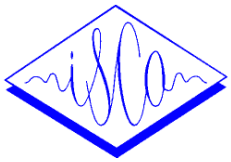
Practical Language Modeling: An Interpolating Method



Xiaohu Liu and **Douglas O'Shaughnessy**
INRS-Telecommunications, University of Quebec
{lxiaohu, dougo}@inrs-telecom.uquebec.ca

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

1 Abstract

Language modeling is a key component in speech and handwriting recognition. N-gram language modeling is used as the formalism of choice for a wide range of domains. Although a high order N can reduce perplexity greatly, it is unrealistic in many practical cases to get statistically reliable N-grams. We propose an interpolated model by introducing signal words and clue words into the baseline N-gram model. The initial word in a word pair with high mutual information is chosen as a signal word. In the same way, we define such words that have high mutual information with a certain morphological form as clue words. In a given context, we select a signal word with the highest score to compute the probability of the current word, and a clue word with the highest score to estimate the probability of the form of the current word. We discuss the basic requirements of designing an interpolating language model and see how our models satisfy the requirements. We got considerable reduction in perplexity, compared to the baseline model. Because both signal words and clue words are easy to collect and handle, the proposed method is practical.

2 Introduction

Language modeling (LM) is very important for research involved in natural language. One of the most successful language models for speech recognition (SR) is the statistical language model (O'Shaughnessy et al., 1997). Statistical language modeling attempts to identify regularities in natural language and capture them in a statistical model.

For the task of SR, the purpose of the decoder is to find the sequence of words W having the largest posterior probability given the acoustic observation A (O'Shaughnessy et al., 1997). The probability $P(W)$ and $P(A|W)$ are computed by a language model (LM) and acoustic model respectively.

In the commonly used trigram model, the distribution of the current word is highly dependent on the previous two words, and the posterior probability can be written as

$$\begin{aligned} P(W) &= P(w_1, w_2, \ldots, w_m) \\ &\approx \prod_{i=1}^{m} P(w_i|w_{i-2}, w_{i-1}). \end{aligned} \quad (1)$$

How to compute the probability of $P(w_i|w_{i-2}, w_{i-1})$ is the central problem in a trigram language model. The difficulty comes from the scarcity of training data. Many smoothing methods and class based models have been adopted to solve the problem (Brown et al., 1990)

Even if there are enough training data for trigram modeling in a specific application, research shows that the trigram model is not good enough to grasp long-distance language dependence (Bellegarda, 1998). One solution is to use a high-order N-gram model. However, the parameter space increases exponentially with N. Theoretically, it is $|V|^n$. Another alternative way is to use a distance language model (Langlois and Smaili, 1999). In that case, the multi-conditional probability $P(w_i|w_1, \ldots, w_{i-1})$ is regarded as a combination of each $P(w_i|w_j)$, where $j = 1, \ldots, i-1$. Additionally, there has been

increasing interest in variable-order N-gram models (Riccardi et al., 1995; Guyon and Pereira, 1995). Their models can represent a long distance history, but models are usually very large.

Before introducing our statistical language model, we discuss the basic requirements of designing an interpolating language model. Using the interpolating method, we introduce distribution information of signal words and clue words into the baseline model and get our improved language model. We prove our model satisfies the basic requirements.

We give experimental results in section 4 and make a conclusion in the last section.

3 Improved interpolated models

3.1 Basic requirements of statistical language models

The critical problem to design a language model is to find a way to obtain the estimation of conditional probability, which is defined as $\hat{P}(w_i|h_i)$, where h_i denotes the history. There are two basic requirements that must be satisfied to design a logical and effective language model.

First, the sum of all posterior probabilities of all words given a certain context should be 1.0, that is,

$$\sum_i \hat{P}(w_i|h) = 1.0 \tag{2}$$

where h is any given history (context) and w_i is any word in the vocabulary.

Second, the sum of probabilities of all possible sentences must be 1.0. That means

$$\sum_i \hat{P}(W_i) = 1.0 \tag{3}$$

where W_i is any possible sentence.

It is easy to see the two requirements are natural statistic laws, so any logical language models must give good estimation to sentences so that requirements are satisfied.

If the sum of all posterior probabilities is larger than 1.0, we will get a smaller perplexity, but it does not necessarily mean the language model is better for speech recognition.

