Bilingual Cluster Based Models for Statistical Machine Translation

Hirofumi Yamamoto
National Institute of Information
and Communications Technology
/ 2-2-2 Hikaridai Seika-cho
Soraku-gun Kyoto Japan
and ATR Spoken Language
Communication Research Labs.
hirofumi.yamamoto@nict.go.jp

Eiichiro Sumita
National Institute of Information
and Communications Technology
/ 2-2-2 Hikaridai Seika-cho
Soraku-gun Kyoto Japan
and ATR Spoken Language
Communication Research Labs.
eiichiro.sumita@nict.go.jp

Abstract

We propose a domain specific model for statistical machine translation. It is well-known that domain specific language models perform well in automatic speech recognition. We show that domain specific language and translation models also benefit statistical machine translation. However, there are two problems with using domain specific models. The first is the data sparseness problem. We employ an adaptation technique to overcome this problem. The second issue is domain prediction. In order to perform adaptation, the domain must be provided, however in many cases, the domain is not known or changes dynamically. For these cases, not only the translation target sentence but also the domain must be predicted. This paper focuses on the domain prediction problem for statistical machine translation. In the proposed method, a bilingual training corpus, is automatically clustered into sub-corpora. Each sub-corpus is deemed to be a domain. The domain of a source sentence is predicted by using its similarity to the sub-corpora. The predicted domain (sub-corpus) specific language and translation models are then used for the translation decoding. This approach gave an improvement of 2.7 in BLEU (Papineni et al., 2002) score on the IWSLT05 Japanese to English evaluation corpus (improving the score from 52.4 to 55.1). This is a substantial gain and indicates the validity of the proposed bilingual cluster based models.

1 Introduction

Statistical models, such as n-gram models, are widely used in natural language processing, for example in speech recognition and statistical machine translation (SMT). The performance of a statistical model has been shown to improve when domain specific models are used, since similarity of statistical characteristics between model and target is higher. For utilize of domain specific models, a training data sparseness and target domain estimation problems must be resolved. In this paper, we try to estimate target domain sentence by sentence, considering cases where the domain changes dynamically. After sentence by sentence domain estimation, domain specific models are used for translation using the adaptation technique(Seymore et al., 1997).

In order to train a classifier to predict the domain, we used an unsupervised clustering technique on an unlabelled bilingual training corpus. We regarded each cluster (sub-corpus) as a domain. Prior to translation, the domain of the source sentence is first predicted and this prediction is then used for model selection. The most similar sub-corpus to the translation source sentence is used to represent its domain. After the prediction is made, domain specific language and translation models are used for the translation.

In Section 2 we present the formal basis for our domain specific translation method. In Section 3 we provide a general overview of the two sub-tasks of domain specific translation: domain prediction, and domain specific decoding. Section 4 presents the domain prediction task in depth. Section 5 offers a more detailed description of the details of domain specific decoding. Section 6 gives details of the experiments and presents the results. Finally, Section

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 514–523, Prague, June 2007.

7 offers a summary and some concluding remarks.

2 Domain Specific Models in SMT

The purpose of statistical machine translation is to find the most probable translation in the target language e of a given source language sentence f. This search process can be expressed formally by:

$$\underset{e}{argmax}\ P(e|f) \tag{1}$$

In this formula, the target word sequence (sentence) e is determined only by the source language word sequence f. However, e is heavily dependent on not only on f but also on the domain D. When the domain D is given, formula (1) can be rewritten as the following formula with the introduction of a new probabilistic variable D.

$$\underset{e}{argmax}\ P(e|f, D) \tag{2}$$

This formula can be re-expressed using Bayes' Law.

$$\underset{e}{argmax}\ P(e|D)P(f|e, D) \tag{3}$$

Here, $P(f|e, D)$ represents the domain D specific translation model and $P(e|D)$ represents the domain D specific language model.

When the domain D is known, domain specific models can be created and used in the translation decoding process. However, in many cases, domain D is unknown or changes dynamically. In these cases, both the translation target language sentence e and the domain D must be dynamically predicted at the same time. The following equation represents the process of domain specific translation when the domain D is being dynamically predicted.

$$\begin{aligned} & \underset{e,D}{argmax}\ P(e, D|f) \\ = & \underset{e,D}{argmax}\ P(D|f)P(e|f, D) \end{aligned} \tag{4}$$

The major difference between this equation and formula (3) is that the probabilistic variable D is the prediction target in equation (4). In this equation, $P(D|f)$ represents the domain prediction and $P(e|f, D)$ represents the domain specific translation.

