The TCH Machine Translation System for IWSLT 2008

Haifeng Wang, Hua Wu, Xiaoguang Hu, Zhanyi Liu, Jianfeng Li, Dengjun Ren, Zhengyu Niu

Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, Beijing, 100738, China
{wanghaifeng, wuhua, huxiaoguang, liuzhanyi, lijianfeng, rendengjun, niuzhengyu}@rdc.toshiba.com.cn

Abstract

This paper reports on the first participation of TCH (Toshiba (China) Research and Development Center) at the IWSLT evaluation campaign. We participated in all the 5 translation tasks with Chinese as source language or target language. For Chinese-English and English-Chinese translation, we used hybrid systems that combine rule-based machine translation (RBMT) method and statistical machine translation (SMT) method. For Chinese-Spanish translation, phrase-based SMT models were used. For the pivot task, we combined the translations generated by a pivot based statistical translation model and a statistical transfer translation model (firstly, translating from Chinese to English, and then from English to Spanish). Moreover, for better performance of MT, we improved each module in the MT systems as follows: adapting Chinese word segmentation to spoken language translation, selecting out-of-domain corpus to build language models, using bilingual dictionaries to correct word alignment results, handling NE translation and selecting translations from the outputs of multiple systems. According to the automatic evaluation results on the full test sets, we top in all the 5 tasks.

1. Introduction

This paper presents the algorithms and the experimental results of the TCH spoken language translation systems for IWSLT 2008. We participated in all the translation tasks with Chinese as source language or target language, shown as follows:

- Challenge tasks
  - ➢ Chinese-English: CT_CE (SS and CRR)
  - ➢ English-Chinese: CT_EC (SS and CRR)
- Pivot tasks
  - ➢ Chinese-English-Spanish: PIVOT_CES (RS and CRR)
- BTEC tasks
  - ➢ Chinese-English: BTEC_CE (RS and CRR)
  - ➢ Chinese-Spanish: BTEC_CS (RS and CRR)

SS, RS and CRR represent different input conditions, namely spontaneous speech, read speech and correct recognition result, respectively.

For different translation directions, we used different translation strategies. For Chinese-English and English-Chinese translation, we used hybrid systems that combine rule-based machine translation (RBMT) method and statistical machine translation (SMT) method. For Chinese-Spanish translation, phrase-based SMT models were used. For the pivot task, we combined the translations of a pivot based statistical translation model and a statistical transfer translation model (firstly, translating from Chinese to English, and then from English to Spanish).

In addition, we also individually investigated the contribution of each module in MT systems and adapted these modules to spoken language translation. These modules include Chinese word segmentation, word alignment, named entity (NE) translation and language model.

The remainder of this paper is organized as follows. Section 2 describes the core algorithms of our systems for the 5 tasks. Section 3 focuses on the specific methods adapted to spoken language translation. Sections 4 to 7 provide the details of our experiments for each task. Section 8 presents the evaluation results of our primary submissions. Section 9 concludes our work for IWSLT 2008.

2. System description

2.1. SMT system

We used the phrase-based SMT system: Moses [1]. In Moses, phrase translation probabilities, reordering probabilities, and language model probabilities are combined in the log-linear model to obtain the best translation $\mathbf{e}_{\text{best}}$ of the source sentence $\mathbf{f}$:

$$\begin{aligned}\mathbf{e}_{\text{best}} &= \arg\max_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f}) \\ &\approx \arg\max_{\mathbf{e}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e},\mathbf{f})\end{aligned} \tag{1}$$

The models or features which are employed by the decoder consist of (a) one or several phrases tables, (b) one or more language models trained with SRILM toolkit [2], (c) distance-based and lexicalized reordering models, (d) word penalty and (e) phrase penalty. The weights are set by a discriminative training method on a held-out data set [3].

2.2. Combination of RBMT and SMT

We used two MT systems with different translation strategies for Chinese-English and English-Chinese translation. One is a RBMT software - Dr. eye[1]. The other is a phrase-based SMT system - Moses. Firstly we ran the RBMT system as a black box to translate the source texts into the target language. The translations and original source text were used as a synthetic bilingual corpus to train an SMT system. Using the bilingual corpus available for an evaluation task, we built another SMT model. Then these two translation models were combined together as a hybrid system [4].

