社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－SLP－54 (20)
2004／12／21

異種コーパスの組合せによるトリガー言語モデルの構築

カルロス・トロンコーソ† 河原 達也† 山本 博史‡ 菊井 玄一郎‡

†京都大学情報学研究科 〒606-8501 京都市左京区吉田二本松町
‡ATR 音声言語コミュニケーション研究所 〒619-0288 京都府相楽郡精華町光台 2-2-2
E-mail: †{carlos, kawahara}@ar.media.kyoto-u.ac.jp, ‡{hirofumi.yamamoto, genichiro.kikui}@atr.jp

あらまし 大語彙連続音声認識において，*n*-gram モデルより長距離の単語共起をモデル化するトリガー言語モデルについて検討する．一般に言語モデルの構築においては，タスクにマッチした学習コーパスのサイズは小さいため，統計量の学習が十分に行えず，逆に，大規模なコーパスでは一般的過ぎて，タスク依存性がなくなるという問題がある．本研究では，タスクにマッチしたコーパスからトリガーペアを抽出し，大規模なテキストコーパスからトリガーペアの生起確率を推定するアプローチを提案する．ATR の旅行会話コーパス(BTEC)，及び日本語話し言葉コーパス(CSJ)の模擬講演において評価を行った結果を報告する．

キーワード 言語モデル，音声認識，トリガー言語モデル，テキストコーパス

Trigger-Based Language Model Construction by Combining Different Corpora

Carlos TRONCOSO† Tatsuya KAWAHARA† Hirofumi YAMAMOTO‡ and Genichiro KIKUI‡

† School of Informatics, Kyoto University, Yoshida Nihonmatsu-cho, Sakyo-ku, Kyoto, 606-8501 Japan
‡ Spoken Language Translation Research Laboratories, ATR, 2-2-2 Hikaridai, Seika-cho, Kyoto, 619-0288 Japan
E-mail: †{carlos, kawahara}@ar.media.kyoto-u.ac.jp, ‡{hirofumi.yamamoto, genichiro.kikui}@atr.jp

Abstract We study the trigger-based language model (LM) for large vocabulary continuous speech recognition (LVCSR), which can model dependencies between words longer than those modeled by the *n*-gram LM. In general, in language modeling for LVCSR, when the training corpus matches the target task, its size is typically small, and therefore insufficient for providing us with reliable probability estimates. On the other hand, large corpora are often too general to capture task dependency. The proposed approach tries to overcome this generality-sparseness trade-off problem by constructing a trigger-based LM in which task-dependent trigger pairs are first extracted from the corpus that matches the task, and then the occurrence probabilities of the pairs are estimated from a huge text corpus. We report evaluation results in ATR's Basic Travel Expression Corpus (BTEC) as well as in the Corpus of Spontaneous Japanese (CSJ).

Keywords Language Model, Speech Recognition, Trigger-Based Language Model, Text Corpus

1. Introduction

The statistical language model (LM) is an integral part of the state-of-the-art automatic speech recognition (ASR) systems. The most widely used LM in LVCSR is the *n*-gram model, where *n* typically ranges from 2 (bigram) to 4 (4-gram). The *n*-gram LM models the occurrence probability of *n* consecutive words in the text, and its parameters are usually estimated from a large text corpus. This model is known to be effective, but it is apparently limited in scope, because it is unable to model dependencies longer than *n*.

Some works in the literature, such as the trigger-based LM [1] and the cache-based LM [2], tried to broaden the scope of the *n*-gram by modeling long-distance dependencies between words. The trigger-based LM uses a set of co-related word pairs, known as trigger pairs, to raise the probability of the words related through a given pair. When training this model, however, we usually find a fundamental problem, depending on the nature of the training data. When the trigger pairs are trained from a large corpus, many of the pairs are not task-dependent, because the corpus is usually too general. On the other hand, when the training data set is from the same domain as the target task, its size is usually

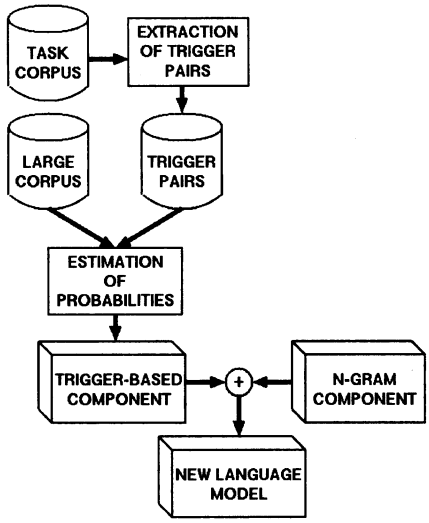

Figure 1: *Outline of the proposed approach.*

insufficient and the probability estimates are unreliable.

