論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	The Use of Finite-State Transducers for Modeling Phonological and Morphological Constraints in Automatic Speech Recognition
著者(和文)	古井 貞熙
Authors(English)	Mate Szarvas, Sadaoki Furui
出典(和文)	日本音響学会2001年秋季講演論文集, Vol. , No. 2-1-20, pp. 87-88
Citation(English)	2001 Fall Acoustical Society of Japan Meeting, Vol. , No. 2-1-20, pp. 87-88
発行日 / Pub. date	2001, 10

THE USE OF FINITE-STATE TRANSDUCERS FOR MODELING PHONOLOGICAL AND MORPHOLOGICAL CONSTRAINTS IN AUTOMATIC SPEECH RECOGNITION

© *Máté Szarvas and Sadaoki Furui (Tokyo Institute of Technology)*

1. INTRODUCTION

This paper is proposing two extensions to the recent weighted finite-state transducer- (WFST-) based framework of automatic speech recognition [4].

It has been shown in [3] that modeling cross-word phonological constraints in the recognition network may decrease the sentence error rate by 12% in a connected digit recognition task. Modeling of these constraints is supposed to be important in large vocabulary recognition (LVCSR) as well. Up to now, however, it was not possible to evaluate this technique in LVCSR due to the lack of a method to incorporate the rules into the recognition network automatically. The first part of this paper proposes a solution for this problem.

Modeling morphosyntactical constraints can also significantly reduce the recognition error rate. It was found in [5] to reduce the word error rate in a small isolated word recognition task by over 50%. Though both theoretical expectations and this experimental result support the importance of the method, it could not be evaluated on a realistic-size task because of the lack of a large-scale implementation. The second part of the paper describes a method of implementation by WFSTs.

1.1. Review of WFST-based ASR

ASR in the WFST framework can be defined as finding the path with the highest likelihood in the integrated recognition network

$$\text{ASR} = \text{best\_path A} \circ \text{CD} \circ \text{D} \circ \text{LM} \qquad (1)$$

where A is the acoustic model, CD is the context dependency transducer, D is the dictionary transducer and LM is the language model, while $\circ$ represents transducer composition.

One of the strengths of the WFST-based approach to ASR is the ease it provides for combining different knowledge sources (KSs). The main task for incorporating a new type of information into the recognition system is to convert it to WFST format, while the combination is carried out with the elegant composition algorithm [4]. Though the bulk of previous WFST research in ASR has been focusing on optimizing recognition systems based on traditional KSs, in a recent work Tsukada introduced a new component for modeling pronunciation variants, reducing the error rate in his system by 45% [6].

In the following we introduce other two components, one for modeling phonological and the other for modeling morphosyntactical rules.

2. MODELING PHONOLOGICAL CONSTRAINTS

The dictionary transducer is constructed by directly converting a list-form pronunciation dictionary into transducer format. The limitation of this approach to pronunciation modeling is that it cannot accurately describe the phonetic realization of a word sequence because it may depend on the adjoining words.

2.1. An example in Hungarian

The usual strategy to treat words with alternative pronunciations is to include multiple pronunciations in each word entry. Applying this strategy would allow all the pronunciations in the recognition network for all (word) contexts, ignoring the fact that the realized pronunciation depends on the context. For example, if we use the following two pronunciation alternatives of the word "folt" in combination with 2 suffixes:

folt	f o l t	— basic pronunciation
folt	f o l d	— alternative pronunciation
ban	b a n	— suffix with 1 pronunciation
ra	r a	— suffix with 1 pronunciation,

the resulting recognition network would permit the correct pronunciations "foltban → f o l d b a n" and "foltra → f o l t r a", but it would permit the incorrect "foltban → f o l t b a n" and "foltra → f o l d r a" as well.

2.2. Context dependent rewrite rules

The phonological phenomenon illustrated in the previous example is quite regular: it is an example of the more general rule of consonant voicing if they are followed by a voiced stop consonant. Such phonological variations can be described by a set of compact rules for most languages. A popular technique in the linguistic literature for describing phonological regularities is the use of context dependent rewrite rules [2] of the form

$$a \rightarrow b / c\_d \qquad (2)$$

The interpretation of such a rule is as follows. Each of the expressions a, b, c and d denote a regular expression. The meaning of the rule is that a is to be replaced by b whenever it is preceded by c and followed by d.

The following rule describes the phonological constraint in the previous example: t → d / _ b, that is, the phoneme t is replaced by d whenever it is followed by b.

Such rewrite rules are easy to apply to a linear string of phoneme symbols to determine the pronunciation of any particular word sequence, but it is not straightforward to apply them to a recognition network because different rules need to be activated depending on the successor word.

2.3. Implementing rewrite rules by transducers

It has been shown that these rewrite rules can be converted into finite state transducers [2]. And once they are in a transducer representation we can integrate them into our recognition network by the composition operation.

Japanese title is to be printed here

Let P_1, P_2, ..., P_k denote the specific phonological rules of the form in Eq. (2) and let $P = P_1 \circ P_2 \circ ...P_k$ represent the transducer for the sequential application of all of the simple rules. The composition of the phonological transducer P with the dictionary transducer D will result in a phoneme to word transducer that is consistent with the phonological rules of the language.

