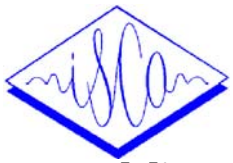

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

# An Integrated Language Modeling with n-gram model and WA model for Speech Recognition*

Shuwu ZHANG , Taiyi HUANG
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100080, P.R.China
Email: {zsw,huang}@prldec3.ia.ac.cn

10.21437/Eurospeech.1997-681

## ABSTRACT

As to traditional n-gram model, smaller n value is an inherent defect for estimating language probabilities in speech recognition, simply because that estimation could not be executed over farther word association but by means of short sequential word correlated information. This has an strong effect on the performance of speech recognition. This paper introduces an integrated language modeling with n-gram model and word association model (abbreviated as WA model). This model integrated two kind of joint probabilities, traditional n-gram probability and word association probability, to estimate actual output probability. WA model are based on a combined probability estimation of orderly word association without distant and strict sequential limitation. In addition, two kinds of local linguistic constraints have also been incorporated into n-gram estimation for smoothing date sparse and adjusting special language unit score locally. A substantial improvement for the performance of Chinese phonetic-to-text transcription in speech recognition has been obtained.

## 1. INTRODUCTION

At present, n-gram language model has been regarded as an effective solution for language decoding in speech recognition. But there still are some peculiar problems we have to process for this model. First of all, because of limited by smaller n value, n-gram language model is difficult to express farther language constraints. In addition, some other problems we concerned are : I. how to smooth data sparse in n-gram. II. some words (such as affixes, surnames, numbers, measures, etc. ) are very powerful to be combined with other words at random but are more scattered distribution in statistical data. So these words can not be ensured with accurate estimation only by current n-gram model.

In this paper, we proposed an integrated language model with n-gram model and WA model. WA model is a kind of word association probability model without the limitation of distance and strict time point. It is quite available for expressing combined relation between words within sentence. It was overlapped on n-gram model by the multiplicity of two joint probabilities.

Meanwhile, we also took two level integration of linguistic constraints in n-gram model to smooth data sparse and adjust probability score of particular language unit respectively. Taking this integrating information, performance of language model has been improved significantly. Some experimental results are shown in comparison with traditional schemes of solution .

## 2. IMPROVED N-GRAM LANGUAGE MODEL

### 2.1 Traditional n-gram

n-gram model has been used extensively as Language Modeling for speech recognition. In general, its mathematical expression can be written as: $P(W) \approx$

$$\prod_{i=1}^{n-1} P(w_i / w_1,\ldots,w_{i-1}) \cdot \prod_{i=n}^{m} P(w_i / w_{i-n+1},\ldots,w_{i-1}) \qquad (1)$$

In practice, the probability $P(w_i / w_{i-n+1},\ldots,w_{i-1})$ is impossible to estimate, simply because of limited by computing complexity . So true application is to keep the last few words just like n value of n-gram is equal to 2 or 3. Thus, the probability $P(w_i / w_{i-n+1},\ldots,w_{i-1})$ is approximated by $P(w_i / w_{i-1})$ or $P(w_i / w_{i-2}, w_{i-1})$.

Sometimes, some discounting model or POS-based n-gram model can be also applied to avoid the problem of zero probability and smooth data sparse as formulas (2 ) and ( 3 ).

$$P'(w_i / w_{i-2}, w_{i-1}) \approx \lambda_0 p(w_i) + \lambda_1 p(w_i / w_{i-1}) + \lambda_2 p(w_i / w_{i-2}, w_{i-1})$$

$$\lambda_0 + \lambda_1 + \lambda_2 = 1 \qquad (2)$$

and

* Supported by National Natural Science Foundation of China. No. 69575018

$$P(w_i / g(w_i) = g_j, g_{i-2}, g_{i-1})$$
$$\approx P(w / g_j) \times (\alpha_0 P(g_j / g_{i-1}) + \alpha_1 P(g_j / g_{i-2}, g_{i-1}))$$
$$\alpha_0 + \alpha_1 = 1 \qquad (3)$$

