Language Model Adaptation with Additional Text Generated by Machine Translation[†]

Hideharu NAKAJIMA
NTT Cyber Space Laboratories, NTT Corporation
420C 1-1 Hikarino-oka Yokosuka-shi, Kanagawa, 239-0847, JAPAN
nakajima.hideharu@lab.ntt.co.jp

Hirofumi YAMAMOTO and **Taro WATANABE**
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun, Kyoto, 619-0288, JAPAN
{hirofumi.yamamoto,taro.watanabe}@atr.co.jp

Abstract

Statistical language modeling requires a large corpus for the application domain. When a large corpus is not available, the language model adaptation technique has often been used in the speech recognition research domain. This adaptation needs only a small corpus of the application domain (the "target corpus") and the corpus should be written in the language of the model. However, it is sometimes difficult to collect even a small corpus, especially of spoken language, due to its high cost. To address this problem, this paper proposes a novel scheme that generates a small target corpus in the language of the model by machine translation of the target corpus in another language. As information about adjacent words, which is necessary for a statistical language model, is stored in the translation knowledge, it can be extracted by machine translation and used for adaptation. Experiments showed that the language model improvement was about half of that which was obtained with a human collected corpus, and this provided some initial proof of the concept experiments.

1 Introduction

Statistical language modeling requires a large corpus for the application domain. In case a large corpus is not available, the language model adaptation technique is often used. In this adaptation, first, a model is estimated with a large corpus or a corpora mixture which is not specific to the application domain (i.e., a "general" corpus). Then, the model is "adapted" to work well in the application domain (i.e., the "(adaptation) target task") with a small corpus of the domain. In multi-lingual speech-to-speech translators, each language model needs a small adaptation target task corpus in each language. However, it is difficult to build even a small corpus, much less multi-lingual corpora due to high costs. This paper tries to address this problem. That is, we propose a novel scheme that generates a small corpus in one language by using machine translation of a target task corpus in another language for model adaptation. This paper also shows that the adaptation with the generated corpus significantly improves models, in terms of test set perplexities.

Lexica, N-grams, and examples are used as the knowledge for machine translation. They can be expected to contain information about wordings in adjacent word contexts. If the translation result contains unnatural expressions or errors, the quality of the whole sentence may be spoiled. However, many local contexts, such as adjacent words, can be expected to maintain appropriate word orders. The proposed scheme translates the corpus of an adaptation target task. Therefore, the translation results can be expected to maintain the topic and sentence style of the target task corpus, and to be useful for adaptation.

Section 2 explains the situation and the proposed scheme of language model adaptation, and an overview of the machine translator used in the evaluation is described in section 3. Section 4 describes experimental conditions and

[†] This research was carried out at ATR. This research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

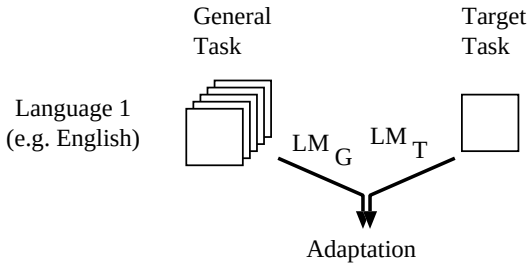

Figure 1: Problem of language model adaptation

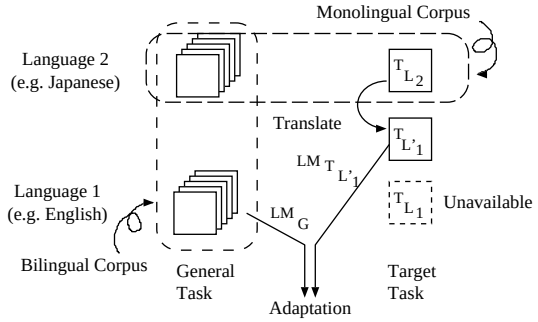

Figure 2: Language model adaptation with additional text generated by machine translation

results. The relation between this paper and other, conventional research is described in section 5.

2 Language Model Adaptation

First, this section reviews the problem of language model adaptation. As shown in **Figure 1**, language model adaptation consists of collecting a small corpus of the adaptation target ("*Target Task*" in Figure 1), using it with a large corpus of a general task (or task independent corpus; "*General Task*" in Figure 1), and making another new language model that works well for the target task.

On the other hand, the language model adaptation treated by this research is as shown in **Figure 2**. That is, in the language model adaptation of "*Language* 1" in Figure 2, when there is no small corpus of the adaptation target task, such as T_{L_1}, a quasi-corpus of T_{L_1} ($T_{L'_1}$) is generated by machine translation of a small corpus T_{L_2} of the target task in another language "*Language* 2." Then, this quasi-corpus is used with a large general corpus (corpus on the side of *Language* 1 in *General Task* in Figure 2) to make a task adapted language model.

