論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Modeling Morphosyntax with Finite-State Transducers and its Application to Hungarian LVCSR
著者(和文)	古井 貞熙
Authors(English)	Mate Szarvas, Sadaoki Furui
出典(和文)	日本音響学会 2003年春季講演論文集, Vol. , No. 3-4-3, pp. 133-134
Citation(English)	Spring Meeting, Acoust. Soc. Jpn, Vol. , No. 3-4-3, pp. 133-134
発行日 / Pub. date	2003, 3

MODELING MORPHOSYNTAX WITH FINITE-STATE TRANSDUCERS AND ITS APPLICATION TO HUNGARIAN LVCSR *

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

1. INTRODUCTION

Large vocabulary speech recognition systems for several languages have to use morphemes as the basic recognition units. Such systems are frequently suffering from the over-generation property of the smoothed N-gram language model. The source of the problem is that most of the function-morphemes are very short and their unigram likelihood is high. These morphemes are inserted frequently in the recognition result when the correct word is misrecognized. We observed in our Hungarian dictation system that in many cases the result of such recognition errors is an invalid morpheme combination, e.g., a noun suffix attached to a verb.

In this article we describe and evaluate a method that filters out these invalid morpheme combinations. The error rate of the recognizer decreased between 5 and 18% when the baseline N-gram language model was replaced by the proposed *stochastic morphosyntactic language* model.

Our recognizer (as well as the proposed method) is based on the weighted finite-state transducer (WFST) paradigm, therefore in the next section we review the basics of WFST-based speech recognition. Then we describe the proposed language model in Section 3 and present the experimental results in Section 4. We conclude the paper in Section 5 with a brief summary.

2. REVIEW OF WFST-BASED ASR

In WFST-based speech recognition [1] each of the knowledge sources is represented as a weighted finite-state transducer. The search-space of the given task is obtained by combining the basic components using the composition operation of WFSTs. The main components of our current system besides the decoder itself are the acoustic model A, the context dependency mapping CD, the phonological rules P, the basic pronunciation dictionary D, the morphosyntactic rule set MS and the N-gram language model LM_N.

Using these components the recognition task is defined as finding the path with the highest likelihood in the integrated recognition network:

$$\text{ASR} = \text{best\_path}\ A \circ CD \circ P \circ D \circ MS \circ LM_N, \quad (1)$$

where $\circ$ represents transducer composition.

3. LANGUAGE MODELING

3.1. The morpheme N-gram model

The standard approach to morpheme unit based language modeling is to use a morpheme N-gram model. In the first step all the words in the language model (LM) training data are split up to their constituent morphemes. In the second step an N-gram model is estimated using the morpheme sequence instead of the original word sequence.

In this approach the permitted word-forms (morpheme combinations) are represented implicitly by the transition likelihoods of the N-gram model. The advantage of this method is that it is easy to use because only a morpheme analyzer is needed to split up the words of the training data and the LM is estimated automatically.

The disadvantage is that all practical LM estimation algorithms have to apply a smoothing mechanism to avoid assigning zero likelihoods to valid but unseen word or morpheme combinations. Smoothing algorithms, however, cannot distinguish between inflected word-forms missing due to data scarcity and inflected forms prohibited by the rules of the language. As a result, the smoothed language model assigns a positive likelihood to invalid inflected forms, increasing the chance of misrecognition. For example, the back-off bigram model of Figure 1 permits the obviously incorrect "*drank* [Verb+Past]-*s*[Present+Pers3]-*ing*[Geround]" sequence.

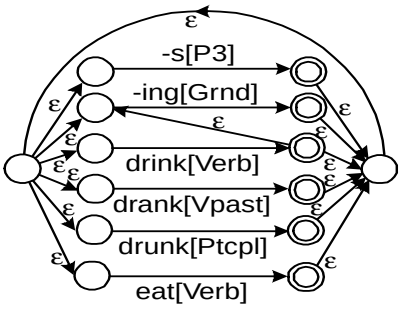

Figure 1. Representation of inflected word-forms by the back-off bigram language model, LM_N.