3 Outline of the Proposed Method

Our method can be analysed into two processes: an off-line process and an on-line process. The processes are depicted in figure 1. In the off-line process, bilingual sub-corpora are created by clustering and these clusters represent domains. Domain specific models are then created from the data contained in the sub-corpora in a batch process. In the on-line process, the domain of the source sentence is first predicted and following this the sentence is translated using models built on data from the appropriate domain.

3.1 Off-line process

In this process, the training corpus is clustered to sub-corpora, which are regarded as domains. In SMT, a bilingual corpus is used to create the translation model, and typically, bilingual data together with additional monolingual corpora are used to create the language model. In our method, both the bilingual and monolingual corpora are clustered. After clustering, cluster dependent (domain specific) language and translation models are created from the data in the clusters.

1. A bilingual corpus which is comprised of the training data for the translation model, or equivalently the bilingual part of the training data for the language model is clustered (see Section 4.2).

2. Each sentence of the additional monolingual corpora (if any) is assigned to a bilingual cluster (see Section 4.3).

3. For each cluster, the domain specific (cluster dependent) language models are created.

4. The domain specific translation model is created using only the clusters formed from clustering bilingual data.

3.2 On-line process

This process is comprised of domain prediction and the domain specific translation components. The following steps are taken for each source sentence.

1. Select the cluster to which the source sentence belongs.

2. Translate the source sentence using the appropriate domain specific language and translation models.

4 Domain Prediction

This section details the domain prediction process. To satisfy equation (4), both the domain D and the translation target word sequence e, which maximizes both $P(D|f)$ and $P(e|f, D)$ must be calculated at the same time. However, it is difficult to make the calculations without an approximation. Therefore, in the first step, we find the best candidates for D given the input sentence f. In the next step, $P(e|f, D)$ is maximized over the candidates for D using the following formula.

$$\underset{e}{argmax}\ P(e|f, \underset{D}{argmax}\ P(D|f)) \qquad (5)$$

Equation (5) is approximation of following equation in that can D is regarded as a hidden variable.

$$\underset{e}{argmax} \sum_{D} P(D|f)P(e|D)P(f|e, D)) \qquad (6)$$

When the following assumptions are introduced to equation (6), equation (5) is obtained as an approximation. For only one domain D_i, $P(D_i|f)$ is nearly equal to one. For other domains, $P(D|f)$ are almost zero. $P(D|f)$ can be re-written as following equation.

$$\begin{aligned} & P(D|f) \\ = \ & P(D, f)/P(f) \\ = \ & P(f, D)/P(D) \times P(D)/P(f) \\ = \ & P(f|D)P(D)/P(f) \end{aligned} \qquad (7)$$

Therefore, we can confirm reasonability of this assumption by calculating $P(f|D)P(D)$ all domains ($P(f)$ is constant).

4.1 Domain Definition

When the domain is known in advance, it is usually expressible, for example it could be a topic that matches a human-defined category like "sport". On the other hand, when the domain is delimited in an unsupervised manner, it is used only as a probabilistic variable and does not need to be expressed. Equation (4) illustrates that a good model will provide high probabilities to $P(D|f)P(e|f, D)$ for bilingual sentence pairs (f, e). For the same reason, a good domain definition will lead to a higher probability for the term: $P(D|f)P(e|f, D)$. Therefore, we define the domain D as that which maximizes $P(D|f)P(e|D)$ (an approximation of $P(D|f)P(e|f, D)$). This approximation ensures that the domain definition is optimal for only the language model rather than both the language and translation models. $P(D|f)P(e|D)$ can be re-written as the following equation using Bayes' Law.

$$\begin{aligned} & P(D|f)P(e|D) \\ = \ & P(e|D)P(f|D)P(D)/P(f) \end{aligned} \qquad (8)$$

Here, $P(f)$ is independent of domain D. Furthermore, we assume $P(D)$ to be constant. The following formula embodies the search for the optimal domain.

$$\underset{D}{argmax}\ P(e|D)P(f|D) \qquad (9)$$

This formula ensures that the search for the domain maximizes the domain specific probabilities of both e and f simultaneously.

4.2 Clustering of the bilingual corpus

As mentioned above, we maximize the domain specific probabilities of e and f to ascertain the domain. We define our domains as sub-corpora of the bilingual corpus, and these sub-corpora are formed by clustering bilingually by entropy reduction. For this clustering, the following extension of monolingual corpus clustering is employed (Carter 1994).

