Internal Use Only（非公開）

TR－SLT－0074

Trigger-Based Language Model Adaptation Using Two Different Corpora

異種コーパスを用いたトリガー言語モデルの適応

Carlos Troncoso Alarcon, Hirofumi Yamamoto, Genichiro Kikui
カルロス　トロンコッソ　アラルコン、山本　博史、菊井玄一郎

2004年5月24日

概要

We present a novel approach to trigger-based language model adaptation for large vocabulary continuous speech recognition (LVCSR) that uses two different corpora to construct the set of trigger pairs. In language modeling for LVCSR, when the training data set is considerably big, it is usually too general and the task dependency is lost. On the other hand, when the training data are task-dependent, they are usually insufficient and the probability estimates are unreliable. The proposed approach tries to overcome this generality-sparseness trade-off problem by first building task-dependent trigger pairs from a Japanese conversational text corpus, which is the target task, and then avoiding data sparseness by calculating the likelihoods of the pairs from a huge text corpus. A small improvement in word recognition accuracy was achieved when using the two corpora, while accuracy degradation was obtained when we used either only the conversational text corpus or the huge corpus to both extract the pairs and calculate their likelihoods.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　　:+81-774-95-1308

Trigger-Based Language Model Adaptation Using Two Different Corpora

Carlos Troncoso Alarcón, Hirofumi Yamamoto, Genichiro Kikui

May 24, 2004

Abstract

We present a novel approach to trigger-based language model adaptation for large vocabulary continuous speech recognition (LVCSR) that uses two different corpora to construct the set of trigger pairs. In language modeling for LVCSR, when the training data set is considerably big, it is usually too general and the task dependency is lost. On the other hand, when the training data are task-dependent, they are usually insufficient and the probability estimates are unreliable. The proposed approach tries to overcome this generality-sparseness trade-off problem by first building task-dependent trigger pairs from a Japanese conversational text corpus, which is the target task, and then avoiding data sparseness by calculating the likelihoods of the pairs from a huge text corpus. A small improvement in word recognition accuracy was achieved when using the two corpora, while accuracy degradation was obtained when we used either only the conversational text corpus or the huge corpus to both extract the pairs and calculate their likelihoods.

1 Introduction

Statistical language models are an integral part of state-of-the-art automatic speech recognition (ASR) systems. The most widely used language model in LVCSR is the n-gram language model, where n typically ranges from 2 (bigram) to 5 (5-gram). n-grams model the occurrence probability of n consecutive words in the text, and their parameters are estimated from a large text corpus. These models have fixed probabilities that are independent of the document being predicted, and they are usually very general in order to cover many different topics or domains.

Adaptation tries to improve language modeling by creating language models closer in style to the target task. In the n-gram language model, when the model is trained from a big corpus, we can obtain a good general n-gram model, but a rather poor task-dependent model. Conversely, when the training data and the target task are from the same domain, the data are usually sufficient for building a task-dependent n-gram, but not for building a general one.

Some works in the literature, such as the trigger-based language model [1], tried to broaden the scope of the n-gram by modeling long-range dependencies between words. In this model, however, when we train the trigger pairs (content word pairs that are related to each other) we usually find two problems, depending on the nature of the training data. When the trigger pairs are trained from a considerably big corpus, even though we can find good probability estimates for them, since the corpus is usually too general we cannot know which of the pairs are task-dependent. On the other hand, when the training data are from the same domain as the target task, although the trigger pairs are task-dependent, the data are usually insufficient and the probability estimates are unreliable. Therefore, there is often a trade-off between generality and sparseness.

The proposed approach takes advantage of two different corpora to create a trigger-based language model whose trigger pairs are adapted to the target task and have reliable estimates.

2 Proposed approach

In order to overcome the generality-sparseness trade-off problem, the trigger pairs were first extracted from a conversational text corpus, which is the target task, and then searched for in a large text corpus, to compute their likelihoods based on their co-occurrence frequency within a text window. Since the trigger pairs are task-dependent, because they were built from the target domain, we solve the generality problem. Furthermore, since the likelihoods of the trigger pairs were calculated from a huge corpus, we avoid the data sparseness problem. By overcoming the generality-sparseness trade-off problem, a significant improvement in speech recognition accuracy is pursued when using the new language model for rescoring N-best lists.

2.1 Extraction of trigger pairs

The trigger pairs were extracted from the Japanese Basic Travel Expression Corpus (BTEC) [2]. The BTEC 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 words, and the latter comprises 24,682 utterances and 184 thousand words.

The BTEC was segmented using the ChaSen morphological analysis system 2.02 [3] to match the segmentation of the Mainichi Shimbun corpus.