Estimation methods include the Maximum a posteriori (MAP) adaptation approach (Masataki et al., 1997) and the linear combination of models (Rudnicky, 1995). The experiments in this paper use the linear combination.

The procedure of language model adaptation in this research is summarized as follows.

Step 1 Prepare a machine translator.

Step 2 Prepare a language model with only a general task (LM_G in Figure 2).

Step 3 Generate a quasi corpus of the target task ($T_{L'_1}$ in Figure 2) with the above machine translator.

Step 4 Make a language model ($LM_{T_{L'_1}}$ in Figure 2) with only the corpus generated in Step 3 above.

Step 5 Combine the language model (LM_G) prepared in Step 2 and the language model ($LM_{T_{L'_1}}$) obtained in Step 4 linearly to create a language model adapted to the new target task.

3 Machine Translator

The proposed scheme can use several kinds of machine translator. For the experiment in this paper, we used a statistical machine translator because of the clearness of its translation principle.

The statistical model used in the machine translator in the evaluation is based on the IBM Model 4 (Brown et al., 1993). Because this model considers phrasal correspondence in a bilingual corpus, we can expect higher translation quality than that of the IBM Model 3, which takes into account only positional alignment. Also, because this model uses fewer parameters than the IBM Model 5, we can expect the parameter estimation to converge more quickly and accurate than that of the IBM Model 5, which is more complicated. Hence, the IBM Model 4 is suitable for experiments and was chosen for use here. GIZA++(Och and Ney, 2000a; Och and Ney, 2000b) is used for translation model parameter

estimation, and the similar beam searching to that used by Tillmann et al. (Tillmann and Ney, 2000) is used. All of the parameters are estimated with a bilingual corpus (e.g., "Bilingual Corpus" for "General Task" in Figure 2).

Here, we describe the overview of the statistical translator that is necessary for the following explanation.

In statistical translation, the problem of translation is considered as a problem of decoding a signal conveyed through a noisy channel. For example, considering translation from Japanese (J) to English (E), the most probable English sentence (E^*) as a translation result can be represented by

$E^* = argmax\ P(E|J)$.

Usually, this equation is transformed, by Bayes rule, to

$E^* = argmax\ \{P(J|E)P(E)/P(J)\}$,

and because the denominator is constant when a Japanese sentence is entered as an input, only the numerator decides the most probable translation result E^*. $P(J|E)$ is called a "translation model" and $P(E)$ is a "language model" for which the adjacent word N-gram is often used.

In the following experiments, this language model $P(E)$ is used as LM_G in Figure 2. It is also to be adapted to some new task ("Target Task" in Figure 2), and the performance of its adaptation is to be evaluated.

If models are estimated correctly, the sentences even in a new task can be translated relatively correctly under the constraints of the language model (as $P(E)$). Then, many local contexts can be expected to maintain comparatively correct word sequences. Therefore, these translations can be expected to be useful for language model adaptation.

4 Experiments

We evaluate our scheme with word perplexities, which indicate word predictability (hereafter denoted as PP), on an open test set.

4.1 General Task and Adaptation Target Task Data

We used a bilingual corpus consisting of paired sentences between Japanese and English. The sentences are related to traveling. About 160,000 sentences were used in the following experiments.

Table 1: Categories of sentences

basic	airport
airplane	returning to home countries
exchange	staying
restaurant	transportation
drinks	light meal
shopping	sightseeing
troubles	beauty treatment
information	business
communication	

Table 2: The size of the general, target and test corpora (Numbers are counted on the English side. Nb. of S denotes "number of sentences." Nb. of W denotes "number of words")

Corpus Name	Nb. of S	Nb. of W
General	152,857	1,197,691
a1000	1,000	7,269
a2000	2,000	15,415
a4739	4,739	36,737
$test_A$	4,739	36,191
b1000	1,000	7,894
$test_B$	3,720	28,974

Each expression was classified, by humans, into categories mainly based on the place where the expression was used, such as "airport", "airplane", and "restaurant." These categories are shown in **Table 1**. The categories of "airport" and "business" in Table 1 were used as adaptation target tasks, and the remaining tasks were used as a general task. The number of sentences in these tasks is shown in **Table 2**. The "General" in Table 2 denotes "General Task." To investigate the relationship between the size of data for the adaptation target task and the adaptation effect, we prepared three sizes for the adaptation target task corpus from the "airport" category. These are "a1000", "a2000", and "a4739" in Table 2. Also to investigate the difference in adaptation effect, we prepared "b1000" from the "business" category as another target task corpus. For example, for the translation from Japanese to English, Japanese word sequences on the Japanese side of each "a1000", "a2000", "a4739", and "b1000" were entered into the machine translator. Then, the translation results were used as small corpora

for task adaptation. Sets "$test_A$" and "$test_B$" were also prepared for evaluation. Both the number of sentences and the number of words were counted on the English side of the bilingual corpus.