1. The total number of clusters (domains) is given by the user.

2. Each bilingual sentence pair is randomly assigned to a cluster.

3. For each cluster, language models for e and f are created using the bilingual sentence pairs that belong to the cluster.

4. For each cluster, the entropy for e and f is calculated by applying the language models from the previous step to the sentences in the cluster. The total entropy is defined as the total sum of entropy (for both source and target) for each cluster.

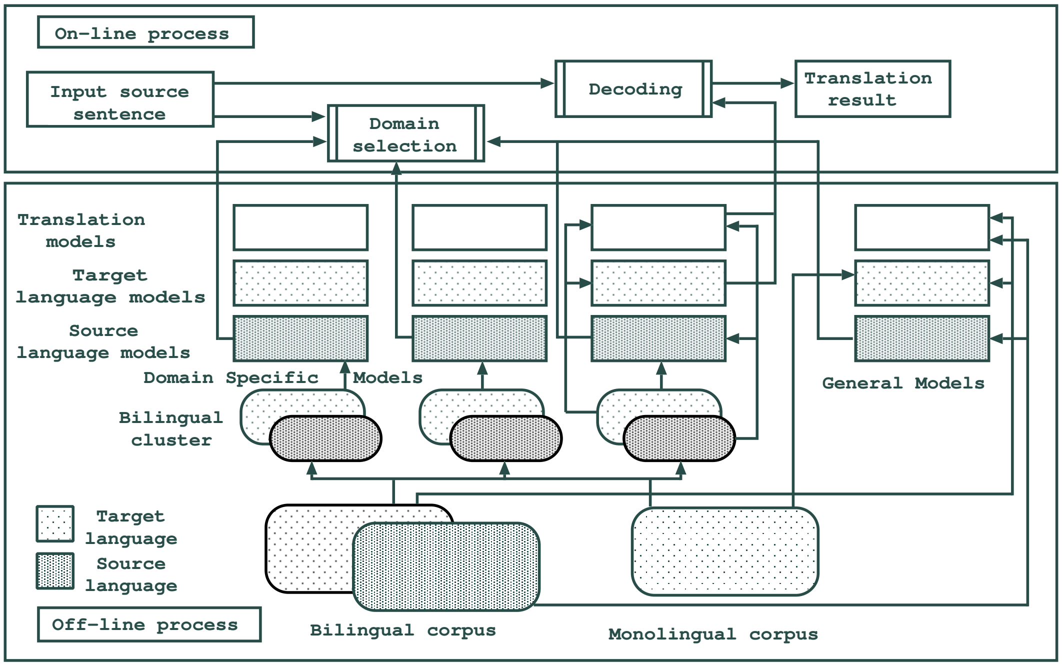

Figure 1: Outline of the Proposed Method

5. Each bilingual sentence pair is re-assigned to a cluster such that the assignment minimizes the total entropy.

6. The process is repeated from step (3) until the entropy reduction is smaller than a given threshold.

4.3 Clustering the monolingual corpus

Any additional monolingual corpora used to train the language model are also clustered. For this clustering, the following process is used.

1. First, bilingual clusters are created using the above process.

2. For each monolingual sentence its entropy is calculated using all the bilingual cluster dependent language models and also the general language model (see Figure 1 for a description of the general language model).

3. If the entropy of the general language model is the lowest, this sentence is not used in the cluster dependent language models.

4. Otherwise, the monolingual sentence is added to the bilingual cluster that results in the lowest entropy.

4.4 Domain prediction

In the process described in the previous section we describe how clusters are created, and we define our domains in terms of these clusters. In this step, domain D is predicted using the given source sentence f. This prediction is equivalent to finding the D that maximizes $P(D|f)$. $P(D|f)$ can be re-written as $P(f|D)P(D)/P(f)$ using Bayes' law. Here, $P(f)$ is a constant, and if $P(D)$ is assumed to be constant (this approximation is also used in the clustering of the bilingual corpus), maximizing the target is reduced to the maximization of $P(f|D)$. To maximize $P(f|D)$ we simply select the cluster D, that gives the highest likelihood of a given source sentence f.