In addition, from the viewpoint of language model evaluation, certain requirements are needed. The entropy can be calculated with

$$H = -\frac{1}{N}\sum_W P_T(W) log_2 \hat{P}(W) \tag{4}$$

where $P_T(W)$ is the probability computed according the testing data, and $\hat{P}(W)$ is the probability estimated from the language model.

If $\sum_i \hat{P}(w_i|h) > 1.0$, or $\sum_i \hat{P}(W_i) > 1.0$, the probability of $\hat{P}(W)$ is overestimated, a lower entropy and smaller perplexity are obtained. However, it is hard to say this language model is better even though it gives larger probability to sentences. With this language model, it is still hard to choose best candidate sentences from the N-best list because it gives larger and false probability to most of the candidates. A good language model should estimate the likely sentence with higher probability, and give lower value to the unlikely sentences.

There are many reasons that a language model with lower perplexity does not necessarily help a recognizer. One of the possible reasons is that the model does not satisfy the basic requirements.

3.2 Signal words and clue words

We believe a good language model should take local and global constraints into account and utilize several levels of knowledge (Bellegarda, 1998). We improve the basic N-gram model by introducing two kinds of words into our models: signal words and clue words.

A word is called a signal word in the sense that the occurrence of the signal word implies a special context or scenario. This word acts like a signal. Concretely, a word is chosen as a signal word if the word has high mutual information in a word pair. For a word pair (w_f, w_s), the mutual information$MI(w_f, w_s)$ is computed according to the frequency of distance bigrams

(Langlois and Smaili, 1999; Liu et al., 1999), and the frequency of the words in the word pair.

If $MI(w_f, w_s) > \theta$, where θ is a threshold, the word w_f is regarded as a signal word. All the signal words are collected from the statistics of training data, and constitute a set of signal words.

In the same way, we define such words that have high mutual information with a certain morphological form as clue words. For instance, the clue word *"have"* prefers an *"-ed"* form rather than the infinitive form for a following verb.

3.3 Improved model

By introducing distribution information of signal words and clue words into the baseline mode, we get our improved language model:

$$\begin{aligned} P_m(w_i|w_{i-2}, w_{i-1}) &= \alpha_1 P_b(w_i|w_{i-2}, w_{i-1}) \\ &\quad +\alpha_2 P(w_i|w_{signal}) \\ &\quad +\alpha_3 P(w_i|w_{clue}) \quad (5) \end{aligned}$$

where $P_b(w_i|w_{i-2}, w_{i-1})$ is the probability given by the baseline model, and $\alpha_1 + \alpha_2 + \alpha_3 = 1.0$. w_{signal} and w_{clue} are the signal word and clue word chosen from the current context according to a criterion of maximum mutual information or minimum entropy.

Our improved language model satisfies the first requirement, because

$$\begin{aligned} &\sum_{w_i} P_m(w_i|w_{i-2}, w_{i-1}) \\ = \;& \alpha_1 * \sum_{w_i} P_b(w_i|w_{i-2}, w_{i-1}) \\ &+\alpha_2 * \sum_{w_i} P(w_i|w_{signal}) \\ &+\alpha_3 * \sum_{w_i} P(w_i|w_{clue}) \\ = \;& \alpha_1 + \alpha_2 + \alpha_3 \\ = \;& 1.0 \quad (6) \end{aligned}$$

It satisfies the second requirement too. We prove the assumption using an inductive method on the basis of length of the sentence. Let the length of the sentence be l, and it is easy to see the assumption is true when $l = 1, 2$. Suppose it is true when $l = k$, where $k > 2$, and we only need to prove that the assumption stands when $l = k + 1$.

$$\begin{aligned} &\sum_{W} P_m(w_1, w_2, \ldots, w_k, w_{k+1}) \\ = \;& \sum_{w_1 \ldots w_k} \sum_{w_{k+1}} P_m(w_1, w_2 \ldots w_k) \\ &*P_m(w_{k+1}|w_{k-1}, w_k) \\ = \;& \sum_{w_1 \ldots w_k} \sum_{w_{k+1}} P_m(w_1, w_2 \ldots w_k) \\ &*\alpha_1 * P_b(w_{k+1}|w_{k-1}, w_k) \\ &+ \sum_{w_1 \ldots w_k} \sum_{w_{k+1}} P_m(w_1, w_2 \ldots w_k) \\ &*\alpha_2 * P(w_k|w_{signal}) \\ &+ \sum_{w_1 \ldots w_k} \sum_{w_{k+1}} P_m(w_1, w_2 \ldots w_k) \\ &*\alpha_3 * P(w_k|w_{clue}) \\ = \;& \sum_{w_1 \ldots w_k} \sum_{w_{k+1}} P_m(w_1, w_2 \ldots w_k) \\ = \;& \sum_{w_1 \ldots w_k} P_m(w_1, w_2 \ldots w_k) \\ = \;& 1.0 \quad (7) \end{aligned}$$

