IEICE TRANS. INF. & SYST., VOL.E91–D, NO.3 MARCH 2008

PAPER Special Section on Robust Speech Processing in Realistic Environments

Bilingual Cluster Based Models for Statistical Machine Translation

Hirofumi YAMAMOTO[†,††a)] *and* Eiichiro SUMITA[†,††], *Members*

SUMMARY 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 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.

key words: *statistical machine translation, domain specific model, domain estimation, sentence clustering*

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 [1].

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. After here, we use "cluster" instead of "domain". Prior to translation, the cluster 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 cluster. After the prediction is made, cluster specific language and translation models are used for the translation.

In Sect. 2 we present the formal basis for our cluster specific translation method. In Sect. 3 we provide a general overview of the two sub-tasks of cluster specific translation: cluster prediction, and cluster specific decoding. Section 4 presents the cluster prediction task in depth. Section 5 offers a more detailed description of the details of cluster specific decoding. Section 6 gives details of the experiments and presents the results. Finally, Sect. 7 offers a summary and some concluding remarks.

2. Cluster 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:

$$\begin{aligned} & \underset{e}{argmax}\ P(e|f) \\ = & \underset{e}{argmax}\ P(e)P(f|e) \end{aligned} \tag{1}$$

Here, $P(e)$ represents language model and $P(f|e)$ represents translation model. When source language sentence f is replaced by acoustic feature sequence A and target language sentence e is replaced by recognition target sentence W, Eq. (1) just represents speech recognition. Currently, phrase-based model [2] is the most standard translation model. For another type of translation model, there is a bilingual n-gram based model [3]. In Eq. (1), the target word sequence (sentence) e is determined only by the source language word sequence f. However, e is heavily dependent not only on f but also on the Cluster C. When the Cluster C is given, formula (1) can be rewritten as the following formula with the introduction of a new probabilistic variable C.

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

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

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

Here, $P(f|e, C)$ represents the cluster C specific translation model and $P(e|C)$ represents the cluster C specific language

Manuscript received July 6, 2007.
Manuscript revised September 12, 2007.

[†]The authors are with National Institute of Communications Technology, Kyoto-fu, 619–0288 Japan.
[††]The authors are with ATR Spoken Language Translation Research Laboratories, Kyoto-fu, 619–0288 Japan.
a) E-mail: hirofumi.yamamoto@nict.go.jp

DOI: 10.1093/ietisy/e91–d.3.588

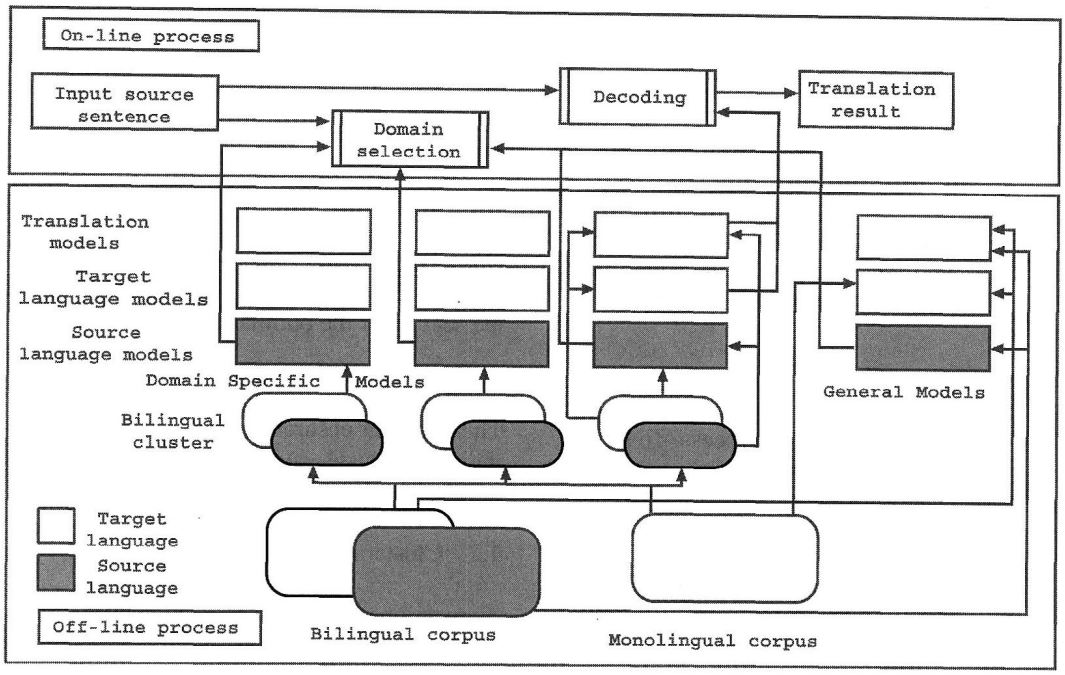

Fig. 1 Outline of the proposed method.

model.