5 Domain specific decoding

After domain prediction, domain specific decoding to maximize $P(e|f,D)$, is conducted. $P(e|f,D)$ can be re-written as the following equation using Bayes' law.

$$\begin{aligned} & P(e|f,D) \\ = \; & P(f|e,D)P(e,D)/P(f,D) \end{aligned}$$

$$= \quad P(f|e,D)P(e|D)P(D)/P(f,D) \quad (10)$$

Here, f is a given constant and D has already been selected by the domain prediction process. Therefore, maximizing $P(f|e,D)P(e|D)$ is equivalent to maximizing the above equation. In $P(f|e,D)P(e|D)$, $P(f|e,D)$ is the domain specific translation model and $P(e|D)$ is the domain specific language model. Equation (10) represents the whole process of translation of f into e using domain D specific models $P(f|e,D)$ and $P(e|D)$.

5.1 Differences from previous methods

5.1.1 Cluster language model

Hasan et al. (2005) proposed a cluster language model for finding the domain D. This method has three steps. In the first step, the translation target language corpus is clustered using human-defined regular expressions. In the second step, a regular expression is created from the source sentence f. In the last step, the cluster that corresponds to the extracted regular expression is selected, and the cluster specific language model built from the data in this cluster is used for the translation. The points of difference are:

- In the cluster language model, clusters are defined by human-defined regular expressions. On the other hand, with the proposed method, clusters are automatically (without human knowledge) defined and created by the entropy reduction based method.
- In the cluster language model, only the translation target language corpus is clustered. In the proposed method, both the translation source and target language corpora are clustered (bilingual clusters).
- In the cluster language model, only a domain (cluster) specific language model is used. In the proposed method, both a domain specific language model and a domain specific translation model are used.

5.1.2 Sentence mixture language model

In equation (6), D is regarded as a hidden variable. Furthermore, when $P(D|f)$ is approximated as $P(D) = D_\lambda$, and the general translation model $P(f|e)$ is used instead of the domain specific translation model $P(f|e,D)$, this equation represents the process of translation using sentence mixture language models (Iyer et al., 1993) as follows:

$$\underset{e}{argmax} \sum_D D_\lambda P(e|D)P(f|e) \quad (11)$$

The points that differ from the proposed method are as follows:

- In the sentence mixture model, the mixture weight parameters D_λ are constant. On the other hand, in the proposed method, weight parameters $P(D|f)$ are estimated separately for each sentence.
- In the sentence mixture model, the probabilities of all cluster dependent language models are summed. In the proposed model, only the cluster that gives the highest probability is considered as approximation.
- In the proposed method, a domain specific translation model is also used.

6 Experiments

6.1 Japanese to English translation

6.1.1 Experimental corpus

To evaluate the proposed model, we conducted experiments based on a travel conversation task corpus. The experimental corpus was the travel arrangements task of the BTEC corpus (Takezawa et al., 2002),(Kikui et al., 2003) and the language pair was Japanese and English. The training, development, and evaluation corpora are shown in Table 1. The development and evaluation corpora each had sixteen reference translations for each sentence. This training corpus was also used for the IWSLT06 Evaluation Campaign on Spoken Language Translation (Paul 2006) J-E open track, and the evaluation corpus was used as the IWSLT05 evaluation set.

6.1.2 Experimental conditions

For bilingual corpus clustering, the sentence entropy must be calculated. Unigram language models were used for this calculation. The translation models were pharse-based (Zen et al., 2002) created using the GIZA++ toolkit (Och et al., 2003). The language models for the domain prediction and translation decoding were word trigram with Good-Turing

Table 1: Japanese to English experimental corpus

	# of sentence	Total words	# of word entry
Japanese Training	40K	355K	12.5K
English Training	40K	315K	9.2K
Japanese Development	510	3,525	918
English Development	510×16	57,388	2,118
Japanese Evaluation	506	3,647	951

backoff (Katz 1987). Ten cluster specific source language models and a general language model were used for the domain prediction. If the general language model provided the lowest perplexity for an input sentence, the domain specific models were not used for this sentence. The SRI language modeling toolkit (Stolcke) was used for the creation of all language models. The PHARAOH phrase-based decoder (Koehn 2004) was used for the translation decoding.

For tuning of the decoder's parameters, including the language model weight, minimum error training (Och 2003) with respect to the BLEU score using was conducted using the development corpus. These parameters were used for the baseline conditions. During translation decoding, the domain specific language model was used as an additional feature in the log-linear combination according to the PHARAOH decoder's option. That is, the general and domain specific language models are combined by log-linear rather than linear interpolation. The weight parameters for the general and domain specific language models were manually tuned using the development corpus. The sum of these language model weights was equal to the language model weight in the baseline. For the translation model, the general translation model (phrase table) and domain specific translation model were linearly combined. The interpolation parameter was again manually tuned using the development corpus.