The trigger pairs were extracted by using two different methods: a method based on the term frequency/inverse document frequency (TF/IDF) measure [4] and another based on log likelihood ratios [5]. We used the former for preliminary experimentation because of its simplicity, while the latter was used due to its powerfulness.

2.1.1 TF/IDF

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

$$v_{ik} = \frac{tf_{ik}\log(N/n_k)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2[\log(N/n_k)]^2}} \tag{1}$$

where tf_{ik} is the frequency of occurrence of T_k in D_i, N is the total number of documents, n_k is the number of documents that contain T_k, and t is the number of terms in D_i.

For each utterance of the corpus, we took the base forms and parts of speech (POS) of the two words w_1 and w_2 with the highest TF/IDF value above a threshold, using the utterance as the document unit. We used POS-based filtering to discard function words, as well as a stop list to ignore high frequency words. Then, for every w_1 and w_2, we constructed the pairs (w_1, w_2) and (w_2, w_1). The threshold was chosen to be 0.2 so that the coverage in the BTEC evaluation corpus of the trigger pairs created with only the POS-based filtering were 62%.

2.1.2 Log likelihood ratios

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

$$\begin{array}{ll} a) & A+B \\ b) & A+\neg B \\ c) & \neg A+B \\ d) & \neg A+\neg B \end{array}$$

where $A+\neg B$ represents the two pairs $(A,\neg B),(\neg B, A)$ formed by A and any word that is not B, the log likelihood ratio (LLR) of the pair (A, 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 utterance of the corpus, we first created all possible pairs from all words in the utterance. Again, POS-based filtering and a stop list were used to remove function words and high frequency words, respectively. Then, we computed the LLR for each pair and chose the trigger pairs with a ratio greater than a threshold. This threshold 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 during the parameter optimization stage. The coverage in the evaluation corpus of the trigger pairs created by using only the POS-based filtering was 58%.

2.2 Calculation of likelihoods

The likelihoods of the trigger pairs were computed from the Mainichi Shimbun text corpus. This corpus consists of five years (1991-1995) of this general Japanese newspaper, and it comprises 130 million words.

We created ANS files from this corpus in order to have a common representation for both the BTEC and the Mainichi Shimbun corpus.

In order to compute the likelihoods of the pairs, we used a text window to calculate the co-occurrence frequency of the pairs inside it. This text window consisted of the previous and next 10 words to the one being processed, excluding the previous and next 2 words because we are trying to model long-distance dependencies that are not modeled by the trigram (3-gram) language model.

The likelihood of each trigger pair (w_1, w_2) was computed as follows:

$$L_{TP}(w_2|w_1) = \frac{N(w_1, w_2)}{N(w_1) \times K} \tag{3}$$

where $N(w_1, w_2)$ denotes the number of times the words w_1 and w_2 co-occur within the text window, $N(w_1)$ is the number of times w_1 occurs in the corpus, and K is the window size, in this case 16. When $N(w_1)$ was less than 10, the corresponding likelihood was set to 0.

2.3 Language model

The likelihoods of the trigger pairs were interpolated with the baseline n-gram model, so that both long and short-range dependencies could be captured, to create a new language model. We tried three different interpolation schemes: word linear interpolation (WLI), sentence linear interpolation (SLI), and log linear interpolation (LLI).

In the WLI scheme, the total score of the new language model for a sentence $W = w_1, w_2, ..., w_m$ was computed in the following way:

$$S_{LM}(W) = 10 \log_{1.0001}\Big(\prod_{i=1}^{m}(\lambda S_{NG}(w_i) + (1-\lambda)S_{TP}(w_i))\Big) \tag{4}$$

where 10 is the language model scaling factor, λ is the interpolation weight, S_{NG} is the score of the n-gram component, and S_{TP} is the score of the trigger-based component, which can be calculated as follows:

$$S_{TP}(w_i|H) = \sum_{h \in H} L_{TP}(w_i|h) \tag{5}$$

where H is the word history and L_{TP} is the likelihood of each trigger pair, defined in equation 3.

In SLI, the total language model score for a sentence W was computed as follows:

$$S_{LM}(W) = 10 \log_{1.0001}\Big(\lambda \prod_{i=1}^{m} S_{NG}(w_i) + (1-\lambda) \prod_{i=1}^{m} S_{TP}(w_i)\Big) \tag{6}$$

where S_{TP} was calculated as:

$$S_{TP}(w_i|H) = \begin{cases} \epsilon & L_{TP}(w_i|h) = 0, \forall h \in H \\ \sum_{h \in H} L_{TP}(w_i|h) & \text{otherwise} \end{cases} \tag{7}$$

and ϵ was a value close to 0 and less than the smallest likelihood found in the set of trigger pairs.