4.2 Procedure of Experiment

The following two cases are tested:

- English language model adaptation with quasi English sentences obtained from Japanese to English translation (JE-translation),
- Japanese language model adaptation with quasi Japanese sentences obtained from English to Japanese translation (EJ-translation).

In both cases, the perplexities before and after adaptation are compared. Word trigrams* are used as language models in experiments.

For comparison, we also evaluate test set perplexities using the language model adapted with sentences described by humans in the above mentioned parallel corpus.

The adapted language model is created according to the procedure previously summarized at the end of section 2. That is, first, a statistical machine translator (hereafter denoted as SMT) is created with a general task bilingual corpus (both language corpora in general task in Figure 2, "General" in Table 2). Next, a language model of the general task in the translation target language, LM_G, is created with the general task corpus in the target language ("General" in Table 2). This language model is also used in the SMT. Then, the adaptation target task corpus (e.g., "a1000") is entered into the SMT. This experiment uses only the first best translation result for each input to make the target task specific lanugage model LM_T. Finally, the two language models of LM_G and LM_T are combined linearly. Please note that human translations are not used when SMT is used.

In this data, the out-of-vocabulary (oov) rate on test set "$test_A$" is 0.79% before adaptation with the lexicon made only of the general task corpus, and "$test_B$" is 0.0%. After the addition of the adaptation target corpus which is made by humans, the oov rate for $test_A$ changes to 0.77%, 0.76%, and 0.75%, respectively. In this experiment, in order to measure the adaptation effect and to eliminate changes in the probabilities of unknown words depending on differences in vocabulary size, we used a lexicon made only of the general task corpus so that we could fix the vocabulary size before and after adaptation.

The performance of a machine translator is sometimes measured with "Word Error Rate (WER)" the same as speech recognition. WER is defined as

$WER[\%] = 100.0 \times (Sub + Ins + Del)/T$,

where T denotes the total number of words in the correct translation, and Sub, Ins, and Del denote the number of substitution errors, insertion errors, and deletion errors, respectively. Under these experimental conditions, the WER for "airport" task is about 80%, and that for "business" is about 74%. Here, the input sentence is considered to have multiple correct sentences, if one input sentence has multiple output sentences by automatic comparison in the bilingual corpus. As the number of translated sentences is very large, human subjective evaluation was not conducted, but comparatively correct sequences in local contexts were observed.

4.3 Results

4.3.1 English Language Model Adaptation with JE-translation

The perplexities for the English language model adaptation with JE-translation results are shown in **Table 3**. The line of "G" denotes the general corpus in Table 3, and "+a1000" and "+b1000" etc., denote the number of sen-

Table 3: Relationship between amount of data and perplexities (lower limit and results of proposed method (for the translation from Japanese to English))

airport	PP_1	R_1 [%]	PP_2	R_2 [%]
G	32.0	-	32.0	-
G + a1000	23.5	26.7	27.8	13.1
G + a2000	21.8	31.9	27.9	12.8
G + a4739	19.8	38.1	27.9	12.8
business	PP_1	R_1 [%]	PP_2	R_2 [%]
G	55.2	-	55.2	-
G + b1000	50.0	9.42	53.2	3.62

*For smoothing, bigrams and unigrams are also used.

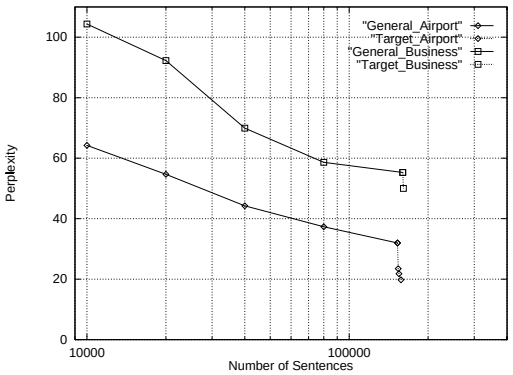

Figure 3: Relationship between corpus size and test set perplexities (for English)