### 2.2 Smoothing data sparse using weighted average discounting among similar words

For statistical method, data sparse is a major difficulty. Although POS-based n-gram (as shown above in formula (3) ) could play a smoothing function in some extent, but it would bring larger information loss and probably cause reduction of recognition accuracy simply because it compact information from a large space of words to a relative small space of POS. In this paper, we suggested a kind of weighted average discounting among similar words to do smoothing. The basic idea of this method is to make estimation with weighted average of all its similar words instead of single word. The smoothing formula based on weighted average discounting among similar words could be shown as :

$$\hat{p}(w_i / w_{i-n+1} \ldots w_{i-1}) = \frac{f(w_{i-n+1} \ldots w_{i-1}, w_i)}{NS} p(w_i / w_{i-n+1} \ldots w_{i-1}) + \sum_{v \in sim(w_i)} \frac{f(w_{i-n+1} \ldots w_{i-1}, v)}{NS} p(v / w_{i-n+1} \ldots w_{i-1}) \qquad (4)$$

Where,

$$NS = f(w_{i-n+1} \ldots w_{i-1}, w_i) + \sum_{v \in sim(w_i)} f(w_{i-n+1} \ldots w_i, v)$$

$sim(w_i)$ denotes a word set in which each word is similar to word $w_i$, and $f(w_{i-n+1}, \ldots, w_{i-1}, w_i)$ denotes the co-occurrence frequency of $w_i$ and its predecessor word string.

Especially , a trigram discounting is :

$$\hat{p}(w_i / w_{i-2}, w_{i-1}) = \frac{f(w_{i-2}, w_{i-1}, w_i)}{NS} p(w_i / w_{i-2}, w_{i-1}) + \sum_{v \in sim(w_i)} \frac{f(w_{i-2}, w_{i-1}, v)}{NS} p(v / w_{i-2}, w_{i-1}) \qquad (5)$$

### 2.3 Improving bi-gram estimation with the mixture of words and POS

In the study of language processing, we have noticed that some of special words ( such as surnames, place names, organization names, prefix, suffix, numbers as well as measures, etc. ) have definite influence for the correction rate of language recognition. Part of errors are often occurred in there. This is mainly due to the features of their scattered distribution and powerful combination with others at random. But a good feature of them is their excellent class coherence.

In reality, each of these words can be put into a corresponding equivalence class. So we can develop only several special POS (parts of speech) for those words and only estimate the bigram discounting probabilities on their POS. Thus, an improved estimation with the mixture of words and POS can be shown as formula ( 6 ).

$$P^*(w_i / w_{i-2}, w_{i-1}) \approx \begin{cases} \lambda_0 P(w_i) + K_t \times \lambda_1 \times P(g_t / w_{i-1}) + \lambda_2 \hat{P}(w_i / w_{i-2}, w_{i-1}) \\ \quad if \ \ w_i \in g_t \ in \ \ G \ \ \& g_t \ \ with \ \ Lc; \\ \lambda_0 P(w_i) + K_l \times \lambda_1 \times P(w_i / g_l) + \lambda_2 \hat{P}(w_i / w_{i-2}, w_{i-1}) \\ \quad if \ \ w_{i-1} \in g_l \ in \ \ G \ \ \& g_l \ \ with \ \ Rc; \\ \lambda_0 P(w_i) + K_l \times K_m \times \lambda_1 P(g_l / g_m) / + \lambda_2 \hat{P}(w_i / w_{i-2}, w_{i-1}) \\ \quad if (w_i \in g_l \ \& \ w_{i-1} \in g_m) \ \& \ \ g_l, g_m \ in \ \ G \\ \quad \& \ \ g_l \ \ with \ Lc, \ \ g_m \ with \ \ Rc; \\ \lambda_0 P(w_i) + \lambda_1 P(w_i / w_{i-1}) + \lambda_2 \hat{P}(w_i / w_{i-2}, w_{i-1}) \\ \quad otherwise. \end{cases}$$
$$\lambda_0 + \lambda_1 + \lambda_2 = 1 \qquad (6)$$

In the formula, $g_i$ is a kind of POS in POS set G, $Lc$ is denoted that a POS has a property of close correlation with some words on the left of it, and similarly, $Rc$ means a close right correlation. Meanwhile , $K$ is an appropriate discount proportion for a POS, and simply it may be equal to reciprocal of the number of words belong to the corresponding POS or weighted average with those words. That is $K = 1 / \sum_{w_i \in g} c_i \times f(w_i)$ .