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

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

The major difference between this equation and formula (3) is that the probabilistic variable C is the prediction target in Eq. (4). In this equation, $P(C|f)$ represents the cluster prediction and $P(e|f, C)$ represents the cluster specific translation. When Eq. (4) is applied to speech recognition, the following equation is obtained by replacing f to A and e to W.

$$\underset{W,C}{argmax}\ P(C|A)P(W|A, C) \tag{5}$$

In the above formula, $P(C|A)$ represents the cluster prediction. However, it is very difficult to predict a cluster that represents language information from acoustic information A. Therefore, in speech recognition, C is indirectly predicted. For example, in the first step, W is decoded using each cluster (multiple decoding). In this step, n (n is number of clusters) candidates for W are obtained. In the next step, candidate for W which gives the highest likelihood is selected as final recognition result for W [4]. On the other hand, in statistical machine translation, C can be directly predicted from f, since f is language information.

3. Outline of the Proposed Method

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

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 specific (dependent) 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 Sect. 4.2).
2. Each sentence of the additional monolingual corpora (if any) is assigned to a bilingual cluster (see Sect. 4.3).
3. For each cluster, the cluster specific language models are created.
4. The cluster specific translation model is created using only the clusters formed from clustering bilingual data.

3.2 On-Line Process

This process is comprised of cluster prediction and the cluster 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 cluster specific language and translation models.

4. Cluster Prediction

This section details the cluster prediction process. To satisfy Eq. (4), both the cluster C and the translation target word sequence e, which maximizes both $P(C|f)$ and $P(e|f,C)$ 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 C given the input sentence f. In the next step, $P(e|f,C)$ is maximized over the candidates for C using the following formula.

$$\underset{e}{argmax}\, P(e|f, \underset{C}{argmax}\, P(C|f)) \tag{6}$$

Equation (6) is approximation of following equation in that can C is regarded as a hidden variable.

$$\underset{e}{argmax} \sum_{C} P(C|f)P(e|C)P(f|e,C)) \tag{7}$$

When the following assumptions are introduced to Eq. (7), Eq. (6) is obtained as an approximation. For only one cluster C_i, $P(C_i|f)$ is nearly equal to one. For other clusters, $P(C|f)$ are almost zero. $P(C|f)$ can be re-written as following equation.

$$\begin{aligned} & P(C|f) \\ &= P(C,f)/P(f) \\ &= P(f,C)/P(C) \times P(C)/P(f) \\ &= P(f|C)P(C)/P(f) \end{aligned} \tag{8}$$

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

4.1 Cluster Definition

When the cluster 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 cluster 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 $P(C|f)P(e|f,C)$ for bilingual sentence pairs (f,e). For the same reason, a good cluster definition will lead to a higher probability for the term: $P(C|f)P(e|f,C)$. Therefore, we define the cluster C as that which maximizes $P(C|f)P(e|C)$ (an approximation of $P(C|f)P(e|f,C)$). This approximation ensures that the cluster definition is optimal for only the language model rather than both the language and translation models. $P(C|f)P(e|C)$ can be re-written as the following equation using Bayes' law.

$$\begin{aligned} & P(C|f)P(e|C) \\ &= P(e|C)P(f|C)P(C)/P(f) \end{aligned} \tag{9}$$

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

$$\underset{C}{argmax}\, P(e|C)P(f|C) \tag{10}$$

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

4.2 Clustering of the Bilingual Corpus

As mentioned above, we maximize the cluster specific probabilities of e and f to ascertain the cluster. We define our clusters as sub-corpora of the bilingual corpus, and these sub-corpora are formed by clustering bilingually by entropy reduction. Clustering methods are roughly categorized to two types, hard and soft clustering. Generally, soft clustering methods give better clustering performance. However, we used a hard clustering method due to the extra complexity involved in soft clustering. That is, in soft clustering, each sentence belongs each cluster with a membership probability. Therefore, the language and translation model for each cluster must be created taking these probabilities into account. For this hard clustering, the following extension of monolingual corpus clustering was employed [5].