After incorporating the phonological transducer, Eq. (1) changes into

$$\text{ASR} = \text{best\_path A} \circ \text{CD} \circ \text{P} \circ \text{D} \circ \text{LM}. \quad (3)$$

3. MODELING MORPHOSYNTACTIC CONSTRAINTS

LVCSR systems for languages with a rich morphology must use morphemes as the basic recognition units in order to keep the computational requirements within practical limits. A difficulty in such systems is that the smoothed morpheme N-gram model is assigning a positive likelihood to all morpheme sequences even though many of the possible combinations are yielding words that are not part of the language.

In most languages some of the morphemes are free (they can be used as a word on their own, such as the word "apple" in English), while the use of others is restricted (such as the use of the sign of plural "-s" in English wich can only be used after nouns but never on its own). Moreover, in Hungarian many of the morphemes have variant forms and the variant is selected by the preceding or succeeding morpheme. There are restrictions based on the part of speech of the word as well. For example nominal case suffixes can follow only nominal stems, while verb inflections can follow only verbs.

It has been verified in a small scale experiment in [5] that modeling these constraints may decrease the word error rate significantly. In this section we propose two FST-based techniques for modeling these restrictions in large scale recognition systems.

3.1. Modeling by morphosyntactical network

It has been verified that morphosyntax can be represented by a fintite state automaton (FSA) for languages as diverse as Finnish, Arabic and Hungarian [1]. The nodes in such an automaton are representing the morphemes and the transitions represent the permitted morpheme combinations of the given language. The strength of such a model is that it accepts all of the permitted combinations but none of the forbidden ones. The weakness is, however, that it cannot assign a likelihood to the permitted combinations. These properties are complementary to the morpheme N-gram model that assigns a likelihood to all combinations but cannot refuse the forbidden ones.

Because both of these two knowledge sources are represented in the form of a (weighted) FSA, they can be combined by the composition operation. The resulting weighted FSA will accept only the permitted morpheme combinations and the weight assigned to any such sequence will be the same as that assigned by the original N-gram model. Denoting the FSA representation of the morphosyntactical network by M, equation Eq. (1) will change to

$$\text{ASR} = \text{best\_path A} \circ \text{CD} \circ \text{D} \circ \text{M} \circ \text{LM}. \quad (4)$$

The strength of this method is that it incorporates all morphosyntactical constraints and decreases perplexity as much as is possible. It is not readily applicable for most languages, however, because of the lack of the morphological analyzer.

3.2. Modeling by FSA subtraction

Though developing an analyzer that accepts exactly the correct morpheme combinations may be very costly, it is usually much easier to create FSAs that cover large subsets of the prohibited forms. For example, denoting the FSA representing the set of all nominal stems by N and the FSA for the verbal suffixes by V, the concatenation of these two transducers, $NV = N.V$ would represent all word forms violating the rule that verb suffixes cannot follow nominal stems.

Creating an FSA for the most frequent prohibited combinations and using the algorithm for WFST-FSA subtraction we can relatively easily obtain a modified language model, $LM' = LM - NV$, that refuses many of the impossible combinations while assigning the original weight to the accepted combinations. Because the modified language model refuses the most frequent wrong combinations but none of the correct ones, it is expected that this improvement will decrease the recognition error rate significantly.

4. CONCLUSION AND FUTURE WORK

In this paper we proposed WFST-based methods for incorporating additional knowledge sources into the standard ASR framework. The first method is modeling phonological rules thereby improving the accuracy of the pronunciation model. The other one is dealing with modeling morphosyntactical rules, thereby decreasing the perplexity of morpheme-based language models. Both methods are addressing problems we met originally during the devlopment of a Hungarian LVCSR system but it is likely that the techniques will find applications in systems for other languages as well. Though the advantages of both phonological and morphosyntactical modeling have already been verified in small scale experiments [3][5], they could not be evaluated on a larger scale due to the difficulty of implementing the system. With the WFST-based techniques proposed in this paper the implementation of both ideas becomes much easier and our current efforts are directed towards evaluating them in a large vocabulary dictation system for Hungarian.

5. REFERENCES

[1] K. Beesley, L. Karttunen. Finite-State Morphology: Xerox Tools and Techniques. Cambridge University Press. 2000.

[2] R. Kaplan, M. Kay. Regular Models of Phonological Rule Systems. *Computational Linguistics*, 20(3):331-379, 1994.

[3] P. Mihajlik, T. Fegyo, P. Tatai, G. Gordos. Pronunciation Modeling in Continuous Number Recognition. To appear in *Proc. Eurospeech 2001.*

[4] M. Mohri. Finite State Transducers in Language and Speech Processing. *Computational Linguistics*, 23:2, 1997.

[5] M. Szarvas, T. Fegyó, P. Mihajlik, P. Tatai. Automatic Recognition of Hungarian: Theory and Practice. *International Journal of Speech Technology*, 3(3/4):237–251, 2000.

[6] H. Tsukada. Automatic Learning of Weighted Finite-State Transducers based on Context Trees. In *Proc. Workshop on Multi-Lingual Speech Communication.*, Kyoto, 2000.