To overcome this trade-off between generality and sparseness, we propose an approach that takes advantage of two different corpora to create a trigger-based LM so that the trigger pairs are dependent on the target task and have reliable estimates.

The organization of this paper is as follows. Section 2 introduces the proposed approach in detail. Its application to the Basic Travel Expression Corpus (BTEC) and the Corpus of Spontaneous Japanese (CSJ) are shown in sections 3 and 4, respectively. Finally, the conclusions and some discussion are given in section 5.

2. Proposed Approach

In the proposed scheme, the trigger pairs are first extracted from a text corpus that matches the target task, and then searched for in a large text corpus in order to estimate their probabilities based on their co-occurrence frequency within a text window. Figure 1 illustrates the outline of the proposed approach.

By extracting the trigger pairs from the target domain, we solve the generality problem, while we avoid the data sparseness problem by calculating the probabilities of the trigger pairs are calculated with a huge corpus.

2.1. Extraction of trigger pairs from task corpus

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 document, it is likely that B come up afterwards.

The trigger pairs are first extracted from the corpus that matches the target domain. For the selection of pairs, we adopt two different criteria: the term frequency/inverse document frequency (TF/IDF) measure [3] and the log likelihood ratio [4]. We used the former for preliminary experiments because of its simplicity, while the latter, although more computationally demanding, was used for its powerfulness.

2.1.1. Extraction based on TF/IDF measure

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

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

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.

For each document, we create all possible pairs, including pairs of the same words (self-triggers), with the base forms and parts of speech (POS) of all the words with a TF/IDF value above a threshold. We use POS-based filtering to discard function words, as well as a stop list to ignore very frequent words.

2.1.2. Extraction based on log likelihood ratio

Given a contingency table with the frequency of the following co-occurrence pairs:

$$a)\ A+B \qquad c)\ \neg A+B$$

$$b)\ A+\neg B \qquad d)\ \neg A+\neg B,$$

where $A+\neg B$ represents the two pairs $A \rightarrow \neg B$, $\neg B \rightarrow A$ formed by A and any word that is not B, the log likelihood ratio (LLR) of the pair $A \rightarrow B$ is calculated as follows:

$$
\begin{aligned}
-2\log\alpha = 2[&a\log a + b\log b + c\log c + d\log d \\
&-(a+b)\log(a+b)-(a+c)\log(a+c) \\
&-(b+d)\log(b+d)-(c+d)\log(c+d) \\
&+(a+b+c+d)\log(a+b+c+d)]
\end{aligned} \tag{2}
$$

For each document, we first create all possible pairs with the base forms and POS of all the words in it, including self-triggers. Again, POS-based filtering and a stop list are used to remove function words and high frequency words, respectively. Then, we compute the LLR for each pair and choose the trigger pairs with a ratio greater than a threshold.

2.2. Probability estimation from large corpus

The probabilities of the trigger pairs are estimated from a huge corpus by using a text window to calculate the co-occurrence frequency of the pairs inside it. This text window consists of the 20 words previous to the one being processed, excluding the previous 2 words, which are covered by the trigram (3-gram) LM.

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

$$
P_{TP}(w_2 \mid w_1) = \frac{N(w_1, w_2)}{\sum_j N(w_1, w_j)}, \tag{3}
$$

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 through all words triggered by w_1. When the denominator is less than 40, the corresponding pair is discarded by setting its probability to 0.

2.3. Proposed trigger-based language model

The probabilities of the trigger pairs are then linearly interpolated with the baseline trigram model, so that both long and short-distance dependencies can be captured.

Thus, the score of the new LM for a word w_i given the word history H is computed in the following way:

$$
P_{LM}(w_i \mid H) = \frac{1}{|H|}\sum_{h \in H} P_{LM}(w_i \mid h), \tag{4}
$$

where $|H|$ means the number of words in the history, and $P_{LM}(w_i \mid h)$ is calculated as follows:

$$
\begin{cases}
P_{NG}(w_i \mid H), & \text{if } P_{TP}(w_j \mid h) = 0, \forall j \\
\lambda P_{NG}(w_i \mid H) + (1-\lambda) P_{TP}(w_i \mid h), & \text{otherwise}
\end{cases} \tag{5}
$$

Here λ is the interpolation weight, P_{NG} is the probability of the n-gram component, and P_{TP} is the probability of the trigger-based component, given by equation 3.