1. The total number of clusters 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.
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 Fig. 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 Cluster Prediction

In the process described in the previous section we describe how clusters are created. In this step, cluster C is predicted using the given source sentence f. This prediction is equivalent to finding the C that maximizes $P(C|f)$. $P(C|f)$ can be re-written as $P(f|C)P(C)/P(f)$ using Bayes' law. Here, $P(f)$ is a constant, and if $P(C)$ 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 $P(f|C)$. To maximize $P(f|C)$ we simply select the cluster C, that gives the highest likelihood of a given source sentence f.

5. Cluster Specific Decoding

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

$$\begin{aligned} & P(e|f, C) \\ = \; & P(f|e, C)P(e, C)/P(f, C) \\ = \; & P(f|e, C)P(e|C)P(C)/P(f, C) \end{aligned} \tag{11}$$

Here, f is a given constant and C has already been selected by the cluster prediction process. Therefore, maximizing $P(f|e, C)P(e|C)$ is equivalent to maximizing the above equation. In $P(f|e, C)P(e|C)$, $P(f|e, C)$ is the cluster specific translation model and $P(e|C)$ is the cluster specific language model. Equation (11) represents the whole process of translation of f into e using cluster C specific models $P(f|e, C)$ and $P(e|C)$.

5.1 Differences from Previous Methods

5.1.1 Cluster Language Model

Hasan et al.[6] proposed a cluster language model for finding the cluster C. 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 cluster specific language model is used. In the proposed method, both a cluster specific language model and a cluster specific translation model are used.

5.1.2 Sentence Mixture Language Model

In Eq. (7), C is regarded as a hidden variable. Furthermore, when $P(C|f)$ is approximated as $P(C) = C_\lambda$, and the general translation model $P(f|e)$ is used instead of the cluster specific translation model $P(f|e, C)$, this equation represents the process of translation using sentence mixture language models [7] as follows:

$$\underset{e}{argmax} \sum_{C} C_\lambda P(e|C)P(f|e) \tag{12}$$

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

- In the sentence mixture model, the mixture weight parameters C_λ are constant. On the other hand, in the proposed method, weight parameters $P(C|f)$ 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 cluster 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 [8], [9] 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 [10] J-E open track, and the evaluation corpus was used as the IWSLT05 evaluation set.

Table 1 Japanese to English experimental corpus.

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

Table 2 Number of entries in translation and language models.

	translation (phrase table)	1-grams	2-grams	3-grams
# of entries	124 K	28 K	340 K	200 K

6.1.2 Experimental Conditions

For bilingual corpus clustering, the sentence entropy must be calculated. Unigram language models were used for this calculation. The experimental target translation model type is phrase-based that is created using the GIZA++ toolkit [11]. Logically, the proposed method is independent from translation model type. However effectiveness for another type of translation model is not confirmed in the experiment. The language models for the cluster prediction and translation decoding were word trigram with Good-Turing backoff [12]. Ten cluster specific source language models and a general language model were used for the cluster prediction. If the general language model provided the lowest perplexity for an input sentence, the cluster specific models were not used for this sentence. The SRI language modeling toolkit [13] was used for the creation of all language models. The PHARAOH phrase-based decoder [14] was used for the translation decoding. Size of translation and language models are shown in Table 2.

For tuning of the decoder's parameters, including the language model weight, minimum error training [15] with respect to the BLEU [16] score was conducted using the development corpus. These parameters were used for the baseline conditions. During translation decoding, the cluster specific language model and general language model are combined by log-linear according to the PHARAOH decoder's option. The weight parameters for the general and cluster 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 cluster specific translation model were linearly combined also according to the PHARAOH decoder's constraint. 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. We had experimented in five, ten and fifteen clusters. If a small number of clusters were used, cluster specific characteristics cannot be represented. If a large number of clusters were used, data sparseness problems are severe, especially in translation models. In any number of clusters, the amount of sentences in each cluster is not so different, therefore the approximation that $P(C)$ is constant seems to be reasonable. Samples of bilingual clusters (number of cluster is ten) are recorded in the appendix "Sample of Cluster". Same clusters seem to represent topics such as money in Cluster 3, food in Cluster 4 and number in Cluster 6. The other clusters seem to represent sentence style such as request sentences in Cluster 2 and question sentences in Cluster 9. However, in some clusters, for example Cluster 0 and Cluster7, the measure of cluster is difficult to understand for human sense.