[1] Available at http://www.dreye.com.cn/prod/cp-pc-download.shtml

In our experiments using the development data for evaluation, we used RBMT system to translate the development data to build the synthetic bilingual corpus. In primary runs at IWSLT 2008, we built the synthetic bilingual corpus by translating the test data using RBMT system. For the pivot task, we also used RBMT system on some training sets, which will be described in Section 7.

2.3. Pivot-based SMT system

For the pivot task Chinese-English-Spanish translation, we built a pivot translation model as described in [5]. Firstly we trained two translation models on the Chinese-English corpus and English-Spanish corpus, and then built a pivot translation model for Chinese-Spanish translation using English as a pivot language. To use a phrase-based translation system such as Moses, we need to obtain a phrase-table for the Chinese-Spanish translation, where two important features are needed: phrase translation probability and lexical weight.

2.3.1. Phrase translation probability

With the Chinese-English and English-Spanish bilingual corpora, we trained two phrase translation probabilities $\phi(\bar{f} \mid \bar{p})$ and $\phi(\bar{p} \mid \bar{e})$, where $\bar{p}$ is the phrase in the pivot language [2]. Given the phrase translation probabilities $\phi(\bar{f} \mid \bar{p})$ and $\phi(\bar{p} \mid \bar{e})$, the phrase translation probability $\phi(\bar{f} \mid \bar{e})$ can be calculated as follows:

$$\begin{aligned}\phi(\bar{f} \mid \bar{e}) &= \sum_{\bar{p}} \phi(\bar{f} \mid \bar{p}, \bar{e}) \phi(\bar{p} \mid \bar{e}) \\ &\approx \sum_{\bar{p}} \phi(\bar{f} \mid \bar{p}) \phi(\bar{p} \mid \bar{e})\end{aligned} \qquad (2)$$

Here, we made an independence assumption: the phrase translation probability $\phi(\bar{f} \mid \bar{p}, \bar{e})$ does not depend on the phase $\bar{e}$ in the target language, since it was estimated from the source-pivot bilingual corpus.

2.3.2. Lexical weight

Given a phrase pair $(\bar{f}, \bar{e})$ and a word alignment a between the source word positions $i = 1,...,n$ and the target word positions $j = 1,...,m$, the lexical weight can be estimated as follows [6].

$$\begin{aligned}& p_{\mathrm{w}}(\bar{f} \mid \bar{e}, a) \\ &= \prod_{i=1}^{n} \frac{1}{|j \mid (i,j) \in a|} \sum_{\forall (i,j) \in a} w(f_i \mid e_j)\end{aligned} \qquad (3)$$

In order to estimate the lexical weight, we first need to obtain the alignment information a between the two phrases $\bar{f}$ and $\bar{e}$, and then estimate the lexical translation probability $w(f \mid e)$ according to the alignment information.

Let a_1 and a_2 represent the word alignment information inside the phrase pairs $(\bar{f}, \bar{p})$ and $(\bar{p}, \bar{e})$ respectively, then the alignment information a inside $(\bar{f}, \bar{e})$ can be obtained as shown in (4).

$$a = \{(f,e) \mid \exists p : (f,p) \in a_1 \;\&\; (p,e) \in a_2\} \qquad (4)$$

We estimated the lexical translation probability directly from the induced phrase pairs using the induced alignment information. Let *K* denote the number of the induced phrase pairs. We estimated the co-occurring frequency of the word pair (f, e) according to the following equation.

$$\begin{aligned}& count(f,e) \\ &= \sum_{k=1}^{K} \phi_k(\bar{f} \mid \bar{e}) \sum_{i=1}^{n} \delta(f, f_i) \delta(e, e_{a_i})\end{aligned} \qquad (5)$$

Where $\phi_k(\bar{f} \mid \bar{e})$ is the phrase translation probability for phrase pair k. $\delta(x,y) = 1$ if $x = y$; otherwise, $\delta(x,y) = 0$. Thus, the lexical translation probability can be estimated as in (6).

$$w(f \mid e) = \frac{count(f,e)}{\sum_{f'} count(f',e)} \qquad (6)$$

3. Methods

In this section, we describe the adaption of different modules in MT systems to spoken language translation tasks, including Chinese word segmentation, translation dictionary extraction, word alignment, NE translation, language model, punctuation restoration, case restoration, and translation selection.

3.1. Chinese word segmentation

Currently, most of Chinese word segmentation systems are not designed for spoken language translation. Thus, we investigated the effect of segmentation granularity and segmentation dictionary for better MT performance on spoken language.