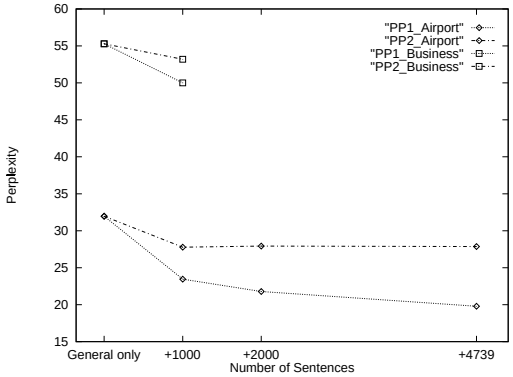

Figure 4: Perplexity comparison (quasi-corpus generated by the proposed method (PP2) v.s. conventional corpus (PP1))

tences to be added. The column of PP_1 lists perplexities under the ideal situation in which a human-made corpus for adaptation is available, and the column of PP_2 lists perplexities obtained by adaptation with machine-translated text. The columns of R_1 and R_2 list the perplexity reduction rates.

Also, the solid lines of **Figure 3** illustrate the relationship between the perplexities and the size of the general corpus. The broken lines illustrate the relationship between the perplexities (PP in PP_1 of Table 3) and the size of the general task corpus plus the size of the target task corpus being added. The upper lines are for the "business" task, and the lower lines are for the "airport." In these figures, the added corpus consists of human-made sentences.

As changes of PP_1 shown in Table 3, the perplexity values obtained with the adapted language model using a general corpus and target task corpus are much smaller (relatively 9.42 to 38%) than the PP obtained only with a general corpus. These results also show the limitation of perplexity reduction if we could have a perfect machine translator that outputs translation as humans do, i.e., relative perplexity reduction is at most 38% after adaptation for "airport," and 9.42% for "business."

Also, Figure 3 predicts that a larger perplexity reduction can be obtained by the adaptation (broken line in Figure 3) than by continuing the collection of data (each extension of the solid line in Figure 3) for the general corpus. Hence, the language model adaptation is useful for the target task, especially for the "airport" task. Moreover, it is clear that we need to prepare an adaptation corpus by some method.

In this research, a statistical machine translator generates the quasi-corpus for adaptation. The results are shown in the column of PP_2 in Table 3. Also, the relationship between additional corpus size and perplexity values are illustrated in **Figure 4**.

A relative 13.1% perplexity reduction was obtained for the "airport" task, and 3.62% for the "business" task, by the language model adapted with a quasi-corpus of the target task that is obtained by machine translation. However, as shown in Figure 4, there is no change after the first addition for the "airport" task. This might be because we only used the first (i.e., the best) translation result for each sentence and because expressions which have higher probabilities appeared in the results, and so we could not obtain various expressions.

However, even with the high WER, the perplexity reduction rate (R_2) obtained by the proposed method was about 30 to 50% of the rate (R_1) obtained in the ideal case in which the human-made data is used. Thus, large gains were made.

4.3.2 Japanese Language Model Adaptation with EJ-translation

Under the same conditions as the previous section except for the translation direction, a Japanese language model was adapted and the adaptation effect was measured as shown in **Table 4**.

The usage of "G", "+a1000" etc., PP_1, PP_2, R_1 and R_2 in Table 4 are the same as in Table

Table 4: Relationship between amount of data and perplexities (lower limit and results of proposed method (for the translation from English to Japanese))

airport	PP_1	R_1 [%]	PP_2	R_2 [%]
G	23.0	-	23.0	-
G + a1000	17.7	22.8	20.0	12.8
G + a2000	16.6	27.6	20.1	12.6
G + a4739	15.4	32.9	20.2	11.9
business	PP_1	R_1 [%]	PP_2	R_2 [%]
G	43.5	-	43.5	-
G + b1000	39.5	9.25	41.3	4.9

3.

As in the previous case, a relative 12.8% perplexity reduction was obtained for the "airport" task, and 4.9% for the "business" task, by the language model adapted with a quasi-corpus of the target task that is obtained by machine translation. These gains (R_2) are about 53 to 56% of the gains (R_1) obtained with human-made sentences.

As shown in Table 3 and Table 4 above, the proposed method was proven to be useful without regard to the translation direction.

4.4 Discussion

We calculated the average ranking of probability for each word in the airport task test set in order to observe whether the machine translation errors might influence prediction probabilities. Average rankings after adaptation with machine translation of 1,000, 2,000 sentences, and all the sentence, were 166.0, 160.2 and 154.8 respectively, while that for no adaptation was 175.2. Thus this result is coincident with perplexity improvements and proves the soundness of the adaptation scheme presented in this paper.