2.4. *N*-best rescoring

The new LM is used to rescore the N-best hypotheses output by a baseline ASR system. The system provides us with acoustic and LM scores for each of the words in every hypothesis.

Words in each hypothesis are added in order to a word history buffer, which is cleared when the hypothesis processing is over. The LM score for each hypothesis is updated by using this buffer and the previous equations. The hypothesis with the highest new total score is regarded as the new 1-best sentence.

The number of trigger pairs used during the rescoring process is limited to be those with a probability above a threshold.

3. Application to BTEC

3.1. Corpus and procedure

The Basic Travel Expression Corpus (BTEC) [5] is a conversational text corpus consisting of sentences from many different topics that usually appear in travel conversations. It is divided in two disjoint sets: training and evaluation. The former contains 467,964 utterances and 3.5 million Japanese morphemes (hereafter words), and the latter comprises 24,682 utterances and 184 thousand words.

The trigger pairs were extracted from the Japanese version of the BTEC. We had to use the utterance as the document unit, since utterances in this corpus are not related to each other.

The threshold for the TF/IDF value was chosen to be 0.2 so that the coverage in the evaluation corpus of the trigger pairs created with only the POS-based filtering were 20%, and was empirically tuned later to produce a threshold of 0.15.

The threshold for the LLR was initially chosen to be 10 based on a subjective judgment of the goodness of the pairs from a sample taken at random, and it was later tuned empirically, producing the

value 2. The coverage in the evaluation corpus of the trigger pairs created by using only the POS-based filtering was 19%.

The probabilities of the trigger pairs were estimated from two different corpora: the Mainichi Shimbun newspaper corpus and a conversational text corpus extracted from the World Wide Web (WWW) [6] (hereafter web corpus). We used 5 years (1991-1995) of articles from the Mainichi Shimbun corpus, consisting of 130 million words. The web corpus consists of conversational texts that can be found in the WWW, such as chat logs, and comprises 270 million words, of which we used 122 million words.

3.2. Experimental setup

The ASR system ATRIUMS 2.2 [7] was used to output the N-best lists. The size of the vocabulary was 36K words. This system normally uses a bigram model in a first stage and a trigram afterwards, in an optional rescoring stage. The BTEC bigram was used in the first recognition stage, and a linear interpolation between the BTEC and Mainichi trigrams, with interpolation weights of 0.99 and 0.01, respectively, was used for the second stage. The test set consisted of 1524 utterances (11K words) taken from the BTEC evaluation corpus (sets 1, 2 and 3) and the number of output hypotheses N was 100.

We obtained an average word recognition accuracy of 87.64% with this baseline LM, and the maximum average recognition accuracy that could be attained by choosing the best hypothesis from the N-best each time was 94.53%.

3.3. Perplexity evaluation

We evaluated the perplexity of the proposed LM for different values of the coverage of the trigger pairs in the test set, determined by the threshold for the frequency of the words in the stop list. The values for this threshold were 500, 1000, 2000, 3000, and 5000. We compared the perplexity of the model constructed from both the BTEC and the web corpus, the model built from the BTEC and the Mainichi Shimbun corpus, and the one that used only the BTEC, both to extract the trigger pairs and to calculate their probabilities. We compared these three models for each of the two criteria used for the extraction of the trigger pairs. For the TF/IDF measure, the number of extracted trigger pairs varied from 447,060 to 1,052,342 for the first model, from 418,629 to 976,656 for the second one, and from 325,253 to 880,957 for the third one. For the LLR, the number of extracted trigger pairs varied from 412,678 to 821,093 for the first model, from 388,912 to 767,157 for the second one, and from 300,849 to 668,878 for the third one.