3.1.1. Word segmentation dictionary

A Chinese dictionary is a fundamental element for word segmentation. In our segmenter, we used three kinds of dictionaries: basic dictionary, NE dictionary and in-domain dictionary.

- The basic dictionary contains some commonly-used words.
- The NE dictionary consists of transliterated person names, Japanese first names and last names[3], and location names. They were extracted from the Chinese-English Name Entity Lists Version 1.0 (LDC2005T34).
- The in-domain dictionary contains domain-specific words. These words were extracted from the in-domain corpora such as the Basic Travel Expression Corpus (BTEC) and the HIT Olympic corpus (2004-863-008)[4].

[2] We use **f**, **e** and **p** to represent the source language, the target language, and the pivot language, respectively.

[3] We added the Japanese names because they cannot be transliterated.

[4] Available at http://www.chineseldc.org/EN/purchasing.htm

3.1.2. Segmentation algorithm

According to our initial experiments, only a few segmentation ambiguities occur in Chinese spoken language in travel domain. Therefore, for this domain, we built an ambiguous fragments database in two steps. Firstly we identify the ambiguous fragments from the texts in this domain by using Forward Maximum-Matching (FMM) and Back One Character (BOC) methods [7]. Then we annotate the correct segmentations for these ambiguous fragments. For example, given a Chinese string "有空席". With the FMM method, "有空" was segmented as a word. Then the segmentation should be continued from the Chinese character "席". While with the BOC method, we continue the segmentation from the previous character of "席", i.e. "空". Then "空席" is also indentified as a candidate word. So "有空席" is indentified as a ambiguous fragment . Then we put the correct segmentation "有 空席" into the ambiguous fragments database.

In our experiments, we used the FMM and BOC to segment a Chinese sentence and used ambiguous fragments database to resolve the ambiguities.

3.1.3. Word granularity

There are lots of arguments about the definition of a Chinese word. In fact, only a few researchers investigated the effect of word granularity on machine translation [8]. In this work, we followed the guidelines shown below to define what is a word.

- Its translation is a word in the target language.
- Its translation is a multi-word expression or frequently used phrase in target language. For example:
  - ➢ 炸薯条: French fries
  - ➢ 失物招领处: lost and found

3.1.4. Word normalization

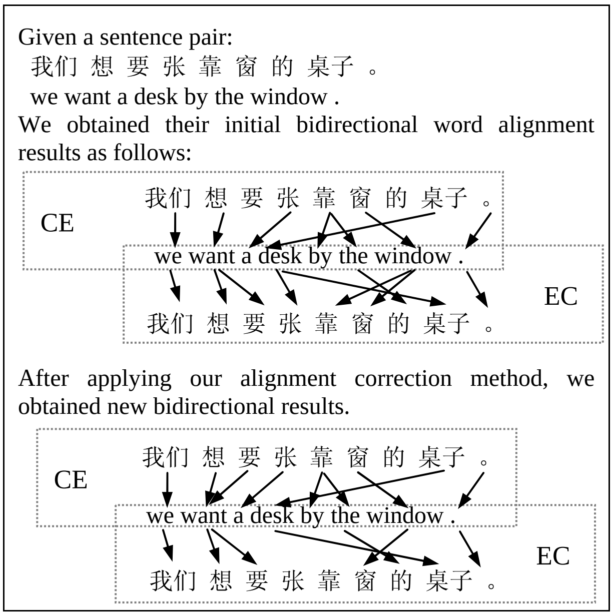
To deal with data sparseness problem due to the small size of training data, we normalized the paraphrases of the same concepts into the same words if the paraphrases in source language are unambiguous and have the same translations in target language.

For example, the Chinese words "打火机" and "火机" are translated into the same English word "lighter". Therefore, in both training data and test data, the two words were replaced with "火机".

This word normalization was only executed in the tasks where Chinese is the source language. For tasks with Chinese as target language, we did not conduct this preprocessing.

3.2. Translation dictionary

The translation dictionaries used in this paper include a general-domain dictionary (Chinese-English Translation Lexicon Version 3.0 (LDC2002L27)), an NE dictionary (Chinese-English Name Entity Lists Version 1.0 (LDC2005T34)), and a translation dictionary extracted from the training corpora[5]. For the NE dictionary, we only kept the person names and location names that are unambiguous and consistent with the word segmentation dictionary.