Next, we confirm the reasonability of the assumption used in Eq. (6). For this confirmation, we calculate $P(C|f)$ for all C for each f (P(C) is approximated as constant). For almost f, only one cluster C_i has a very large value compared with other clusters. Therefore, this approximation is confirmed to be reasonable.

In these experiments, we compare three ways of deploying our cluster specific models to a baseline. In the first method, only the cluster specific language model is used. The ratio of the weight parameter for the general model to the cluster specific model was 6:4 for all the cluster specific language models. In the second method, only the cluster specific translation model was used. The ratio of the interpolation parameter of the general model to the cluster specific model was 3:7 for all the cluster specific models. In the last method, both the cluster 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 3. Under all of the conditions and for all of the evaluation measures, the proposed cluster specific models gave better performance than the baseline. The highest performance came from the system that used both the cluster specific language and translation models in ten clusters, resulting in a 2.7 point BLEU score gain over the baseline. Under this condition, cluster prediction accuracy (if predicted cluster also gives the highest probability for reference English sentence, cluster prediction is regarded as correct) is 69% (347 in 506 sentences). For 86% sentences (437 in 506 sentences), cluster specific models are used. It is a very respectable improvement. Appendix "Sample of Different Translation Results" records samples of different translation results with and without the cluster specific language and translation models. In many cases, better word order is obtained in the cluster specific models.

Table 3 Japanese to English translation evaluation scores.

	# of clusters	BLEU	NIST	WER	PER	Meteor	TER
Baseline		52.38	9.316	42.87	33.21	70.63	35.46
Cluster Specific LM	5	53.61	9.325	41.88	32,49	71.30	34.45
Cluster Specific LM	10	53.66	9.349	41.73	32,27	71.39	34.17
Cluster Specific LM	15	53.22	9.255	42.49	33,21	70.98	34.86
Cluster Specific TM	5	53.10	9.263	42.03	32.56	71.25	34.49
Cluster Specific TM	10	54.30	9.333	41.64	32.50	71.77	33.80
Cluster Specific TM	15	53.49	9.347	41.40	32.29	71.04	34.29
Cluster Specific LM+TM	5	53.77	9.344	41.17	32.00	71.90	33.89
Cluster Specific LM+TM	10	**55.09**	**9.451**	**41.05**	**31.63**	**72.09**	**33.20**
Cluster Specific LM+TM	15	54.18	9.352	**41.05**	32.15	70.92	33.91

Table 4 Evaluation using ASR output.

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

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 cluster 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% . In this experiment, number of cluster is ten. The results shown in Table 4 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 contemporary 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 [17]. 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 5. 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 with ten clusters are shown in Table 6. The improvements over the baseline of our method in both BLEU and NIST [18] score were

Table 5 Training and evaluation corpora used for comparison with previous methods.

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

Table 6 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

greater than those for both CLM and SMix. In particular, our method showed improvements 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.

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 [19]. The size of the bilingual training corpus was 2.9 M sentence pairs. For the language model training, an additional monolingual corpus of 1.5 M 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

Table 7 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

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 cluster specific models, 50 clusters (this number of clusters is not optimized) were created from the bilingual and monolingual corpora. In this experiment, only the cluster specific language model was used. The experimental results are shown in Table 7. 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.

7. Conclusion

We have proposed a technique that utilizes cluster specific models based on bilingual clustering for statistical machine translation. It is well-known that cluster specific modeling can result in better performance. However, in many cases, the target cluster is not known or can change dynamically. In such cases, cluster determination and cluster 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 clusters. Cluster specific language and translation models are created from the data within each bilingual cluster. When a source sentence is to be translated, its cluster is first predicted. The cluster 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 cluster predicted for the source sentence.

In our experiments we used a corpus from the travel cluster (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 available 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.