6.1.3 Experimental results

In our bilingual clustering, the number of clusters must be fixed in advance. Based on the results of preliminary experiments to estimate model order, ten clusters were used. If less than ten clusters were used, domain specific characteristics cannot be represented. If more than ten clusters were used, data sparseness problems are severe, especially in translation models. The amount of sentences in each cluster is not so different, therefore the approximation that $P(D)$ is reasonable. Two samples of bilingual clusters are recorded in the appendix "Sample of Cluster". The cluster A.1 includes many interrogative sentences. The reason is that special words "です か (desu ka)" or "ます か (masu ka)" are used at the end of Japanese sentence with no corresponding word used in English. The cluster A.2 includes numeric expressions in both English and Japanese.

Next, we confirm the reasonability of the assumption used in equation(5). For this confirmation, we calculate $P(D|f)$ for all D for each f (P(D) is approximated as constant). For almost f, only one domain D_i has a vary large value compared with other domains. Therefore, this approximation is confirmed to be reasonable.

In this experiments, we compare three ways of deploying our domain specific models to a baseline. In the first method, only the domain specific language model is used. The ratio of the weight parameter for the general model to the domain specific model was 6:4 for all the domain specific language models. In the second method, only the domain specific translation model was used. The ratio of the interpolation parameter of the general model to the domain specific model was 3:7 for all the domain specific models. In the last method, both the domain specific language and translation models (LM+TM) were used. The weights and interpolation parameters were the same as in the first and second methods. The experimental results are shown in Table 2. Under all of the conditions and for all of the evaluation measures, the proposed domain specific models gave better performance than the baseline. The highest performance came from the system that used both the domain specific language and translation models, resulting in a

2.7 point BLEU score gain over the baseline. It is a very respectable improvement. Appendix ”Sample of Different Translation Results” recodes samples of different translation results with and without the domain specific language and translation models. In many cases, better word order is obtained in with the domain specific models.

6.2 Translation of ASR output

In this experiment, the source sentence used as input to the machine translation system was the direct textual output from an automatic speech recognition (ASR) decoder that was a component of a speech-to-speech translation system. The input to our system therefore contained the kinds of recognition errors and disfluencies typically found in ASR output. This experiment serves to determine the robustness of the domain prediction to real-world speech input. The speech recognition process in this experiment had a word accuracy of 88.4% and a sentence accuracy of 67.2% . The results shown in Table 3 clearly demonstrate that the proposed method is able to improve the translation performance, even when speech recognition errors are present in the input sentence.

6.3 Comparison with previous methods

In this section we compare the proposed method to other comtemporary methods: the cluster language model (CLM) and the sentence mixture model (SMix). The experimental results for these methods were reported by RWTH Aachen University in IWSLT06 (Mauser et al., 2006). We evaluated our method using the same training and evaluation corpora. These corpora were used as the training and development corpora in the IWSLT06 Chinese to English open track, the details are given in Table 4. The English side of the training corpus was the same as that used in the earlier Japanese to English experiments reported in this paper. Each sentence in the evaluation corpus had seven reference translations. Our baseline performance was slightly different from that reported in the RWTH experiments (21.9 BLEU socre for RWTH's system and 21.7 for our system). Therefore, their improved baseline is shown for comparison. The results are shown in Table 5. The improvements over the baseline of our method in both BLEU and NIST (Doddington 2002) score were greater than those for both CLM and SMix. In particular, our method showed improvent in both the BLEU and NIST scores, this is in contrast to the CLM and SMix methods which both degraded the translation performance in terms of the NIST score.