Given a sentence pair:
我们 想 要 张 靠 窗 的 桌子 。
we want a desk by the window .
We obtained their initial bidirectional word alignment results as follows:

CE
我们 想 要 张 靠 窗 的 桌子 。
we want a desk by the window .
EC
我们 想 要 张 靠 窗 的 桌子 。

After applying our alignment correction method, we obtained new bidirectional results.

CE
我们 想 要 张 靠 窗 的 桌子 。
we want a desk by the window .
EC
我们 想 要 张 靠 窗 的 桌子 。

Figure 1. An example of improving word alignment results

For covering domain-specific word or phrase translations, we extracted a translation dictionary from an in-domain training corpus as an additional translation dictionary. Firstly we obtained word translation pairs from bidirectional word alignments generated by the GIZA++ toolkits[6]. Then to filter the noisy translation pairs caused by the incorrect alignment links, we automatically removed the translation pairs with low translation probabilities or low co-occurring frequencies, and then manually checked the extracted dictionaries to remove incorrect translation pairs.

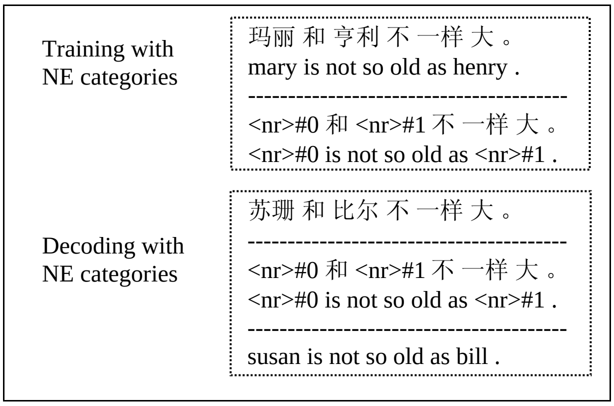
3.3. Word alignment

In order to obtain a phrase table from training data, we first ran the GIZA++ toolkits to obtain a baseline word alignment result. In this baseline word alignment result, there were many alignment errors, which resulted in a noisy phrase table. Here we used translation dictionaries to improve the alignment result. The translation dictionaries include a domain specific dictionary extracted from the training data and/or a publicly available general dictionary such as the LDC Chinese-English Translation Lexicon Version 3.0 (LDC2002L27). We improved the word alignment result as follows:

- We ran the GIZA++ toolkits to obtain bidirectional word alignment results.
- We kept the links in the intersection set of the bidirectional word alignment results.
- For those alignment links occurring in bilingual dictionaries, we added them into the final alignment set.
- For the links conflicting with the links in the final alignment set, we simply deleted them.
- For the remained links, we kept them in the bidirectional results.

[5] For the tasks using Spanish as the target language, we only used in-domain dictionaries extracted from training data.

[6] Available at http://www.fjoch.com/GIZA++.html



Training with NE categories

玛丽 和 亨利 不 一样 大 。

mary is not so old as henry .

<nr>#0 和 <nr>#1 不 一样 大 。

<nr>#0 is not so old as <nr>#1 .

Decoding with NE categories

苏珊 和 比尔 不 一样 大 。

<nr>#0 和 <nr>#1 不 一样 大 。

<nr>#0 is not so old as <nr>#1 .

----

susan is not so old as bill .

Figure 2. An example of NE translation

- Finally we selected an alignment heuristics to get symmetrized alignment result based on the corrected bidirectional alignment results.

Figure 1 shows an example on how to improve the word alignment result. In this example, since (我们, we) and (。,.) are alignment links in the intersection sets, we kept them. Then we selected the alignment links (想 要, want), (窗, window) and (桌子, desk) as final links according to the bilingual dictionary. Since the alignment links (想, want), (靠, window) and (张, desk) conflicted with the selected links, we deleted them. Finally, we kept the remained links and obtained the improved bidirectional alignment result.

3.4. Named entity translation

For named entities such as digits, dates, times, person names and location names, it is difficult to translate them with a SMT method since some of them never appear in training data. So we used some hand-crafted rules to recognize and translate digits, dates and times. For person names and location names, we replaced them in the bilingual training data with specific NE tags. Then, we trained translation models on the data with these NE tags. Given a test text in the source language, we replaced the NEs in the test text with NE tags and translated the text into the target sentence. Finally, we generated the translations for the NE tags in the target text using a translation dictionary and rules. Figure 2 shows an example.

3.5. Language model

Although the BTEC corpus is an in-domain corpus, its size is quite small. Thus we cannot get a well-trained language model on it. Therefore we adopted a selection-and-interpolation method to build language models (LM) by leveraging in-domain data and a large amount of out-of-domain data.