References

[1] K. Seymore and R. Rosenfeld, "Using story topics for language model adaptation," Proc. EUROSPEECH, pp.1987–1990, 1997.
[2] R. Zens, F.J. Och, and H. Ney, "Phrase-based statistical machine translation," 25th German Conference on Artificial Intelligence, Sept. 2002.
[3] J. Marino, R. Banchs, J. Grego, A. de Gispert, P. Lambert, M. Costa-jussa, and J. Fonollosa, "Bilingual N-gram statistical machine translation," Proc. MT Summit X, pp.275–282, Sept. 2005.
[4] T. Shimizu, S. Kuroiwa, and N. Higuchi, "Conversational speech recognition using sentence style related multi N-grams," Proc. ASRU, pp.245–248, 1999.
[5] D. Carter, "Improving language models by clustering training sentences," Proc. ACL, pp.59–64, 1994.
[6] S. Hasan and H. Ney, "Clustered language models based on regular expressions for SMT," Proc. EAMT, Budapest, Hungary, May 2005.
[7] R.M. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixture versus dynamic cache models," IEEE Trans. Speech Audio Process., vol.7, no.1, pp.30–39, 1994.
[8] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," Proc. Conference Language Resource and Evaluation, May 2002.
[9] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," Proc. EUROSPEECH, pp.381–384, 2003.
[10] M. Paul, "Overview of the IWSLT 2006 evaluation campaign," IWSLT 2006, Nov. 2006.
[11] F.J. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, no.1, vol.29, pp.19–51, 2003.
[12] S.M. Katz, "Estimation of probabilities from sparse data for language model component of a speech recognizer," IEEE Trans. Acoust. Speech Signal Process., vol.35, no.3, pp.400–401, 1987.
[13] A. Stolcke, "SRILM — An extensible language model toolkit," http://www.speech.sri.com/projects/srilm/
[14] P. Koehn, "PHARAOH: A beam search decoder for phrase-based statistical machine translation models," http://www.isi.edu/publications/licensed-sw/pharaoh/
[15] F.J. Och, "Minimum error rate training for statistical machine trainslation," Proc. ACL, 2003.
[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," Proc. ACL, 2002.
[17] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH statistical machine translation system for IWSLT 2006 evaluation," IWSLT 2006, Nov. 2006.
[18] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," Proc. ARPA Workshop on Human Language Technology, 2002.
[19] NIST 2006 Machine Translation Evaluation, http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html

Appendix A: Sample of Cluster

A.1 Cluster 0

Training sentence

- E: we want to have a table near the window
  J: 窓際の席を御願いします (modogiwa no seki wo o negai shi masu)
- E: i don't want it extracted
  J: 抜かないでください (nuka naidekudasai)
- E: pickpocket
  J: すりだ (suri da)

- E: can you sew on this button
  J: このボタンを付けて頂けますか (kono botan wo tsuke teitadake masu ka)
- E: i requested a no smoking seat but this is a smoking seat
  J: 禁煙席を指定したのにここは喫煙席です (kinnen seki wo shitei shi ta noni koko wo kitsuen seki desu)

Evaluation sentence

- 鞄を開けて頂けますか (kaban wo ake teitadake masu ka)
- ここを押すだけです (koko wo osu dake desu)
- タクシーにバッグを置き忘れてしまいました (takushii ni baggu wo oki wasure teshimai masu ta)

A.2 Cluster 1

Training sentence

- E: no worry about that i'll take it and you need not wrap it up
  J: 結構ですそれを頂きましょう包まなくても構いません (kekkou desu sore wo itadaki masho u tsutsuma naku te mo kamai masen)
- E: i can't guarantee that but we will try our best
  J: 御約束はできませんができるだけその様にさせて頂きます (o yakusoku wa deki masen ga dekiru-dake sonoyouni sa se teitadaki masu)
- E: sorry this space is already taken
  J: ごめんなさいこの場所は取ってあるんです (gomen nasai kono basho wa to ttearu n desu)
- E: thank you that's enough
  J: もうこれで十分ですどうもありがとう (mou kore de juubun desu doumo arigatou)
- E: all right . and how about you ma'am
  J: 承知しましたあなたはいかがなさいますか (shouchi shi mashi ta anata wo ikaga nasai masu ka)

Evaluation sentence

- そうですかパウエル通りをどのくらい歩くのですか (sou desu ka pauerudoori wo donokurai aruku no desu ka)
- いつか昼食をご一緒にいかがですホテルの近くに良いレストランを見つけたんですよ (itsu ka choushoku wo goissho ni ikaga desu hoteru no chikaku ni yoi resutoran wo mitsuke ta n desu yo)
- どうしてですか予約しておいたんですよ (dou shi te desu ka yoyaku shi teoi ta n desu yo)