As shown above, despite the translation accuracy, the following facts are clear:

1. The absolute value of the perplexity reduction differs with the target task, but
2. The use of machine-translated text as the task adaptation corpus obtains about 50% of the perplexity reduction rate that is obtained with human-made translations, hence,
3. The proposed scheme is useful for the generation of quasi-corpora as adaptation target task corpora.

As pointed out in 2 above, the gain of adaptation is limited to around 50% of the ideal case in which a human-made corpus of the adaptation target is available. This was because we only used the best translation result for each sentence. We expect that the gain will be improved by using various expressions from the N-best or lattice, or by using results from multiple translators. This is one topic for future work. Moreover, to obtain the same high perplexity reduction as that obtained with human-made translation, machine translators need to be improved until they are able to output multiple expressions that have the same meaning. This is a future research topic for machine translation.

The SMT model training used a large corpus except for the adaptation target task corpus. Although expressions which resemble those in the adaptation target task might be included in the corpus for SMT training, those evaluations did not use this kind of information. In this research, expressions that are useful for adaptation are extracted by SMT and large gains in the language model adaptation are achieved.

When language models are actually used, the topic of the language is coherent to a few topics, for example, as conversations in restaurants are mostly those between restaurant employees and customers, such as ordering and paying for meals, the topics are consistent. However, there is some risk of the appearance of another topic. To allow for this risk, topic identification is introduced before the use of a topic adapted language model, or a mixture of adapted language models is used (Iyer et al., 1994). This paper focused on the language model that is used after the topic identification or mixture. Especially, we focused on the generation of a quasi-corpus necessary for adaptation by machine translation. We consider pre-identification and mixture to be important, but they are other research topics.

5 Related Works

In principal, creating statistical language models requires a large corpus of the application domain, but it is sometimes difficult to obtain. In-

stead, conventional research used limited size, small corpora (Rudnicky, 1995) or documents obtained from the World Wide Web (WWW) (Berger and Miller, 1998) to adapt existing models to the target task. Most researches have tried to improve the performance of dictation tasks and used documents in written language, which are in greater abundance than those in spoken language. Corpora in spoken language are difficult to collect. Also, in the domains of dictation for medicine and human-machine interfaces, where corpus building is also difficult, system developers wrote context free grammars (CFG) for describing the language in these domains, and used them to artificially generate languge data for language model adaptation (Ito et al., 1998; Wang et al., 2000). On the other hand, there are few spoken language corpora, and often, even small new task data do not exist. Moreover, it is difficult to write sufficient grammars for spoken language. To date, there has only been a few attempts such as our proposed scheme that translates a corpus written in the first language to one in a second language and uses the translated corpus for the second language model adaptation.

6 Conclusion

To achieve language model adaptation of one language even when the target task corpus in the same language is not available, this paper proposed a scheme for corpus generation. The proposed scheme generates a quasi-corpus of the target task in the language by using machine translation of the corpus in other languages.

Evaluation showed significantly large perplexity reductions. Perplexity reductions of about 50% were obtained in comparison with a human-made corpus, proving that the proposed method is useful for generation of the corpus for language model adaptations. Also, with the current translation accuracy, a new application domain for machine translation is presented.

References

Adam Berger and Robert Miller. 1998. Just-in-time language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 705–708.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Nobuyasu Ito, Shiho Ogino, and Hitoshi Niijima. 1998. "bunpoo wo riyoo shita n-gram moderu no tasuku tekioo". In *Proceedings of the Fourth Annual Meeting of the Association for Natural Language Processing*, pages 610–613 (in Japanese).

R. Iyer, M. Ostendorf, and J. R. Rohlicek. 1994. Language modeling with sentence-level mixtures. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 82–87.

Hirokazu Masataki, Yoshinori Sagisaka, Kazuya Hisaki, and Tatsuya Kawahara. 1997. Task adaptation using map estimation in n-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech,and Signal Processing (ICASSP'97)*, pages 783–786.

F. J. Och and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proceedings of COLING-2000*, pages 1086–1090.

F. J. Och and H. Ney. 2000b. Improved statistical alignment models. In *Proceedings of ACL-2000*, pages 440–447.

Alexander I. Rudnicky. 1995. Language modeling with limited domain data. In *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, pages 66–69.

C. Tillmann and H. Ney. 2000. Word reordering and dp-based search in statistical machine translation. In *Proceedings of COLING-2000*, pages 850–856.

Ye-Yi Wang, Milind Mahajan, and Xuedong Huang. 2000. A unified context-free grammar and n-gram model for spoken language processing. In *Proceedings of the International Conference on Acoustics, Speech,and Signal Processing (ICASSP'00)*, pages 23–30.