Probably, there is different $K$ value for different class $g$ . on the other hand, $\lambda_0 + \lambda_1 + \lambda_2 = 1$ should be equal to 1 for normalization. $\hat{p}(w_i / w_{i-2}, w_{i-1})$ can be referred to formula (5 )

## 3. WORD ASSOCIATION (WA) LANGUAGE MODEL

As we have seen from above, n-gram is to estimate output probability only by adjacent words in time order. So it is difficult to reflect some word associations in father distance. However, this loose word association could be a kind of very useful information for the determination of word sequence. Thus, based on a large corpus, we can neglect the limitation of distance and time point between words and count only these word associations within the compound sentences, then incorporate this kind of information source into the statistical language model for a joint probability estimation. As to using of word association, there ever were some methods has been suggested [6] [7]. Those methods still considered the distant limitation between words directly or indirectly. Here, we introduce an improved method to estimate word association probability.

**DEFINTION:**

Let $w_i$ denote the current word, $w_h =< w_1,...,w_{i-1} >$ is its historical word set. If some words among $w_h$ have combined association with $w_i$ and their combined probabilities satisfy $a(w_j^h, w_i) \geq \delta$, we identify them as the combined association set of $w_i$, which be denoted as $< w_{i1}^h,..., w_{il}^h >$.

For example: a sentence,

*We try to use a kind of new method to solve this problem.*

If having known that each of word set $try, use, method, solve$ has combined association with word *"problem"*, then we regard word set $< try, use, method, solve >$ as the combined association set of $w_i$.

The joint output probability of word sequence by word association information still can be formulated as:

$$a(w_1,...,w_n) = p(w_1)\prod_{i=2}^{n} p(w_i / w_{i1}^h,..., w_{il}^h) \quad (7)$$

In fact, this joint probability can be regarded as another language model like n-gram model. We called it word association model (abbreviated as WA model). In WA model, local conditional dependencies can be decomposed into pairwise interaction between $w_i$ and its association words and expressed as:

$$a(w_i / w_{i1}^h,..., w_{il}^h) \approx \sum_{m=1}^{l} k_m a(w_i / w_{im}^h) \quad (8)$$

Where, $a(w_i / w_{im}^h) = \dfrac{N_{wa}(w_{im}^h, w_i)}{N_{wa}(w_{im}^h)}$ has stochastic constraint: $\sum_w a(w / v) = 1$ for each $v$. $N_{wa}(w_{il}^h, w_i)$ is the count of word combined association between $w_{il}^h$ and $w_i$. Without the distance limitation between words, it can be obtained by the statistics of combined counting of each two words within the compound sentence space.

The weight $k_m = \dfrac{C_{2l}^m}{\sum_{t=1}^{m} C_{2l}^t}$ can be scaled by coefficients valued of binomial distribution, and should satisfies the stochastic constraint: $\sum_{m=1}^{l} k_m = 1$. It is a half of bell-shaped window and decreased progressively with far away $w_i$.

## 4. INTEGRATED LANGUAGE MODELING WITH N-GRAM MODEL & WA MODEL

Based on above two model, we can integrate two kind of different information , sequential n-gram joint probability $p(w_1,...,w_n)$ and orderly word association joint probability $a(w_1,...,w_n)$, to compute the integrated joint log-probability with:

$$\log P_{joint}(w_1,...,w_n) = \log[p(w_1,...w_n) \times a(w_1,...,w_n)] \quad (9)$$

According to formula (1) and (7), local conditional log-probabilities can be inferred as:

$$\log P(w_i / \cdot) = \log[p^*(w_i / w_{i-2}, w_{i-1}) \times a(w_i / w_{i1}^h,..., w_{il}^h)] = \log p^*(w_i / w_{i-2}, w_{i-1}) + \log a(w_i / w_{i1}^h,..., w_{il}^h) \quad (10)$$

$p^*(w_i / w_{i-2}, w_{i-1})$ can be referred to formula (6), and $a(w_i / w_{i1}^h,..., w_{il}^h)$ to formula (8). Thus, this integrated model is more powerful to predict succeed word.

In practical application, we can set up a temporary cache memory to store these historical words within a compound sentence and estimate the correlation value of current candidate with all of its associated words.

## 5. EXPERIMENTS & RESULTS

Based on above integrated model, some experiments have been conducted for the transcription from phonetic string to word string in Chinese speech recognition.