Table 5: Comparison results with previous methods

	BLEU	NIST	WER	PER
RWTH	21.9	6.31	66.4	50.8
Our	21.7	6.79	70.9	51.2
CLM	+0.6	-0.22	-2.7	-1.1
SMix	+0.2	-0.06	-1.1	-0.9
Proposed	+1.1	+0.17	-1.1	-0.5

6.4 Clustering of the monolingual corpus

Finally, we evaluated the proposed method when an additional monolingual corpus was incorporated. For this experiment, we used the Chinese and English bilingual corpora that were used in the NIST MT06 evaluation (NIST 2006). The size of the bilingual training corpus was 2.9M sentence pairs. For the language model training, an additional monolingual corpus of 1.5M English sentences was used. NIST 2006 development (evaluation set for NIST 2005) is used for evaluation. In this experiment, the test set language model perplexity of a model built on only the monolingual corpus was considerably lower than that of a model built from only the target language sentences from the bilingual corpus. Therefore, we would expect the use of this monolingual corpus to be an important factor affecting the quality of the translation system. These perplexities were 299.9 for the model built on only the bilingual corpus, 200.1 for the model built on only the monolingual corpus, and 192.5 for the model built on a combination of the bilingual and monolingual corpora. For the domain specific models, 50 clusters were created from the bilingual and monolingual corpora. In this experiment, only the domain specific language model was used. The experimental results are shown in Table 6. The results in the table show that the incorporation of the additional monolingual data has a pronounced beneficial effect on performance, the performance improved according to all of the evaluation measures.

Table 2: Japanese to English translation evaluation scores

	BLEU	NIST	WER	PER	Meteor	TER
Baseline	52.38	9.316	42.87	33.21	70.63	35.46
Domain Specific LM	53.66	9.349	41.73	32,27	71.39	34.17
Domain Specific TM	54.30	9.333	41.64	32.50	71.77	33.80
Domain Specific LM+TM	55.09	9.451	41.05	31.63	72.09	33.20

Table 3: Evaluation using ASR output

	BLEU	NIST	WER	PER	Meteor	TER
Baseline	48.17	8.892	47.05	36.86	67.40	39.36
Domain Specific LM	48.94	8.900	46.26	36.37	67.98	38.42
Domain Specific TM	49.11	8.842	45.78	36.55	68.01	37.88
Domain Specific LM+TM	50.12	9.001	45.26	35.80	68.05	37.22

7 Conclusion

We have proposed a technique that utilizes domain specific models based on bilingual clustering for statistical machine translation. It is well-known that domain specific modeling can result in better performance. However, in many cases, the target domain is not known or can change dynamically. In such cases, domain determination and domain specific translation must be performed simultaneously during the translation process. In the proposed method, a bilingual corpus was clustered using an entropy reduction based method. The resulting bilingual clusters are regarded as domains. Domain specific language and translation models are created from the data within each bilingual cluster. When a source sentence is to be translated, its domain is first predicted. The domain prediction method selects the cluster that assigns the lowest language model perplexity to the given source sentence. Translation then proceeds using a language model and translation model that are specific to the domain predicted for the source sentence.

In our experiments we used a corpus from the travel domain (the subset of the BTEC corpus that was used in IWSLT06). Our experimental results clearly demonstrate the effectiveness of our method. In the Japanese to English translation experiments, the use of our proposed method improved the BLEU score by 2.7 points (from 52.4 to 55.1). We compared our approach to two previous methods, the cluster language model and sentence mixture model. In our experiments the proposed method yielded higher scores than either of the competitive methods in terms of both BLEU and NIST. Moreover, our method may also be augmented when an additional monolingual corpus is avaliable for building the language model. Using this approach we were able to further improve translation performance on the data from the NIST MT06 evaluation task.

A Sample of Cluster

A.1 Cluster 1

- E: do you do alterations
  J: 直し は し てい ます か (naoshi wa shi tei masu ka)

- E: what's the newest color in this season
  J: 今年 の 新色 は どれ です か (kotoshi no shinshoku wa dore desu ka)

- E: are there any baseball games today
  J: 今日 野球 の 試合 は あり ます か (kyou yakyu no shiai wa ari masu ka)

- E: where's the nearest perfumery
  J: 最寄り の 香水 店 は どこ です か (moyori no kousui ten wa doko desu ka)

- E: how much is the breakfast
  J: 朝食 は いくら です か (choshoku wa ikura desu ka)

Table 4: Training and evaluation corpora used for comparison with previous methods

	# of sentence	Total words	Vocabulary size
English Training	40K	315K	9.2K
Chinese Training	40K	304K	18.7K
Chinese Evaluation	489	5,110	1.3K

Table 6: Experimental results with monolingual corpus

	BLEU	NIST	WER	PER	Meteor	TER
Baseline	24.39	7.918	86.51	61.65	53.36	68.21
Proposed	24.95	8.030	85.89	61.27	53.86	67.48