Firstly, a baseline LM was built on the BTEC training corpus. Then we used the baseline LM to select relevant sentences from the out-of-domain corpora. Relevance was measured by perplexity, shown as follows:

$$PPL = 2^{-\frac{1}{T}\sum_{i=1}^{T}\log_2 p(w_i|w_{i-n+1}\cdots w_{i-1})} \qquad (7)$$

Where T is the word number of each out-of-domain sentence, and the conditional probabilities were calculated according to the baseline LM. Smaller perplexity means that the sentence is more similar to the in-domain corpus. Sentences with small perplexity were selected to build an out-of-domain LM.

LMs trained on different data sets (e.g., in-domain data set and a few other out-of-domain data sets) were integrated into one LM by linear interpolation using the SRILM toolkit. The interpolation weights of input models were tuned on development sets.

3.6. Punctuation restoration

For recovering the punctuations in English ASR input, we used the *hidden-ngram* tool in SRILM toolkit to insert four possible English punctuations in the English text, including comma, period, question mark, and exclamation mark. The language model used by *hidden-ngram* was trained on the training data. After that, a few hand-crafted rules were employed to revise the inserted punctuations. For example, if the sentence begins with "can", "would", "what", etc., then the ending period will be replaced by a question mark.

For Chinese ASR input, we used a maximum entropy (ME) model (a modified version of the *Maxent* tool in the OpenNLP Toolkit[7]) to recover punctuations since our initial study indicated that *Maxent* outperformed *hidden-ngram* on the Chinese text. The process for punctuation restoration included two steps. First, to determine the position to split a sentence into sub-sentences, we used 1-gram to 6-gram features. Specifically, three proceeding and three following words around each word boundary are involved. Secondly, to determine which punctuation should be inserted at the boundary of the sub-sentence, in addition to the features to determine sub-sentence boundary, we also used boundary word features. A boundary word is the word at the beginning or the end of a sub-sentence.

3.7. Case restoration

For case restoration, we used the tool *recaser* available in the training scripts of Moses to perform recasing on western language text (Spanish or English text). That tool dealt with the case restoration problem as a machine translation problem and then used Moses to do case restoration. For each task with Spanish or English as target language, we trained a recasing tool respectively using the training text with case information in the task. Moreover, for English, we looked up a lexicon to identify words or phrases that should be capitalized from the text to be recased. That lexicon includes named entities obtained from publicly available resources (e.g., training text in respective tasks, Tanaka corpus, and HIT corpus).

3.8. Translation selection

To select final translation from several translation systems, we employed two methods. One is to use a 5-gram LM for selection [9]. The other is to employ the information of target sentence average length (TSAL) calculated on the BTEC training corpus. TSAL is the average length of the target sentence for a source sentence with a given length. The translation with the length closest to the calculated TSAL was selected as the final translation. On different tasks, we selected one of them for translation selection that obtained better translation results on the development sets.

[7] Available at http://opennlp.sourceforge.net/projects.html

Corpus	Sentence pairs	Source words	Target words
BTEC	19,972	177,168	182,627
HIT	80,868	802,454	822,508
CLDC	200,732	2,113,534	2,096,731
Tanaka	149,207	-	1,351,645

Table 1. Statistics of training data in CE task

4. Chinese to English system

We developed the same Chinese to English (CE) translation system for both the BTEC CE task and Challenge CE task. Although there are differences between the CRR and automatic speech recognition (ASR) inputs, we mainly tuned our system under the CRR input of the BTEC task.

4.1. Data

The resources include bilingual dictionaries, bilingual training data, monolingual training data, development sets and test sets. The bilingual dictionaries include the Chinese-English Translation Lexicon Version 3.0 (LDC2002L27), an in-domain translation dictionary extracted from the training corpus and an NE dictionary extracted from Chinese-English Name Entity Lists Version 1.0 (LDC2005T34). The LDC translation lexicon contains 54,170 entries, and the in-domain dictionary contains 38,620 entries. For the NE list, we only kept the person names and location names that are not ambiguous, resulting in 47,692 entries. The LDC translation lexicon and the in-domain dictionary were used to correct word alignments. And the NE dictionary was used to recognize and translate person names and location names.