3.2. The stochastic morphosyntactic language model

The accurate representation of the vocabulary of a language is a recurring problem in many applications. A frequent choice is to build a morphosyntactic grammar that represents the permitted combinations of morphemes. The morphosyntactic grammar can be efficiently implemented as a finite-state automaton (FSA). A simplified example of a morphosyntactic grammar in the form of an FSA is represented in Figure 2. As opposed to the N-gram model example of Figure 1, this model does not permit ungrammatical sequences such as "*drank* [Verb+Past]-*s*[Present+Pers3]-*ing*[Geround]."

This representation is suitable for use in applications where the input is a deterministic character sequence. In speech recognition, however, the input is ambiguous and it is not enough to decide if a particular input sequence is valid or not, but we need to assign likelihoods to different input sequences.

In this regard the stochastic N-gram model and the deterministic morphosyntactic grammar are complementary. The smoothed N-gram model can assign a likelihood to any input sequence, but cannot distinguish between permitted and invalid morpheme sequenes. The morphosyntactic grammar, on the other hand, can only decide if a sequence is permitted or not, but it cannot assign a likelihood to permitted sequences. We would like to have a language model that is accepting exactly those sequences that the morphosyntactic grammar is accepting and to the accepted sequences it assigns the same likelihood as the N-gram model. The finite-state intersection $MS \cap LM_N$ of the two FSA-s has exactly this property: by definition, the finite state intersection of two weighted FSA-s is accepting those sequences that both automata accept. But LM_N is accepting all sequenes, therefore $MS \cap LM_N$ is accepting the same set as MS. And the intersection of two weighted FSA-s assigns the sum of the original weights to any accepted sequence. Since the unweighted MS assigns a weight of 0 to any sequence, the sum is the weight assigned by LM_N.

* 有限状態トランスジューサに基づく形態論モデル化とハンガリー語LVCSRへの応用。 サルワシュマーテー、古井貞熙 （東工大）

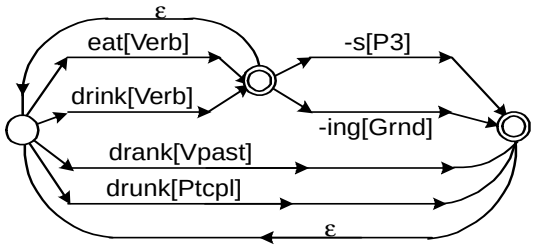

Figure 2. Representation of inflected word-forms by the morphosyntactic grammar, MS.

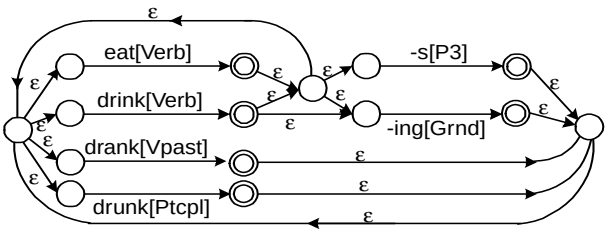

Figure 3. Representation of inflected word-forms by the stochastic morphosyntactic language model, $SMLM$. The morphosyntactic model filtered out the ungrammatical combinations at the price of increasing the size of the network.

Because the resulting language model, $MS \cap LM_N$, integrates the advantages of the stochastic N-gram model and the morphosyntactic model we call it the *stochastic morphosyntactic language model* (SMLM). The SMLM resulting from the intersection of the N-gram model LM_N in Figure 1 and the morphosyntactic grammar MS in Figure 2 is depicted in Figure 3. We can see in the figure, that MS eliminated the invalid combinations from LM_N while retaining the likelihoods of the valid transitions. The final step in making the SMLM a correct language model is to renormalize the weights on the transitions leaving each node because the intersection with MS is removing many transitions from LM_N and the sum of the weights becomes smaller than 1 for many nodes.