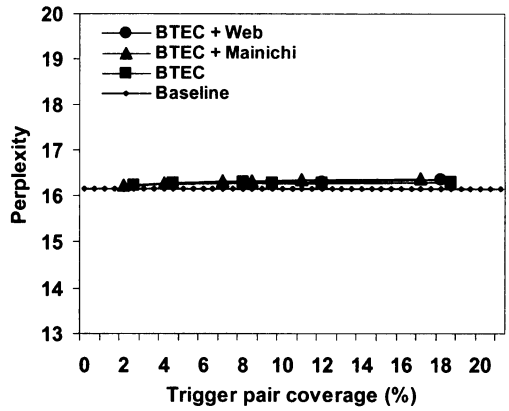

Figure 2: *Perplexity against coverage of trigger pairs for different sets of trigger pairs extracted from the BTEC with the LLR.*

The two criteria gave similar results, and in Figure 2 we show the results when we used the LLR criterion. We can see that the perplexity did not change significantly in any of the cases. One of the possible reasons for this is that, since the utterances of BTEC are unrelated to each other, we could not use the information of the previous sentences for our trigger-based LM. Furthermore, most utterances in BTEC are short, so it is difficult to extract good trigger pairs from them.

3.4. Rescoring experiments

We then carried out rescoring experiments with the output of the baseline system. We compared the word recognition accuracy of the models constructed from the BTEC and the web corpus, the BTEC and the Mainichi Shimbun corpus, and only the BTEC, for each of the two extraction criteria.

Figure 3 shows these results. The recognition accuracy is plotted against the coverage in the test set of the trigger pairs. The best accuracy obtained was 87.71%, that is, we achieved a global 0.07% improvement when we used trigger pairs based on the LLR, a stop list threshold of 5000, and the probabilities were computed from the web corpus.

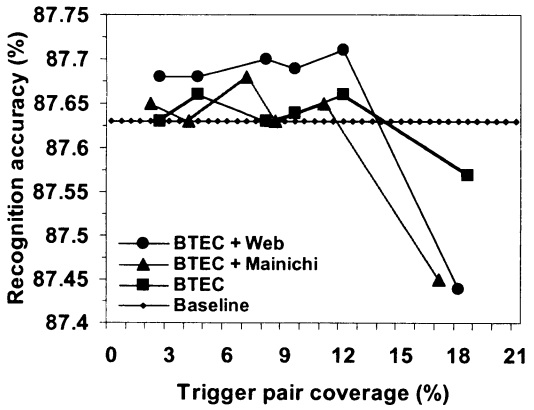

Figure 3: *Word recognition accuracy against coverage of trigger pairs for different sets of trigger pairs extracted from the BTEC with the LLR.*

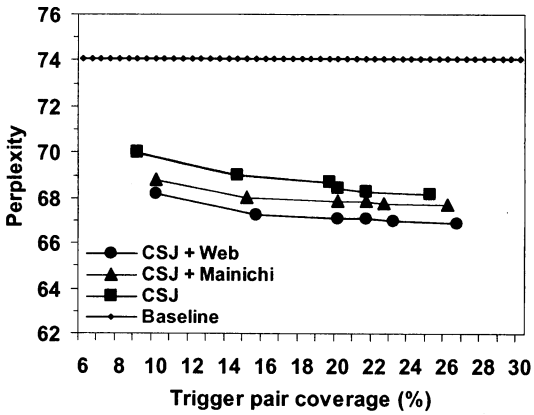

Figure 4: *Perplexity against coverage of trigger pairs for different sets of trigger pairs extracted from the CSJ with the TF/IDF measure.*

4. Application to CSJ

4.1. Corpus and procedure

The Corpus of Spontaneous Japanese (CSJ) [8] is a conversational corpus consisting of lectures on various academic subjects and extemporaneous speeches on different topics. We used the extemporaneous speeches, which are divided into 1705 speeches of training data, comprising 3.5 million words, and 10 speeches of evaluation data, containing 16 thousand words.

The trigger pairs were extracted from the training data. In this case, we used only the TF/IDF measure, because the LLR approach was computationally expensive, and in the BTEC experiments both criteria gave similar results. We used the lecture as the document unit. The threshold for the TF/IDF value was initially chosen to be 0.015 based on a subjective judgment of the goodness of the pairs from a sample taken at random, and it was later tuned empirically, producing the value 0.031.

For estimating the probabilities, we used two different corpora: 11 years (1991-2001, 289 million words) of the Mainichi Shimbun corpus and the whole web corpus (270 million words).

4.2. Experimental setup

For the CSJ experiments, we used the ASR system Julius 3.4.2 [9]. The size of the vocabulary was 30K words. We created a word bigram and a word trigram from the CSJ training corpus, and we used the CSJ test set for the experiments. The number of output hypotheses N was also 100 here. The average word recognition accuracy was 66.76% with this baseline LM.

