2005－SLP－57（3）
2005／7／15

トリガー言語モデルの適応によるパネル討論の音声認識

カルロス・トロンコーソ　河原達也

京都大学情報学研究科
〒606-8501 京都市左京区吉田二本松町

{carlos,kawahara}@ar.media.kyoto-u.ac.jp

あらまし

パネル討論の音声認識を対象として、トリガーモデルを用いた言語モデル適応法を提案する。パネル討論では、与えられた話題について終始話されるので、遠距離でもキーワードの相関が期待できる。トリガー言語モデルはそのような遠距離の依存関係をとらえるためのものであるが、従来は新聞記事などの一般的すぎる大規模コーパスから構築されており、タスクに依存したトリガーペアが十分に得られない。提案手法では、ベースラインモデルによる初期認識結果を使用して、当該討論に特化したトリガーペアを抽出し、またそれらの確率を推定する。確率値については、大規模コーパスから推定される統計量も利用するバックオフ手法も提案する。実験の結果、大規模コーパスから作成した通常のトリガー言語モデルと比較して、テストセットパープレキシティを約 2 倍削減できた。さらに、トライグラム言語モデルの適応と組合わせることにより、ベースラインよりパープレキシティを 41%削減できた。

Automatic Transcription of Panel Discussions Using Trigger-Based Language Model Adaptation

Carlos Troncoso and Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501

{carlos,kawahara}@ar.media.kyoto-u.ac.jp

Abstract

We present a novel trigger-based language model adaptation method oriented to the transcription of meetings. In meetings, the topic is focused and consistent throughout the whole session, therefore keywords can be correlated over long distances. The trigger-based language model is designed to capture such long-distance dependencies, but it is typically constructed from a large corpus, which is usually too general to derive task-dependent trigger pairs. In the proposed method, we make use of the initial speech recognition results to extract task-dependent trigger pairs and to estimate their statistics. Moreover, we introduce a back-off scheme that also exploits the statistics estimated from a large corpus. The proposed model reduced the test-set perplexity twice as much as the typical trigger-based language model constructed from a large corpus, and achieved a remarkable perplexity reduction of 41% over the baseline when combined with an adapted trigram language model.

1. Introduction

In automatic speech recognition (ASR), the most widely used language model is the *n*-gram model, where *n* typically ranges from 2 (bigram) to 4 (4-gram). The *n*-gram language model estimates the occurrence probability of *n* consecutive words in the text. This model is known to be effective, but it is apparently limited in scope, because it is unable to model dependencies longer than *n*.

Alternative approaches, such as the trigger-based language model [1][2] and the cache-based language model [3], try to broaden the scope of the *n*-gram by modeling long-distance dependencies between words. The trigger-based language model uses a set of correlated word pairs, known as trigger pairs, to raise the probability of the words "triggered" by others in the word history.

The conventional trigger-based language model has some limitations. This model has been mostly applied to corpora of newspaper articles. This kind of corpora are usually too general in topic and do not closely match the specific test data. Moreover, it has been observed that much of the potential of trigger-based language models lies in words that trigger themselves, called self-triggers. Self-triggers are virtually equivalent to the cache-based language model, so the original trigger-based language model does not significantly outperform the cache-based model.

This paper addresses an effective implementation of the trigger-based language model mainly targeting at a meeting transcription task to overcome the model's limitations. The transcription of meetings and lectures is one of the promising applications of large vocabulary continuous speech recognition. The subject matter in a meeting is fairly homogeneous during it, so we can expect to find keywords related in their topic throughout the whole session. The trigger-based language model could be used to capture these constraints, but typical large corpora such as newspapers are too general to extract task-specific trigger pairs and their statistics. On the other hand, the data from a single meeting session are large enough to extract trigger pairs from them, and we expect that the probabilities of the trigger pairs can be also estimated from these data.