4. EXPERIMENTAL EVALUATION

We conducted continuous speech recognition experiments in order to evaluate the usefulness of the proposed stochastic morphosyntactic language model in a real task. The conditions of the experiments are described in detail in [2], therefore here we provide only a brief summary in the interest of saving space.

Conditions. The testing database contained read newspaper sentences from the same Hungarian newspaper that was used for training the LM_N-model (with no overlap between testing data and LM-training data). We run two sets of experiments with different vocabulary sizes. The first set used a 1350 word closed vocabulary while the second set used a 25k word open vocabulary.

The acoustic models used in the experiments were speaker and gender independent triphone HMMs trained with 1 hour of read speech from 30 speakers. The phoneme recognition error rate was 39.13%. This relatively high phoneme error rate indicates that the 1 hour training speech was not enough to train the acoustic models robustly.

1350 word closed vocabulary case. We precompiled recognition networks using 4 different language models: a bigram and a trigram model to serve as the baseline, and the corresponding stochastic morphosyntactic models for both cases. The final normalization step described at the end of Section 3.2 was not applied in the case of the SMLM-s. The size of the 4 networks is displayed in Table 1. We can see that the application of the morphosyntactic model, MS, increased the size of both networks by more than a factor of 2. The reason is that the intersection with the MS model introduced several new back-off nodes like the starting node of the two suffixes in the network of Figure 3.

Table 1. Number of arcs in the precompiled recognition network for different language models (k=thousand).

	N-gram LM		SMLM
Bi-gram	582 k	→	1144 k
Tri-gram	1085 k	→	2504 k

Table 2. Comparision of speaker independent morpheme error rates with different language models (1350 word vocabulary).

	N-gram LM	SMLM
Bi-gram	21.20%	18.89% (-10.9%, rel.)
Tri-gram	17.97%	14.75% (-17.9%, rel.)

Table 3. Comparision of speaker independent morpheme error rates with different language models (25k vocabulary).

	N-gram LM	SMLM
Tri-gram	25.10%	23.75% (-5.38%, rel.)

The recognition error rates for the 4 cases are displayed in Table 2. The application of the morphosyntactic grammar decreased the error rate of the bigram model by about 11% relatively. The use of a trigram LM decreased the error rate even more with a slightly smaller network size. However, the improvement provided by trigrams was completely independent from the improvement by morphosyntax, since the use of MS together with the trigram model gave an additional 18% relative improvement, larger than in the case of the bigram model.

25k open vocabulary case. The previous results suggest that the use of the SMLM can significantly improve the accuracy of morpheme-based speech recognition systems. The evaluation domain, however, was quite limited. Therefore we repeated the experiments on a 25k vocabulary task to see how these results generalize for larger tasks. We can see from Table 3 that the use of morphological information proved useful on this larger task as well, though the improvement was much smaller than on the 1350 word task. We did not finish yet the analysis of the errors but preliminary investigations suggest that the reduced improvement is related to the OOV words. In the case of an OOV word the weaker N-gram model could recover from the errors quickly but the stronger SMLM introduced a longer sequence of errors in order to maintain grammaticality.

5. CONCLUSION

In this paper we introduced the novel stochastic morphosyntactic language model that integrates the advantages of N-gram models and morphosyntactic grammars in morpheme-unit based speech recognizers. The main difference of our model to the standard N-gram model is that our model can utilize the strong connection constraints between different morpheme classes, whereas the standard (smoothed) N-gram model permits any combination. We evaluated the method in a Hungarian dictation task and the use of the morphological grammar reduced the baseline N-gram morpheme error rate between 5 and 18%.

6. REFERENCES

[1] M. Mohri, F. Pereira, M. Riley. Weighted Finite-State Transducers in Speech Recognition. In *Proc. ISCA Automatic Speech Recognition 2000.*, pp. 97–106.

[2] M. Szarvas, S. Furui. Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes. In *Proc. ICSLP 2002.*, pp. 1297–1300.