The detailed statistics of the training data are showed in Table 1. For the bilingual corpus, besides the BTEC data, we also used the HIT corpus[8] and other Chinese LDC (CLDC) corpora, including the Chinese-English Sentence Aligned Bilingual Corpus (CLDC-LAC-2003-004) and the Chinese-English/Chinese-Japanese Parallel Corpora (CLDC-LAC-2003-006). From the HIT corpus and the CLDC corpora, we selected some in-domain sentence pairs which are close to those in the BTEC corpus. For the Chinese sentence, the full-width case was converted to the half-width case and the digital string was converted to the textual form. Then the Chinese sentences were segmented. English sentences were preprocessed with tools provided by the IWSLT 2008 organizers and the words were lowercased. We used the English part of BTEC corpus, HIT corpus and Tanaka corpus as monolingual corpora to train the language model and the case restoration model.

We used devset1, devset2 and devset4 of IWSLT 2008 evaluation as development sets and used the evaluation sets of 2005 (devset3), 2006 (devset5) and 2007 (devset6) as test sets.

4.2. Results

The case sensitive BLEU score[9] was used to evaluate the CRR translation quality. Table 2 shows the experimental results. We implemented the minimum error rate training (MERT) method to train a group of weights which would be used in the following experiments. The performance of the RBMT system (Dr. eye) is showed in row "RBMT". Firstly we implemented a baseline translation model (TM) trained on BTEC corpus with original word segmentation using *grow-diag* heuristics. The baseline 5-gram language model (LM1) was trained with the English parts of the BTEC, HIT and Tanaka corpus by using interpolated Kneser-Ney smoothing. The row "Baseline" shows the results of the baseline system.

	devset3	devset5	devset6
RBMT	0.4253	0.2020	0.2086
Baseline	0.5186	0.2013	0.2807
Our segmenter	0.5425	0.2047	0.3029
+HIT	0.5697	0.2323	0.3416
+Dic	0.5819	0.2375	0.3456
+NE	0.5838	0.2396	0.3537
+CLDC	0.5891	0.2445	0.3554
+RBMT	0.6091	0.2536	0.3570
+LM Inter.	0.6223	0.2516	0.3823

Table 2. Results of the CE on development sets

Motivated by the analysis in Section 3.1, we preprocessed the BTEC corpus and test data with our word segmenter. The row with "Our segmenter" in Table 2 shows the performance of baseline system with updated training data. For investigation of the contribution of more training data, the HIT corpus was added into the bilingual corpus to train a new translation model (TM1). The row with "+HIT" in Table 2 shows the performance.

The bilingual lexicon was used as a clue to correct word alignments for translation quality improvement. Results are in the row with "+Dic". Rules and the NE dictionary were used to translate the NEs. The row with "+NE" shows the results.

In order to cover infrequently occurring words, we trained another translation model (TM2) with the CLDC corpora. Since TM2 is not as good as TM1, we punished the phrase pairs from TM2. The phrase pairs in TM2 were used only when their source phrases did not occur in TM1 and their translation probabilities were multiplied by a small coefficient (0.1). The row with "+CLDC" shows the results of system using TM1 and TM2. Furthermore, a rule-based system was used to translate the test sets. The translated test sets were combined with the BTEC corpus to train a new translation model (TM3). Since the RBMT translation results contain some errors, the translation probabilities were also punished by multiplying a small coefficient (0.1). The results of the MT system using TM1, TM2 and TM3 are shown in the row with "+RBMT".

After that, we investigated a new method for language model training. Instead of merging all corpora in LM1, we trained three language models by using the BTEC, HIT and Tanaka corpus respectively. Then, these three language models were interpolated to generate a new language model with weights 0.6, 0.2 and 0.2. The row with "+LM Inter." in Table 2 shows the performance of MT system with the updated language model.

Finally, we tried to find better weights of the log-linear models with the MERT method for the system "+LM Inter.". Since it is very difficult to balance the weights of different

[8] In this paper, the HIT corpus contains the CLDC Olympic corpus (2004-863-008) and the other HIT corpora available at http://mitlab.hit.edu.cn/index.php/resources/29-the-resource/111-share-bilingual-corpus.html.