4.3. Perplexity evaluation

We evaluated the perplexity of the proposed LM for different values of the coverage of the trigger pairs in the test set. We compared three different models: the model that was constructed by using the CSJ and the web corpus, the model constructed with the CSJ and the Mainichi Shimbun corpus, and the model that used only the CSJ corpus, both to extract the trigger pairs and to calculate their probabilities. The number of extracted trigger pairs varied from 11,202,729 to 11,771,760 for the first of the mentioned models, from 10,841,342 to 11,539,218 for the second one, and from 3,446,299 to 5,435,233 for the third one.

The results are illustrated in Figure 4. The highest relative perplexity reduction was 9.7%. We can see that the model that used the CSJ and the web corpus achieved a lower perplexity than both the model that used the CSJ and the Mainichi Shimbun corpus and the model that used only the CSJ. The style of the web corpus is more similar to the target task than the style of the Mainichi Shimbun corpus, and that is why it performed better in terms of perplexity evaluation. Furthermore, the models that used two corpora performed better than the model that used only one.

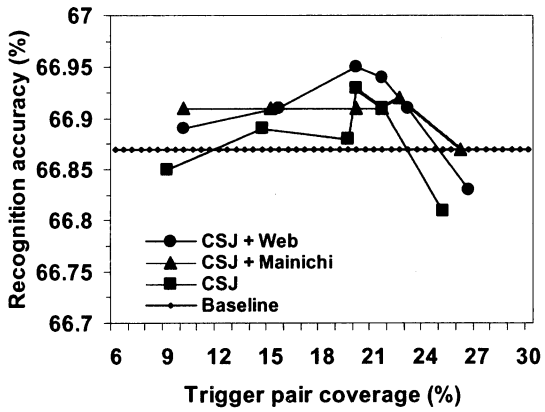

Figure 5: *Word recognition accuracy against coverage of trigger pairs for different sets of trigger pairs extracted from the CSJ with the TF/IDF measure.*

4.4. Rescoring experiments

Next, we performed rescoring experiments with the output of the baseline system. We compared the word recognition accuracy of the models constructed from the CSJ and the web corpus, the CSJ and the Mainichi Shimbun corpus, and only the CSJ. The results are shown in Figure 5. As can be seen, the model that used both the CSJ and the web corpus achieved the highest accuracy. The models that used two corpora performed on average better than the model that used only the CSJ.

5. Conclusion and Discussion

We presented a novel approach to the trigger-based LM based on two different corpora. A significant improvement in perplexity was achieved when using the CSJ and the web corpus, as compared with the baseline trigram and the model that uses only the CSJ.

There can be several possible reasons for the small degree of improvement obtained in the rescoring experiments. The most likely cause is the fact that we used the simple linear interpolation scheme to combine the standard *n*-gram LM with our trigger-based LM. Linear interpolated models make suboptimal use of their components and are generally inconsistent with them [1]. It was used as a quick means to test the proposed approach in this work. Using a more robust method, such as the maximum entropy approach [1], might bring more improvement.

Acknowledgements

We would like to thank Mr. Nobuhiro Kaji and Professor Sadao Kurohashi of University of Tokyo for providing us with the conversational web corpus, and also Dr. Shinsuke Mori of IBM Japan, who gave us very useful comments.

References

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

[2] 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, 1990.

[3] G. Salton, "Developments in Automatic Text Retrieval," Science, vol. 253, pp. 974–980, 1991.

[4] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, vol. 19, no. 1, pp. 61–74, 1993.

[5] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, S. Yamamoto, "Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World," Proceedings LREC, vol. 1, pp. 147–152, 2002.

[6] N. Kaji, M. Okamoto, S. Kurohashi, "Paraphrasing Predicates from Written Language to Spoken Language Using the Web," Proceedings HLT-NAACL, 2004.

[7] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, Y. Sagisaka, "Spontaneous Dialogue Speech Recognition Using Cross-Word Context Constrained Word Graph," Proceedings ICASSP, vol. 1, pp. 145–148, 1996.

[8] K. Maekawa, H. Koiso, S. Furui, H. Isahara, "Spontaneous speech corpus of Japanese," Proceedings LREC, vol. 2, pp. 947–952, 2000.

[9] A. Lee, T. Kawahara, K. Shikano, "Julius – an Open Source Real-Time Large Vocabulary Recognition Engine," Proceedings Eurospeech, vol.3, pp. 1691–1694, 2001.