In the proposed approach, we regard a meeting session as a document unit, and the trigger pairs are extracted from the document's initial speech recognition results. The initial transcription, though containing errors, can provide us with useful information about the topic or speaking style of the meeting. In this method, the trigger pairs are selected from the whole meeting to capture global constraints such as topic information, as opposed to the conventional trigger-based language model, where the word pairs are selected from a text window of fixed length. The statistics of the trigger pairs are also estimated from the initial transcription, but they might not be reliable due to the small amount of data. Thus, we introduce a back-off scheme that incorporates information from a large corpus.

The organization of this paper is as follows. Section 2 introduces the proposed trigger-based language model adaptation in detail. Then, an enhancement based on a back-off scheme using a large corpus is proposed in section 3. Their experimental evaluation in a panel discussion task is reported in section 4. A further enhancement by combining with *n*-gram model adaptation is described in section 5.

2. Trigger-based language model adaptation

Figure 1 illustrates the outline of the proposed approach. First, ASR is performed with a standard *n*-gram as the baseline language model, yielding the initial speech recognition results. The trigger pairs are then extracted and their probabilities are also estimated from the initial transcription. Finally, the resulting trigger-based component is combined with the *n*-gram component to produce a new language model.

2.1. Extraction of trigger pairs from initial transcription

A trigger pair is a pair of content words that are semantically related to each other. Trigger pairs can be represented as $A \rightarrow B$, which means that the occurrence of A "triggers" the appearance of B, that is, if A appears in a text, it is likely that B will come up afterwards.

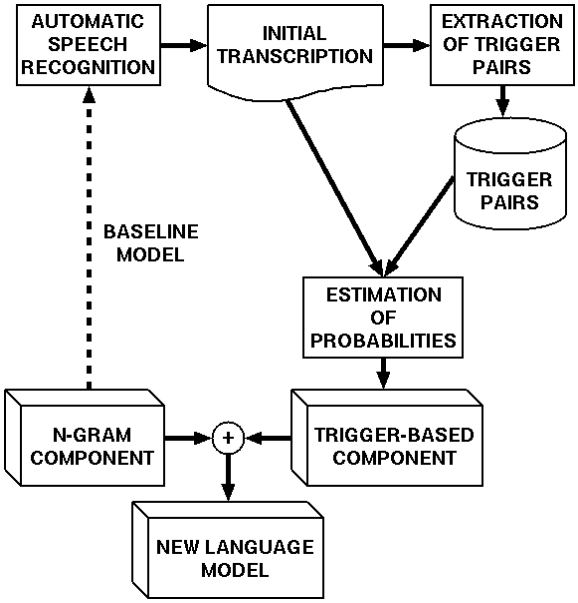

Figure 1: Outline of the proposed approach

Task-dependent trigger pairs are first extracted from the initial transcription, namely the *K*-best ASR hypotheses. For the selection of pairs, instead of the average mutual information used in [1], we use the term frequency/inverse document frequency (TF/IDF) measure [4]. We employ this measure because it is document-based, that is, it lets us extract the trigger pairs from a whole document, rather than from a text window of the corpus. In this way, we can capture global constraints from each document. This measure is also chosen because of its simplicity.

The TF/IDF value of a term t_k in a document D_i is computed as follows:

$$v_{ik} = \frac{tf_{ik} \log(N / df_k)}{\sqrt{\sum_{j=1}^{T} (tf_{ij})^2 [\log(N / df_j)]^2}}, \quad (1)$$

where tf_{ik} is the frequency of occurrence of t_k in D_i, N is the total number of documents, df_k is the number of documents that contain t_k, and T is the number of terms in D_i.

Since the initial transcription intuitively consists of only one document, the TF part is calculated from the *K*-best hypotheses, while the IDF part is computed from a fraction of a large corpus similar to the target task, corresponding to texts (documents) from the same year as the task.

We create all possible word pairs, including pairs of the same words (self-triggers), with the base forms and parts of speech (POS) of all content words with a TF/IDF value above a threshold. POS-based filtering is introduced to discard function words.