[9] Available at https://www.slc.atr.jp/Corpus/IWSLT08/eval/IWSLT08_auto_eval.tgz

	devset3	devset5	devset6
Default	0.5927	0.2547	0.3453
Mert1	0.6061	0.2679	0.3837
Mert2	0.6274	0.2551	0.3863
Select	0.6260	0.2627	0.3882

Table 3. Results of translation selection

models, three different groups of weights were kept, as shown in Table 3. "Default" means that we used the default parameters in Moses. We kept the parameters "Mert1" because we obtained higher performance on devset5 with this group of parameters. We used the parameters in "Mert2" since we obtained satisfactory results on all of the three devsets[10]. Every input sentence was translated into three different target sentences according to the three groups of weights. A simple method was used to select the final translation. If there are two identical sentences, the identical sentence was chosen. Otherwise, we chose the sentence that has the closest length to the source sentence according to the method described in Section 3.8.

5. English to Chinese system

For the English to Chinese (EC) challenge task with SS results or CRR texts as input, we used the same system that combines three translation models trained on different data sets. The parameters for this system were tuned on development set with CRR texts as input.

5.1. Data

Table 4 shows the statistics of the training data in the EC system. The training data consists of the BTEC data provided for this task and the HIT corpus released for IWSLT 2008 as the additional training data. The development sets include the devset3 and devset.

In addition to the BTEC data, 89,318 examples that are similar to the BTEC data were extracted from the HIT corpus. We preprocessed training data as same as that done in the Section 4.1, except that (1) the unambiguous abbreviations in the English sentence were restored here, for example, "I'm" was restored to "I am"; (2) Chinese word normalization in the segmenter was not used for this task.

We reversed the bilingual dictionaries in the Chinese to English task and used it to improve the word alignments and to translate the person names and location names.

5.2. Results

The experimental results with CRR inputs on the devset3 and devset are shown in Table 5. In our experiments, we used *grow* as the heuristic of word alignments, and used the target part of the training data to build a 5-gram language model with interpolated Kneser-Ney smoothing.

The first row "RBMT" in Table 5 shows the results of the RBMT system on the development sets. The row "Baseline" shows the results of the SMT system with translation model and language model trained on the BTEC data, where we used the word-segmented Chinese sentences provided by the organizers. In the results of the row "Our segmenter", the SMT system used the same data as "Baseline", except that the Chinese sentences were preprocessed by our Chinese word segmenter (Section 3.1).

Corpus	Sentence pairs	Source words	Target words
BTEC	19,972	189,041	178,339
HIT	89,318	945,010	914,121

Table 4. Description of training data in the challenge EC task

	devset3	devset
RBMT	0.4362	0.4425
Baseline	0.4455	0.4511
Our segmenter	0.4528	0.4564
+Dic	0.4551	0.4684
+NE	0.4558	0.4773
+HIT	0.4830	0.5325
+RBMT	0.5131	0.5426
+Select	0.5133	0.5551

Table 5. Results of the challenge EC task on development sets

"+Dic" in the fourth row means that the bilingual dictionaries were used to improve word alignment as described in Section 3.3. This translation model is denoted as TM1.

Then, the NE translation strategy was adopted to translate the digits, dates, times, person names and location names in the development sets. The digits, times and dates were translated by using the rules, and person names and location names were translated with the bilingual dictionaries. From the results in the row with "+NE", it can be seen that the translation quality on devset was significantly improved from 0.4684 to 0.4773.

After that, we trained another translation model (TM2) on the selected HIT corpus. Then TM1 and TM2 were interpolated with weights 0.7 and 0.3 to generate a new translation model (TM12). At the same time, the language models trained from the BTEC data and the HIT corpus were also interpolated with weights 0.4 and 0.6. Results in the row "+HIT" indicate that the TM12 and interpolated language model significantly improved the translation quality.

In order to provide more translation candidates, the rule-based system was employed to translate the development sets. The development sets and their translations were used as a synthetic bilingual corpus. The translation model (TM3) trained on the synthetic bilingual corpus and TM12 were interpolated with weights 0.1 and 0.9 to generate a new translation model (TM123). The row with "+RBMT" shows that the translation qualities were improved by about 3 and 1 BLEU scores on devset3 and devset, respectively.

Finally, the translation selection method based on the language model probability described in Section 3.8 was adopted to select a better translation from the two SMT systems: TM123 and TM12. The results of the row "+Select" shows that the translation selection method achieved an improvement of 1.25 BLEU scores on devset.

In the EC system, the weights were not optimized using the MERT method because MERT did not improve the performance of our MT system on the two development sets.

[10] All of the systems in Table 2 used a group of parameters trained on the system "+HIT". Thus, the result of "+LM Inter." is different from those in Table 3.