A.3 Cluster 2

Training sentence

- E: i'll tell her to call you as soon as she returns
  J: 彼女が戻り次第電話させます (kanojo ga modori shidai danwa sa se masu)
- E: help me please call an ambulance please
  J: 助けて下さい救急車を呼んで下さい (tasuke tekudasai kyuukyuusha wo yon dekudasai)
- E: if it should turn up please call me
  J: もし見つかったら電話を下さい (moshi mitsuka ttara denwa wo kudasai)
- E: please write down your address here
  J: 住所をここに書いて下さい (juusho wo koko ni kai tekudasai)
- E: full tank please
  J: 満たんにして下さい (mantan ni shi tekudasai)

Evaluation sentence

- ちょっと待って下さい (chotto ma ttekudasai)
- もう少し大きな声で話して下さい (mou shukoshi ookina koe de hanashi tekudasai)
- この用紙に記入して下さい (kono youshi ni kinyuu shi tekudasai)

A.4 Cluster 3

Training sentence

- E: of course it was ninety nine dollars wasn't it
  J: 分かりました九十九ドルですね (wakari masi ta 99 doru desu ne)
- E: i've paid for meals and hotel charges in advance
  J: 食事代と宿泊料は前払いしてあります (shokuji dai to shukuhaku ryou ha maebarai shi teari masu)
- E: into french francs please
  J: これをフランスフランに換えて下さい (kore wo furansufuran ni kae tekudasai)
- E: could you cash my two hundred dollar check
  J: この二百ドル小切手を現金に換えて下さい (kono 200 doru kogitte wo genkin ni kae tekudasai)
- E: i have about three thousand dollars
  J: 約三千ドル持っています (yaku 3000 doru mo ttei masu)

Evaluation sentence

- 旅行小切手は使えますか (ryokou kogitte wa tsukae masu ka)
- これをユーケーポンドに両替して下さい (kore wo yuukeipondo ni ryougae shi tekudasai)
- クレジットカードで支払えますか (kurejittokaado de shiharae masu ka)

A.5 Cluster 4

Training sentence

- E: pass the bread please
  J: パンを回して下さい (pan wo mawashi tekudasai)
- E: would you clean these clothes
  J: この服の洗濯を御願いできますか (kono fuku no sentaku wo o negai deki masu ka)

- E: which wine would go well with grilled salmon
  J: サーモンの塩焼き に 合う ワイン は どれ です か (saamonnoshioyaki ni au wain wa dore desu ka)
- E: i'd like some crayfish
  J: ザリガニ が 欲しい の です が (zarigani ga hoshii no desu ga)
- E: would you like cream and sugar in your coffee
  J: コーヒー に ミルク と 御 砂糖 を 御 入れ し ます か (koohii ni miruku to o satou wo o ire shi masu ka)

Evaluation sentence

- 玉ねぎ は 入れ ない で (tamanegi wa ire nai de)
- コーラ の 中 を 一つ 下さい (koora no chuu wo hitotsu kudasai)
- ズボン が 要る の です が (zubon ga iru no desu ga)

A.6 Cluster 5

Training sentence

- E: the light was red
  J: 信号 は 赤 でし た (shingou wa aka deshi ta)
- E: i've never heard of this address around here
  J: この 住所 は この 辺 で 聞い た こと ない です ね (kono juusho wa kono hen de kii ta koto nai desu ne)
- E: i have a sore pain here
  J: ここ が ひりひり 痛み ます (koko ga hirihiri itami masu)
- E: it's too heavy
  J: これ は 重 すぎ ます (kore wa omo sugi masu)
- E: she's seriously injured
  J: 彼女 は ひどい けが を し てい ます (kanojo wa hidoi kega wo shi tei masu)

Evaluation sentence

- テレビ が 点き ません (terebi ga tsuki masen)
- 体 が 怠い の です (karada ga darui no desu)
- 下痢 を し てい ます (geri wo shi tei masu)

A.7 Cluster 6

Training sentence

- 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)

Evaluation sentence

- 日本 の 大使館 の 電話番号 は 何 番 です か (nihon no taishikan no denwabangou wa nan ban desu ka)
- 御 客 様 の 部屋 は 二 一 ゼロ 号 室 です (go kyaku sama no heya wa 2 1 0 gou shitsu desu)
- 一 ゼロ 七 ゼロ 号 室 です が (1 0 7 0 gou shitsu desu ga)