2.2. Probability estimation from initial transcription

The probabilities of the trigger pairs are then estimated from the *K*-best hypotheses by using a text window to calculate the co-occurrence frequency of the pairs inside it.

The probability of each trigger pair $w_1 \rightarrow w_2$ is computed as follows:

$$P_{TP}^{IT}(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)}, \quad (2)$$

where $N(w_1, w_2)$ denotes the number of times the words w_1 and w_2 co-occur within the text window, and j runs throughout all words triggered by w_1.

2.3. Proposed trigger-based language model

The proposed trigger-based language model is then constructed by linearly interpolating the probabilities of the trigger pairs with those of the baseline *n*-gram model, so that both long and short-distance dependencies can be captured at the same time.

The probability of the proposed language model for a word w_i given the word history H is computed in the following way:

$$P_{LM}(w_i \mid H) = \frac{1}{|H|} \sum_{h \in H} P_{LM}(w_i \mid h)$$

$$P_{LM}(w_i \mid h) = \begin{cases} P_{NG}(w_i \mid H), & \text{if } P_{TP}^{IT}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda) P_{TP}^{IT}(w_i \mid h), & \text{else} \end{cases} \quad (3)$$

Here $|H|$ means the number of words in the history; P_{NG} is the probability of the *n*-gram component; P_{TP}^{IT} is the probability of the trigger-based component, computed by equation (2); and λ is the language model interpolation weight. When there are no words triggered by h, the *n*-gram model alone is used. Otherwise, the *n*-gram probabilities are linearly interpolated with those from the trigger pairs.

3. Combination with statistics from large corpus

In order to enhance the model's performance, this section introduces a back-off scheme to combine the proposed model with the trigger-based statistics estimated from a large corpus. Since the amount of data provided by the initial transcription may be insufficient to obtain reliable probability estimates, a large corpus is used to cope with this problem.

3.1. Construction of trigger pairs from large corpus

The trigger pairs are first extracted with the TF/IDF measure from a fraction of a large corpus similar to the target task, which is a collection of documents from the same year as the task. By extracting the word pairs from a corpus similar to the target task, we expect to extract trigger pairs that are related in topic to it. This time, the corpus is divided into documents, so the TF/IDF computation is straightforward. Then, the probabilities of the trigger pairs are computed from the whole corpus. We previously demonstrated that the method that selects trigger pairs from a matched corpus and then estimates their statistics with a larger corpus is effective [5].

The resulting trigger pairs are similar to those used in the conventional trigger-based language model, except that the trigger pairs presented here are derived with the TF/IDF measure, instead of the average mutual information, and that they are extracted from a matched portion of the large corpus, instead of from the whole training set.

3.2. Proposed back-off method

Next, we make use of the statistical model derived from the large corpus to complement the proposed model described in section 2. We have two different sets of trigger pairs: the trigger pairs constructed from the initial transcription (hereafter trigger set IT), and the trigger pairs extracted from the large corpus (hereafter trigger set LC). The trigger set IT provides a probability distribution more faithful to the target domain, whereas the trigger set LC offers a more reliable distribution that can cope with the problem of data sparseness discussed in [5].

The probability of the enhanced language model based on the back-off scheme $P_{BO}(w_i \mid h)$ is calculated in the following way:

$$\begin{cases} P_{NG}(w_i \mid H), & \text{if } P_{TP}^{IT}(w_j \mid h) = 0, P_{TP}^{LC}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda) P_{TP}^{LC}(w_i \mid h), & \text{if } P_{TP}^{IT}(w_j \mid h) = 0, \forall j \\ \lambda P_{NG}(w_i \mid H) + (1-\lambda)\left(\delta P_{TP}^{LC}(w_i \mid h) + (1-\delta) P_{TP}^{IT}(w_i \mid h)\right), & \text{otherwise} \end{cases} \quad (4)$$

Here, P_{NG} is the probability of the n-gram component, P_{TP}^{IT} is the probability of the trigger set IT, and P_{TP}^{LC} is the probability of the trigger set LC. When there are no words triggered by h in either of the two trigger sets, the n-gram model alone is used. When there are no trigger pairs for h in the trigger set IT, the n-gram probabilities and the trigger set LC probabilities are linearly interpolated. Otherwise, all language models are linearly interpolated.