Corpus was derived from full text of "People's Daily" in 1993 with about 30,000,000 Chinese characters. Based on the corpus, word collocation information with relative frequencies ( including about 50,040 words, 2.4 M word pairs, and 6.9 M word triplets ) were counted. In the training procedure of WA model, by selecting 2M Characters from the corpus, we got 7.2 M word association pairs without the limitation of distance. Pruned out parts of pairs with lower frequency less than 10 times, there still were 0.82M word association pairs with relative frequency which have been used as the basic information for WA modeling.

The word similarity could be measured in terms of the semantic classes. We have introduced an approach to designate Chinese words into 1898 semantic classes [8]. These classes can be used as similar word set for

smoothing sparse data by weighted average discounting. Local POS consisted of 17 special classes with the property of $Lc$ or $Rc$. Near 3,000 words were put into relevant local POS which takes about 8.9 percent of whole vocabulary. Testing material was selected from other newspapers which were similar in style to training corpus.

The preliminary results can be seen from table I.

| Models | | Transcription Accuracy (%) |
|---|---|---|
| n-gram Model | Interpolation Bigram | 85.30 |
| | Interpolation Trigram | 89.58 |
| | Improved Trigram | 92.63 |
| WA model only | | 82.94 |
| Integrated Modeling | | 95.43 |

*Table I. Performance Comparison of various models*

In table I, It has been shown that there was a 29.7% reduction of transcription error by replacing traditional trigram with improved trigram ( specified in section II) and about 56% reduction with integrated model. So, It can be said that integrated model has a significant improvement for performance of Chinese phonetic-to-text transcription in speech recognition.

It has also been shown that the result of independent WA estimation was relative low than that of bigram, because it took smaller training data than n-gram. In fact , WA model should be theoretical more powerful than bigram, if there were the same large training corpus. .

## 6. CONCLUSION

We have presented an integrated language modeling for speech recognition. This model combined n-gram probability estimation with a new WA model. Meanwhile, in improved n-gram model, a weighted average discounting method based on similar words was used to smooth local trigram estimating in the interpolation trigram model, and a mixed estimation with word and local POS also was applied as local bigram estimating in the same model for adjusting particular units' score. It has been shown that incorporating WA model as well two kinds of local constrains in n-gram has a substantial improvement for performance of phonetic-to-text transcription in Chinese speech recognition.

In current integrated model, word association probabilities were used as only a complement for n-gram. In fact, with the characteristic of unlimited in distance and time point, WA model could capture more and father word correlated information including bigram probabilities. So it should be theoretical more powerful than bigram even trigram alone, if there were the same large training data. Meanwhile, WA model could be represented as almost uniform with n-gram model in the procedures of training and estimating, so it is convenient to compute the output probability as same as that of traditional model without almost additional computing.

Furthermore, some optimal techniques for WA model are being considered. Some further experiments are also being taken to compare the performance of WA model alone with n-gram or integrated model.

## REFERENCES


[1] Bo Xu, Bin Ma, Shuwu Zhang, Taiyi. Huang, "Speaker- independent Dictation of Chinese Speech with 32K Vocabulary ", *ICSLP96, Oct. 1996, USA*

[2] A. M. Derouault and B.Merialdo, "Natural Language Modeling for phoneme-to- text transcription", IEEE Trans. Pattern Anal. Machine Intell., Nov 1986.

[3] M. Federico, et al. "Language modeling for efficient beam-search", Computer Speech and Language (1995)9

[4] F.Jelinek, R.L. Mercer and L.R. Bahl, "A maximum likelihood approach to continuous speech recognition " , *IEEE Trans. Pattern Anal. Machine Intell. , Mar. 1983.*

[5] R. Kuhn, R. De Mori, "A cache-based natural language model for speech recognition" , *IEEE Trans. Pattern Anal. Machine Intell. , June 1990.*

[6] R.Rosenfeld, "Adaptive statistical language modeling: A Maximum Entropy Approach", Technical report in CMU,1994.

[7] H.Ney, U.Essen and R. Kneser, " On structuring probabilistic dependencies in stochastic language modelling", *Computer Speech and Language (1994),8,1-38.*

[8] Shuwu Zhang, Taiyi Huang, "Tagging Semantic Markers for the Word in the Practical Dictionary", *Intl. Conf. on Chinese Computing, June 1996,Singapore.*