Initial and Evaluations of an Open Source WFST-based Phoneticizer *

◎ Josef Novak†, Dong Yang‡, Nobuaki Minematsu†, Keikichi Hirose†
†The University of Tokyo, ‡Tokyo Institute of Technology

1 Introduction

This paper introduces a new open-source, WFST-based Grapheme-to-Phoneme system, named Phonetisaurus. The system is modular and includes support for several third-party components. The system has been implemented primarily in python, but also leverages the OpenFST framework and is intended to support both practical work as well as educational goals. Standard G2P test sets were used to evaluate the performance of the new system and showed that the system performs comparably to state-of-the-art benchmarks.

2 Related Work

Grapheme-to-phoneme conversion is the term applied to the process of automatically generating pronunciation hypotheses given an input orthography. It is an important issue for both speech synthesis and automatic speech recognition for English and many other languages.

Much research has focused on data-driven methods for training G2P systems, for example [1], [2], [3], and [4]. The basic G2P problem is succinctly formulated in [2] as follows: given a grapheme sequence G, find the phoneme sequence P^* that maximizes $Pr(P|G)$:

$$P^* = \underset{P}{argmax}\ Pr(P|G) = \underset{P}{argmax}\ Pr(G,P) \quad (1)$$

One approach to modeling this is via a joint source channel model such as that described in [1]. The current work builds on this approach and also employs the WFST framework in a manner similar to [5]. Specifically, given an orthography, $G = (g_1, g_2, ..., g_N)$, and a pronunciation $P = (p_1, p_2, ..., p_N)$ the model is trained to compute:

$$\prod_{k=1}^{N} Pr(< g,p >_k \mid < g,p >_{1,k-1}) \quad (2)$$

Two improvements have been made to the basic idea outlined in [5]. First the approach has been extended to support the many-to-many G2P alignment procedure described in [3], and second it has been re-written as a modular, open-source project [6].

3 Alignment

Manual G2P alignments are generally not available, thus it is necessary to first align the grapheme and phoneme sequences in a pronunciation dictionary, prior to building a pronunciation model. One approach is to use a simple dynamic programming algorithm such as Needlement-Wunsch. In most previous literature, including [5], a 1-to-1 alignment procedure has been utilized. Instead, in this work we utilized the EM-based many-to-many alignment procedure detailed in [3] that supports alignments from digraphs such as "sh" to a single phoneme, or the reverse case. This should be advantageous for languages like English, where such mappings occur frequently.

4 Pronunciation Model

The pronunciation model followed the same general outline that was laid out in [5]. An aligned pronunciation dictionary was used to produce a joint n-gram model, and this was then used to "decode" pronunciations for novel words. The construction process is quite simple, and involved the following steps: 1. Convert each aligned sequence, $(g_1, g_2, ..., g_n)$, $(p_1, p_2, ..., p_n)$ to a sequence of aligned pairs, $(g_1 : p_1, g_2 : p_2, ..., g_n : p_n)$; 2. Generate an n-gram model from (1); 3. Convert the n-gram model to an equivalent Weighted Finite-State Acceptor; 4. Re-separate the individual grapheme-phoneme pairs into input and output labels, turning the acceptor into a transducer.

Generating a pronunciation for a new word is achieved by compiling the word into a WFSA and composing it with the pronunciation model. The best hypothesis is just the shortest path through the composed WFST.

---

*オープンソースの WFST-駆動 G2P システムの構築と評価、
◎ノバックジョセフ †、楊冬 ‡、峯松信明 †、広瀬啓吉 †、(† 東京大学、‡ 東京工業大学)

Table 1 Results for 6 G2P test sets. Proposed vs. Bisani [4]. Figures in %; (m) means m2m alignment.