Note that if the trigger set IT is empty, that is, if we do not use the trigger pairs extracted from the initial transcription, the resulting model (first two entries in equation (4)) is similar to the conventional trigger-based language model, that is, a model whose trigger pairs are constructed from a large corpus. The differences are those discussed in section 3.1. Hereafter we call this model the quasi-conventional trigger-based language model.

4. Experimental evaluation

4.1. Experimental setup

The task in this work is the transcription of panel discussions from a Japanese TV program called "Sunday Discussion" [6]. The corpus consists of 10 programs of 1 hour recorded from June 2001 to January 2002. The average number of words is 14K.

The ASR system Julius 3.4.2 was used for speech recognition. The baseline language model was a word trigram model constructed from the Corpus of Spontaneous Japanese (CSJ) [7] (3.5M words), and the minutes of the National Diet of Japan [6] (71M words). The size of the vocabulary was 30K words. The average word recognition accuracy with this baseline model was 51.6%. We obtained this relatively low accuracy because the utterances are truly spontaneous and often uttered very fast.

Table 1: Specification of used corpora

Corpus name	Contents	Size
Sunday Discussion	Panel discussions	10 programs of ~14K words
Corpus of Spontaneous Japanese (CSJ)	Extemporaneous speeches	3.5M words
Minutes of the National Diet	Congress meetings	71M words

The minutes of the National Diet from the year 2001 (17M words) were used for calculating the IDF part used in the extraction of the trigger pairs of the set IT, and also to extract the trigger pairs of the set LC.

Table 1 summarizes the corpora used in the present work.

4.2. Parameter optimization

The parameters of all models were optimized by dividing the test set into two. The first 5 programs were used to empirically tune the parameters used in the other 5 programs and vice versa. The parameters were optimized by means of the perplexity.

The resulting average word history size $|H|$ was 26 for the proposed trigger-based model. The optimal language model interpolation weight λ was, on average, 0.56 for the proposed trigger-based model (equation (3)), 0.72 for the quasi-conventional model (equation (4) without last entry), and 0.57 for the back-off method (equation (4)). The value of λ is higher for the quasi-conventional model than for the proposed models, because the trigger pairs are not task-dependent in the former model and, therefore, less beneficial in the interpolation. The optimal trigger set interpolation weight δ was 0.08, and the optimized number of hypotheses from the initial transcription K was 2.

4.3. Perplexity evaluation

We evaluated the test-set perplexity for the 10 programs by three different models: the quasi-conventional trigger-based model using only a large corpus (LC), the proposed trigger-based language model using only the initial transcription (IT), and the back-off method (IT+LC). For reference, we also evaluated the model constructed by deriving the trigger pairs from the correct transcription.

Table 2: Comparison of trigger-based language models constructed by different methods

Model	Perplexity	Reduction (%)
Baseline	95	–
Large corpus (LC)	81	14.74
Initial transcription (IT)	68	28.42
Back-off model (IT+LC)	67	29.47
(cf.) Correct transcription	41	56.84

The perplexity and its reduction averaged over the 10 programs are shown in Table 2. The proposed model achieved a reduction of 28.42%, almost double the reduction obtained with the quasi-conventional model. This demonstrates the effectiveness of the proposed approach.

The back-off method improved the perplexity slightly, but not significantly. This suggests that the initial transcription provides trigger pairs that are much more adapted to the task than those constructed from the large corpus, so the benefit obtained from the latter is minimal. We expect that the proposed back-off scheme can be useful when the initial transcription is smaller in size.