A.8 Cluster 7

Training sentence

- E: in my case it is usually on business seldom for pleasure
  J: 私 の 場合 大抵 仕事 で 滅多 に 遊び で は 行き ません (watashi no baai taitei shigoto de metta ni asobi de wa iki masen)
- E: i'll be staying two days
  J: 二日 間 滞在 する 予定 です (futsuka kan taizai suru yotei desu)
- E: the japanese islands run northeast to southwest in the northwestern part of the pacific ocean
  J: 日本列島 は 太平洋 の 北西 に 北東 から 南西 の 方向 に 伸び てい ます (nihonrettou wa taiheiyou no hokusei ni hokutou nansei no houkou ni nobi tei masu)
- E: mary is not so old as henry
  J: メアリー は ヘンリー ほど の 年 で は あり ません (mearii wa henrii hodo no toshi de wa ari masen)
- E: i'm calling from the airport i'll be there around four
  J: 空港 から 電話 し てい ます 四時 頃 に は そちら に 着き ます (kuukou kara denwa shi tei masu yoshi goro ni wa sochira ni tsuki masu)

Evaluation sentence

- ここ へ は 何 の 用 で 来 ました か (koko e wa nan no you de ki mashi ta ka)
- 船上 の 生活 は どう でし た か (senjou no seikatsu wa dou deshi taka)
- 一 ペニー の 節約 は 一 ペニー の 儲け (1 penii no setsuyaku wa 1 penii no mouke)

A.9 Cluster 8

Training sentence

- E: it's over there just in front of the tourist information
  J: あちら の 旅行 者 案内 所 の 前 です (achira no ryokou sha annai sho no mae desu)
- E: go straight until you see a drugstore
  J: まっすぐ 行く と 薬局 が 見え ます (massugu iku to yakkyoku ga mie masu)
- E: does this bus stop at stoner avenue
  J: この バス は ストーナー街 に 止まり ます か (kono basu wa sutoonaagai ni tomari masu ka)

- E: go left at the third corner
  J: 三つ目の角を左に曲がって下さい (mittu me no kago wo maga ttekudasai)
- E: this car goes to chicago doesn't it
  J: この車両はシカゴ行きですね

Evaluation sentence

- ここに行きたいんですけど (koko ni iki tai n desu kedo)
- どちらへいらっしゃりたいのですか (dochira e irasshari tai no desu ka)
- 市内へ行くバスは何番ですか (shinai e iku basu wa nan ban desu ka)

A.10 Cluster 9

Training sentence

- 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)

Evaluation sentence

- 席はありますか (seki wh ari masu ka)
- 甘口のワインはどれですか (amakuchi no wain wa dore desu ka)
- 一泊いくらですか (i ppaku ikura desu ka)

A.11 General

Evaluation sentance

- 私は昨日確かにこのホテルに電話して予約しました (watashi wa kinou tashika ni kono hoteru ni denwa shi te yoyaku shi mashi ta)
- どういう意味ですか (douyuu imi desu ka)
- ここで口を挟んでもよろしいですか (koko de kuchi wo hasan de mo yoroshii desu ka)

Appendix B: Sample of Different Translation Results

1. Source: 御客様の部屋は二一ゼロ号室です
   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. Source: 魚のたくさんいるポイントはどこですか
   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. Source: デザインが気に入りません
   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. Source: どこに連絡したら良いのでしょうか
   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. Source: 日本語の通じる警察署はどこですか
   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

Hirofumi Yamamoto received the M.S. degree in agriculture from the Tokyo University 1981 and the Ph.D. degree in global information and telecommunication from the Waseda University in 2004. Dr. Yamamoto is currently a researcher at National Institute of Communications Technology. His research interests include speech recognition and machine translation. He is a member of the IEEE, the ASJ and the ANLP.

Eiichiro Sumita received the M.S. degree in computer science from the University of Electro-Communications in 1982 and the Ph.D. degree in engineering from Kyoto University in 1999. Dr. Sumita is NLP department head of ATR/SLC, research manager of NiCT/KCCRC/SLCG, visiting professor of Kobe University and vice-president of ATR-Langue. His research interests include machine translation and e-Learning. He is a member of the IEEE, the ACL, the IPSJ, the ASJ and the ANLP. He serves Associate Editor of ACM/TSLP.