A.2 Cluster 2

- E: mr. aoki yes a single room for two nights
  J: アオキ さん です ねえー シングルルーム で 二 泊 です ね (aoki san desu ne ee shingururumu de 2 haku desu ne)
- E: may i have the key to room two fifteen
  J: 二 一 五 号 室 の 鍵 を 下さい (2 1 5 gou shitsu no kagi o kudasai)
- E: i'd like extension twenty four please
  J: 内線 二 十 四 を 御 願い し ます (naisen 24 o o begai shi masu)
- E: the flight number is se one o three to tokyo on the second of april
  J: フライトナンバー は 東京 行き エスイー 一 ゼロ 三 便 四月 二日 の 便 です (furaitonanba wa tokyo iki s e 1 0 3 bin 4 gatsu futsuka no bin desu)
- E: delta airlines flight one one two boarding is delayed
  J: デルタ航空 一 一 二 便 は 搭乗 が 遅れ てい ます (derutakouku 1 1 2 bin wa tojo ga okure tei masu)

B Sample of Different Translation Results

1. Ref: your room is number two ten
   Base: your room this is room two o one
   LM: your room is this is room two one zero
   TM: your room is room two o one
   LM+TM: your room is this is room two one zero
2. Ref: where is a spot where there are a lot of fish
   Base: i'm spot where is the lot of fish
   LM: where is the spot are a lot of fish
   TM: i'm spot where is the lot of fish
   LM+TM: where is the spot are a lot of fish
3. Ref: i don't like the design
   Base: design i don't like it
   LM: i don't like it design
   TM: i don't like the design
   LM+TM: i don't like the design
4. Ref: where can i contact you
   Base: where contact if i may
   LM: where contact if i can
   TM: where can i contact
   LM+TM: where can i contact
5. Ref: where is a police station where japanese is understood
   Base: japanese where's the police station
   LM: japanese where's the police station
   TM: where's the police station where someone understands japanese
   LM+TM: where's the police station where someone understands japanese

References

K. Seymore, R. Rosenfeld, ”Using Story Topics for Language Model Adaptation,” Proc. EUROSPEECH, pp. 1987-1990, 1997.

David Carter, ”Improving Language Models by Clustering Training Sentences,” Proc. ACL, pp. 59-64, 1994.

S. Hasan, H. Ney, ”Clustered Language Models Based on Regular Expressions for SMT,” Proc. EAMT, Budapest, Hungary, May 2005.

R. M. Iyer and M. Ostendorf, ”Modeling Long Distance Dependence in Language: Topic mixture versus dynamic cache models,” IEEE Transactions on Speech and Audio Processing, 1994.

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,” Proc. Conference on Language Resource and Evaluation, May 2002.

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, Seiichi Yamamoto, ”Creating Corpora for Speech-to-Speech Translation,” Proc. EUROSPEECH, pp. 381-384, 2003.

M. Paul, ”Overview of the IWSLT 2006 Evaluation Campaign,” IWSLT 2006, Nov. 2006.

R. Zens, F. J. Och, H. Ney, ”Phrase-based statistical machine translation,” 25th German Conference on Artificial Inteligence, Sep 2002.

F. J. Och, H. Ney, ”A Systematic Comparison of Various Statistical Alignment Models,” Computational Linguistics, No. 1, Vol. 29, pp. 19-51, 2003.

S. M. Katz, ”Estimation of Probabilities from Sparse Data for Language Model Component of a Speech Recognizer,” IEEE Trans. on Acoustics, Speech, and Signal Processing, pp. 400-401, 1987.

A. Stolcke, ”SRILM - An Extensible Language Model Toolkit,” http://www.speech.sri.com/projects/srilm/

P. Koehn, ”PHARAOH: A beam search decoder for phrase-based statistical machine translation models,” http://www.isi.edu/publications/licensed-sw/pharaoh/

F. J. Och, ”Minimum error rate training for statistical machine trainslation,” Proc. ACL, 2003.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, ”Bleu: a method for automatic evaluation of machine translation,” Proc. ACL, 2002.

A. Mauser, R. Zens, E. Matusov, S. Hasan, H. Ney, ”The RWTH Statistical Machine Translation System for IWSLT 2006 Evaluation,” IWSLT 2006, Nov. 2006.

G. Doddington, ”Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” Proc. ARPA Workshop on Human Language Technology, 2002.

NIST 2006 Machine Translation Evaluation, http://www.nist.gov/speech/tests/mt/mt06eval_officiaL results.html