Interspeech*, pp. 1973-1976, 2005.

[5] Vozila, P., Adams, J., Lobacheva, Y. and Thomas, R., "Grapheme to Phoneme Conversion and Dictionary Verification Using Graphonemes", *in Proc. of Eurospeech*, pp. 2469-2472, 2003.

[6] Galescu, L., Allen, J. F., "Bi-Directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model", *in Proc. of the 4th ISCA Tutorial and Research Workshop on Speech synthesis*, 2001.

[7] Bisani, M. and Ney, H., "Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion," *in Proc. of ICSLP*, pp. 105-108, 2002.

[8] Chen, S. F., "Conditional and Joint Models for Grapheme-to-Phoneme Conversion", *in Proc. of Eurospeech*, pp. 2033-2036, 2003.

[9] Polyakova, T., and Bonafonte, A., "Using Error-Driven Approach to Improve Automatic Grapheme-to-Phoneme Conversion Accuracy", *in Proc. of TC-STAR Workshop on Speech-to-Speech Translation*, pp. 213-217, 2006.

[10] Clark, S., Curran, J. R., and Osborne, M., "Booststrapping POS Taggers Using Unlabelled Data", *in Proc. of CoNLL*, pp. 49-55, 2003.

[11] Blum, A. and Mitchell, T., "Combining Labeled and Unlabeled Data with Co-Training", *in Proc. of 11th Annual Conf. on Comp. Learning Theory (COLT)*, pp. 92-100, 1998.

[12] Zighed, D. A., Lallich, S., and Muhlenbach, F., "Separability Index in Supervised Learning", *in Proc. of the 6th European Conference PKDD*, pp. 475-487, 2002.

[13] Ni, X., Chen, Y., Chu, M., Soong, F. K., Zhao, Y. and Zhang, P., "Agreement Learning for Automatic Accent Annotation", *in Proc. of ICASSP,* pp. 829-832, 2007.

[14] Lepage, Y. and Denoual, E., "Automatic Generation of Paraphrases to Be Used as Translation References in Objective Evaluation Measures of Machine Translation", *in Proc. of IWP*, pp. 57-64, 2005.

[15] Sejnowski, T. J., "The NetTalk Corpus: Phonetic Transcription of 20,008 English Words", 1998.

[16] Weide, R., "The CMU Pronunciation Dictionary, Release 0.6", Carnegie Mellon University, 1998.