4 Experiments

Our experiments are performed on the Switchboard corpus, which includes transcriptions of 2,280 scenarios. We will try four language models depending on whether the signal or clue words are used or not. We compare the performance of different models on variable sizes of training data.

We train the four kinds of language models with variable sizes of training data, and test the models on the same testing data. The experimental results are illustrated in the following graph.

It is obvious that the entropy is decreasing as the size of training data increases. When signal words or clue words are combined into the baseline model, we get better performance of our improved interpolated model. If both signal words and clue words are used, we get the best performance. The importance is more distinctive when we use less training data. When the size of training data is more than 2 million words, the

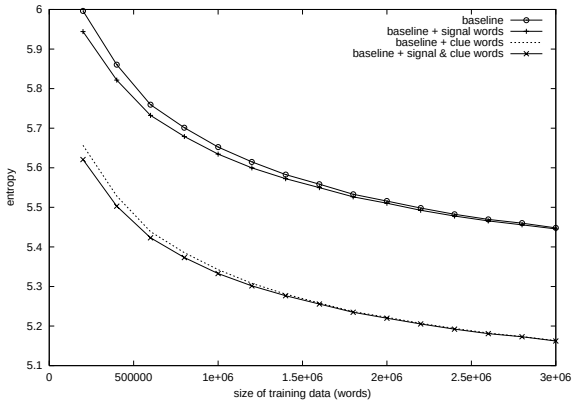


Figure 1: Evaluation results of language models

effect of signal words is very small, but the clue words are still very useful. It shows us that higher level knowledge, such as morphological knowledge, is more useful than lower level knowledge, such as signal words.

We divide the corpora into three parts dynamically, the first is used to compute the occurrence times of trigrams, bigrams and unigrams, the second is used to compute the coefficients and the third part is used to test the performance of language models.

Besides maximum mutual information, we wonder if there is another better way to choose signal words and clue words. We tested another criterion of minimum entropy to select signal words and clue words and we find that mutual information is better.

5 Conclusion

Statistical language modeling is one of the most important directions for speech recognition and handwriting recognition areas. N-gram models are the most commonly used recently. We research how to use longer context and to use no extra training data. We argue that there are two basic requirements which must be satisfied to design a logical language model and explain how our improved models satisfy the requirements.

We improve the basic N-gram model by introducing signal words and clue words. Experiments were performed on the Switchboard corpus, and we got a remarkable reduction in perplexity, compared to the baseline model. Since both signal words and clue words are easy to collect and handle, the proposed method is practical. We will perform more experiments on other corpora and test the performance of our language model in a speech recognizer.

References

D. Langlois and K. Smaili. A new based distance language model for a dictation machine:application to MAUD. 1999. *Eurospeech'99, pp. 1779-1782*

D. O'Shaughnessy, Z. Li, A. Farhat, R. El Meliani, R. Vergin and M. Héon. Recent progress in automatic recognition of continuous speech. 1997. *Canadian Conference on Electrical and Computer Engineering, vol. 1, pp. 51–54*

G. Riccardi, E. Bocchieri and R. Pieraccini. Non-deterministic Stochastic Language Models for Speech Recognition. 1995. *Proceedings of the ICASSP, pp. 237-240*

I. Guyon and F. Pereira. Design of a Linguistic Postprocessor using Variable Memory Length Markov Models. 1995. *Proceedings of the 3rd ICDAR, Montreal, pp. 454-457*

J. R. Bellegarda. Exploiting both local and global constraints for multi-span statistical language modeling. 1998. *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 677-680*

P. F. Brown, V. J. Della Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. Class-based N-gram models of natural language. 1990. *Proceedings of the IBM Natural Language ITL, Paris, France, March 1990*

X. Liu, P. Fung and C. S. Cheung. A Monolingual Semantic Decoder Based on Word Sense Disambiguation for Mixed Language Understanding. 1999. *Eurospeech-99, Vol. 5, pp. 2011-2014*