The perplexity reduction by the proposed method was about half that obtained with the model that used the correct transcription. The baseline word recognition accuracy is 51.6%, meaning that about half of the initial transcription is erroneous, so the results are consistent with this fact.

The average number of trigger pairs was 43K in the trigger set IT, 9158K in the trigger set LC, and 14K from the correct transcription. The average coverage of the trigger pairs in the test set was 29% for the first case, 33% for the second, and 34% for the third. We can see that the set IT efficiently covers the test set with a much smaller number of trigger pairs than the set LC. This is because the pairs from the set LC are not task-dependent. The back-off method had slight impact on the perplexity because the coverage by using the set LC is only a little greater than that by the set IT.

The model constructed from the initial transcription used 547 self-triggers on average during the test-set perplexity evaluation, while 19670 non-self-triggers were used. This is a significant difference with the conventional works on trigger-based language models, where non-self-triggers offered little benefit over self-triggers.

Table 3: Comparison and combination of the proposed method with the adapted *n*-gram

Model	Perplexity	Reduction (%)
Baseline	95	–
Adapted *n*-gram	77	18.95
+Initial transcription (IT)	57	40.00
+Back-off model (IT+LC)	56	41.05

5. Comparison and combination with *n*-gram model adaptation

In this section, we use the initial transcription to create an adapted *n*-gram language model in order to compare its performance with that of the proposed approach. We then combine this with the proposed model for further improvement.

5.1. Construction of n-gram from initial transcription

We take the *K*-best hypotheses from the initial transcription and create a text file by adding the *i*-best in order, that is, first all the 1-best hypotheses, then all the 2-best, and so on. This text file is used for creating a back-off *n*-gram model.

A trigram model was constructed from each of the 10 test sets, and then interpolated with the baseline trigram model. The value *K* was optimized with the method discussed in section 4.2, yielding the value 20.

5.2. Perplexity evaluation

The resulting interpolated trigram was combined with the trigger-based language model. Table 3 shows the results of the perplexity evaluation. We achieved a notable maximum perplexity reduction of 41.05% over the baseline trigram model. Although the improvement is not additive, the *n*-gram model adaptation serves as a good complement for the proposed approach.

6. Conclusion and future work

We presented a novel trigger-based language model adaptation based on initial speech recognition results. A significant improvement in perplexity was achieved over both the baseline trigram and the typical trigger-based model constructed from a large corpus. A further improvement was achieved by combining with *n*-gram model adaptation.

The proposed approach is particularly useful in tasks where large amounts of training data are not readily available but the test set is long, since we have observed that the initial transcription is a good source for deriving the trigger pairs. This is specifically true for many transcription tasks.

We plan to incorporate the proposed language model into the decoder of the ASR system and evaluate it in speech recognition experiments.

7. References

[1] R. Rosenfeld, “A Maximum Entropy Approach to Adaptive Statistical Language Modeling,” Computer, Speech and Language, vol. 10, pp. 187–228, 1996.

[2] C. Tillmann, H. Ney, “Selection Criteria for Word Trigger Pairs in Language Modeling,” International Colloquium on Grammatical Inference, pp. 95-106, 1996.

[3] R. Khun, R. De Mori, “A Cache-Based Natural Language Model for Speech Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 6, pp. 570–583, 1990.

[4] G. Salton, “Developments in Automatic Text Retrieval,” Science, vol. 253, pp. 974–980, 1991.

[5] C. Troncoso, T. Kawahara, H. Yamamoto, G. Kikui, “Trigger-Based Language Model Construction by Combining Different Corpora,” Technical Report of IEICE, vol. 104, no. 542, pp. 25–30, 2004.

[6] Y. Akita, T. Kawahara, “Language Model Adaptation Based on PLSA of Topics and Speakers,” Proceedings ICSLP, pp. 1045–1048, 2004.

[7] K. Maekawa, H. Koiso, S. Furui, H. Isahara, “Spontaneous Speech Corpus of Japanese,” Proceedings LREC, vol. 2, pp. 947–952, 2000.