Finally, in LLI, the total language model score for W was calculated in the following way:

$$S_{LM}(W) = 10 \log_{1.0001}\Big(\prod_{i=1}^{m} S_{NG}(w_i)\Big) + \lambda \log_{1.0001}\Big(\prod_{i=1}^{m} S_{TP}(w_i)\Big) \tag{8}$$

where S_{TP} was computed using equation 7.

2.4 N-best rescoring

The new language model was used to rescore the N-best hypotheses output by an ASR system. This system provided us with acoustic and n-gram scores for each of the words in every hypothesis, as well as total scores for the each hypothesis.

Words in each hypothesis were added in order to a word history buffer, which was cleared when the hypothesis processing was over. The score of the trigger-based component was calculated by using the history buffer to find trigger pairs containing the word being processed, taking their previously computed likelihoods, and using equation 5 for the WLI scheme and equation 7 for the SLI and LLI schemes. The new language model score for a hypothesis was computed by using equations 4, 6, and 8 for the WLI, SLI and LLI interpolation schemes, respectively. Finally, the total score for one hypothesis was the sum of the log acoustic score of the hypothesis and its new language model score. The hypothesis with the highest new total score became the candidate for the new 1-best sentence.

For each utterance, both the original and the candidate 1-best sentences were considered. If the number of trigger pairs found in the original 1-best hypothesis was greater or equal to the number of pairs found in the candidate 1-best, the original hypothesis prevailed; otherwise, the candidate sentence became the new 1-best. This makes sense because if the same trigger pairs are present in both the original and candidate hypotheses, there are no errors that can be corrected with the information provided by these pairs, so there is no need for changing the original hypothesis.

3 Experiments

3.1 Baseline experiments

The ASR system ATRIUMS 2.2 [6] was used to output the N-best lists. This system normally uses a bigram model in a first stage and a trigram afterwards, in an optional rescoring stage. We created a word bigram and a word trigram from the BTEC training corpus, as well as a word trigram from the Mainichi Shimbun corpus. 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), the number of output hypotheses N was 100, and the size of the vocabulary, extracted from the BTEC, was 36K words.

We obtained an average word recognition accuracy of 87.25% for this baseline language model, 87.64% when using the insertion penalty optimized in the parameter optimization stage, and the maximum average recognition accuracy that could be attained by choosing the best hypothesis from the N-best each time was 94.53%.

3.2 Parameter optimization

Next, we performed speech recognition experiments similar to the ones in the previous section, this time with a held-out data set from the BTEC evaluation corpus (set 5), consisting of 466 utterances. This provided us with N-best lists that were rescored with the proposed language model, assigning different values to the system parameters in order to optimize them. In particular, the interpolation weight λ, the insertion penalty, and the threshold for the LLRs of the trigger pairs were tuned. The highest recognition accuracy was achieved when λ was 0.37 and the threshold for the LLRs was 15.

3.3 Experimental results

We then carried out rescoring experiments with the output of the baseline experiments of section 3.1 and with the parameters optimized in the previous section. We compared the word recognition accuracy of the model with the trigger pairs constructed using the method based on the TF/IDF measure with that of the model with the trigger pairs that used the LLRs. For each of these models, we also compared different thresholds for the frequency of the words in the stop list: 500, 1000, 2000, 3000 and 5000. The number of extracted pairs ranged from 104,134 to 195,265 for the method based on the TF/IDF measure, and from 143,928 to 219,001 for the method based on the LLRs.

Finally, in order to prove the usefulness of using two different corpora, all these experiments were repeated using either only the BTEC or the Mainichi Shimbun corpus, both to extract the trigger pairs and to calculate their likelihoods. When we used only the BTEC, the number of trigger pairs ranged from 92,479 to 193,969 for the TF/IDF-based pairs, and from 124,042 to 205,806 for the LLR-based ones. When we used only the Mainichi Shimbun corpus, three different thresholds for the stop list were used: 10000, 30000 and 50000. This time, the number of extracted trigger pairs ranged from 4,560,156 to 5,131,819 for the TF/IDF-based pairs, and from 23,747,412 to 29,655,833 for the LLR-based pairs.