Test set	Author	PER	WER
Celex	Bisani [4]	2.5	11.4
	Proposed(m)	2.6	12.4
OALD	Bisani [4]	3.5	17.5
	Proposed(m)	3.6	18.8
Pronlex	Bisani [4]	6.8	27.3
	Proposed(m)†	6.9	28.4
*NETtalk*15*k*	Bisani [4]	8.3	33.7
	Proposed	***7.3***	34.3
*NETtalk*18*k*	Bisani [4]	7.8	31.8
	Proposed	***6.8***	31.9
*NETtalk*19*k*	Bisani [4]	7.7	31.0
	Proposed	***6.6***	***31.0***

4.1 Experiments

We conducted 6 sets of experiments with the system, utilizing several well-known test sets from the G2P field. The first set of experiments evaluated Phonetisaurus alone on the popular nettalk-15k test set, using 3 alignment methods and n-gram orders from 2 to 7. The results of these experiments are depicted in Fig. 1. $n > 6$ resulted in over-fitting, thus n was set to 6 for the remaining experiments. The many-to-many alignment supported up to 2-2 alignments, but for some test-sets 1-1 alignment performed best.

Further experiments were carried out to compare the proposed system to existing results on other standard test sets. The results from these experiments are summarized in Table 1, where PER refers to Phoneme Error Rate and WER refers to Word Error Rate, and the † indicates the test partitions were not identical. The performance for the proposed and [4] systems is clearly very similar. Experiments also showed that the appropriate alignment procedure depends on the test set. The proposed approach has several advantages however, first it is highly modular, and second it is very fast. The proposed system required 2m 55s training time for the NETtalk-15k data set whereas the [4] approach required many hours. The proposed system also supports many-to-many alignment. Finally, the WFST framework ensures a compact representation of the results, including n-best.

5 Conclusion

In this work we introduced a new, modular open-source phonetizer, called Phonetisaurus, based on the WFST-framework. We showed that it performs

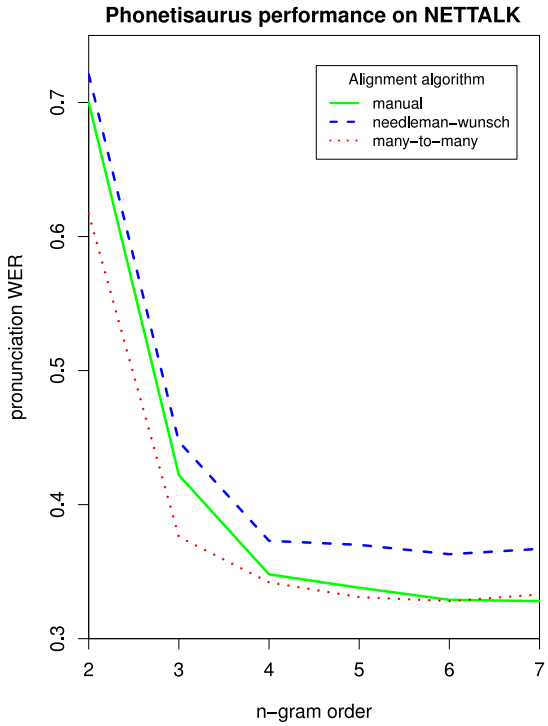

Fig. 1 Results on the nettalk-jiang test set for 3 alignment approaches and n-gram orders 2-7.

comparably to state-of-the-art results on standard test sets. In future we plan to extend it with larger array of native alignment implementations and modeling techniques. We hope it will also be useful as educational software.

参考文献

[1] L. Galescu, et. al., "Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model", Proceedings ISCA Tutorial on TTS, 2001.

[2] S. Chen, "Conditional and Joint Models for Grapheme-to-Phoneme Conversion", Eurospeech, 2003, pages 2033-2036.

[3] S. Jiampojamarn, et. al., "Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion", NAACL HLT, 2007, pages 372-379.

[4] M. Bisani, et. al., "Joint-sequence models for grapheme-to-phoneme conversion", Speech Communication 50, 2008, pages 434-451.

[5] D. Yang, et. al., "Rapid development of a G2P system based on WFST framework", ASJ 2009 Autumn session, 2009, pages 111-112.

[6] J. Novak, http://code.google.com/p/phonetisaurus