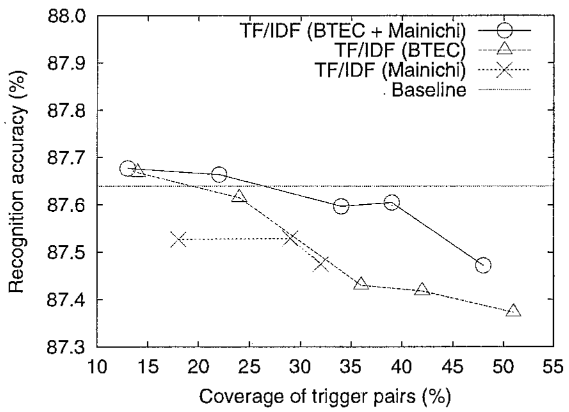

Figure 1: *Speech recognition accuracy for different sets of trigger pairs based on the TF/IDF measure.*

We performed all these experiments using the three different interpolation schemes described in section 2.3. We found that for the SLI and LLI schemes, the speech recognition accuracy of the models that used the two corpora was always very similar to that of the models that used the BTEC only. However, for the WLI scheme, the recognition accuracy of the models based on the two corpora always outperformed the accuracy of the models constructed with only the BTEC.

Figures 1 and 2 show the experiments that used the WLI interpolation scheme. The speech recognition accuracy is compared with the coverage in the evaluation corpus of all the different sets of trigger pairs. The horizontal line is the value of the average recognition accuracy of the baseline language model.

The maximum recognition accuracy obtained was 87.71%, that is, we achieved a global 0.07% improvement when we used trigger pairs based on LLRs and a stop list threshold of 1000, and the likelihoods were computed from the Mainichi Shimbun corpus. On the contrary, using the BTEC to calculate the likelihoods resulted in accuracy degradation of 0.03% for the same case. As a matter of fact, we can see that in the cases where we used the two corpora, the recognition accuracy was always higher than in the cases where we used only one corpus, where the accuracy improved that of the baseline in just one case. We can also notice that the LLR-based trigger pairs performed generally better than the TF/IDF-based ones.

In order to further evaluate the impact of the proposed model in the BTEC task, we discarded from the test set those sentences that did not contain any of our trigger pairs. The baseline recognition accuracy in this case was 88.02%, and the maximum attainable accuracy was 95.13%. We then repeated the previous experiments only for the case where the threshold of the stop list was 1000. We obtained a word recognition accuracy of 89.06% for the TF/IDF-based trigger

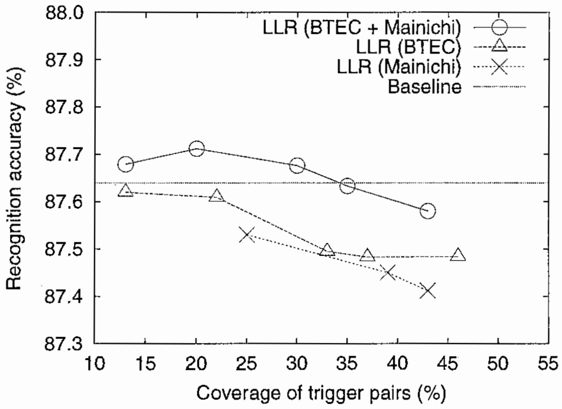

Figure 2: *Speech recognition accuracy for different sets of trigger pairs based on log likelihood ratios.*

pairs, and an accuracy of 88.98% for the LLR-based pairs. For the first case, the global improvement over the baseline is of 0.31%, which represents a 4.36% of the total possible improvement.

4 Discussion

There can be several reasons for the small degree of improvement obtained. The most likely cause is that the target task (BTEC) belongs to the conversational language domain, while the Mainichi Shimbun corpus, from where the likelihoods were computed, belongs to the written language domain instead. This domain mismatch might be the main problem, since the Mainichi Shimbun corpus is probably unable to provide us with good estimates for the likelihoods of trigger pairs adapted to a conversational domain. We decided to use the Mainichi Shimbun corpus because large Japanese conversational corpora are not readily available, so we wanted to try it for preliminary experimentation. It would be possible to build a huge conversational corpus by gathering text data from the World Wide Web similar to the target domain in order to overcome this problem.

Another possible factor that contributed to the small improvement is the fact that we used the simple linear interpolation scheme to combine the standard n-gram language model with our trigger-based language model. Linear interpolated models make suboptimal use of their components and are generally inconsistent with them [1]. Again, we decided to use linear interpolation as a quick means to test the proposed approach. We think that using a more robust method, such as the maximum entropy approach [1], would also contribute to improve recognition accuracy.

Finally, in order to further increase accuracy, based on the hypothesis that long words usually have a higher acoustic confidence than short words, we propose to use acoustic confidence scores derived from the generalized word posterior probability [7] as an additional parameter when calculating the score during the N-best rescoring.

References

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

[2] 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.

[3] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara, "Morphological Analysis System ChaSen version 2.2.1 Manual," http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf, 2000.

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

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

[6] 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.

[7] F. Soong, W. Lo, S. Nakamura, "Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words," Proceedings SWIM, 2004.