	Baseline	Our segmenter	+Dic
BLEU	0.3596	0.3726	0.3839

Table 6. Case sensitive results for the BTEC CS task

6. Chinese to Spanish system

6.1. Data

For the Chinese to Spanish (CS) BTEC task, we only used the BTEC data provided for this task as training data, which contains 19,972 sentence pairs. The preprocessing steps for Chinese sentences were as same as those in Section 4.1. For Spanish sentences, we used the tools provided by the IWSLT 2008 organizers for tokenization. In addition, we extracted a CS translation dictionary from the BTEC corpus using the method described in Section 3.2, which contains 9,990 entries.

For language model training, in addition to the target language part of the BTEC corpus, we also treated Spanish translation for each entry in the translation dictionary as target sentences. We combine them to train a 5-gram language model with interpolated Kneser-Ney smoothing and without count cutoff. The recaser model was trained on the Spanish language part of BTEC corpus.

We used the development set (devset3) as test set since there is only one development set in this task. Moreover, all parameters in the log-linear translation model were not optimized using MERT[11].

6.2. Results

Table 6 shows our experimental results of the CRR input. We used the heuristic *grow-diag* for word alignment in all the MT systems here. The baseline system used the BTEC corpus with default Chinese word segmentation provided by IWSLT 2008 organizers for translation model training and language model training. The column with "Baseline" shows the performance of this baseline system. Motivated by the analysis in section 3.1, we built another MT system trained on the BTEC corpus processed by our word segmenter. The column with "Our segmenter" shows the performance of this system, which is better than that of the baseline system with 1.3 BLEU scores.

Then we added the translation dictionary to the training corpus and used this dictionary to correct word alignment results, which resulted in the third system. The performance of this system is shown in the column with "+Dic" in Table 6. We can see that using the translation dictionary resulted in a significant improvement of about 1 BLEU score. And for the translations in Spanish language, we used hand-crafted rules to revise some punctuation marks. For examples, if the end of the sentences contains a question mark "?", we added the "¿" at the beginning of the sentence if it is missing.

7. Chinese-English-Spanish system

7.1. Data

Table 7 describes the data used for model training in this pivot task, including the Chinese-English (CE) corpus and the English-Spanish (ES) corpus provided by IWSLT 2008 organizers, the HIT corpus, the Europarl corpus and the Tanaka corpus.

Corpus	Sentence pairs	Source words	Target words
BTEC CE	20,000	164,957	182,793
HIT CE	80,868	802,454	822,508
BTEC ES	19,972	182,627	185,527
Europarl ES	400,000	8,485,253	8,219,380
Tanaka	149,207	-	1,351,645

Table 7. Description of training data in the pivot task

For Chinese-English translation, we selected some sentence pairs from the HIT corpus close to the BTEC CE corpus. Then we used the English parts of the BTEC CE corpus and the selected HIT corpus, and the Tanaka corpus to train a 5-gram English language model with interpolated Kneser-Ney smoothing. Moreover, we used the LDC Translation Lexicon Version 3.0 and an in-domain translation dictionary extracted from the BTEC CE corpus and the HIT CE corpus in Table 7. The in-domain dictionary contains 39,010 entries.

For English-Spanish translation, we used the method described in Section 3.5 to select sentences from the Europarl corpus[12] that are close to the English parts of both the BTEC CE corpus and the BTEC ES corpus. Finally, we selected 400k sentence pairs. Then we interpolated an out-of-domain LM trained on the Spanish part of this selected corpus with the in-domain LM trained with the BTEC corpus. The interpolation weights were set to 0.8 and 0.2 for the in-domain LM and out-of-domain LM, respectively. Moreover, we extracted a dictionary from the BTEC ES corpus and the Europarl ES corpus. This dictionary contains 10,426 entries.

For Chinese-English-Spanish translation, we used the development set (devset3) as test set since there is only one development set in this task. All parameters in the log-linear translation model were not optimized using MERT.

7.2. Results

For the pivot tasks, we used two methods. One is to train a pivot model as described in Section 2.3. The other is to use the transfer method: first translating from Chinese to English, and then from English to Spanish.

Table 8 shows the translation results. The baseline system only used the BTEC corpus for both translation model and language model training. The alignment heuristics *grow-diag* and *grow-diag-final* were used for the CE and ES models.

Then, for CE translation model, we took HIT corpus as supplemental training data, and used the CE dictionary for word alignment improvement. For ES translation model, we incorporated the ES translation dictionary into the training data. Then we added the selected Europarl corpus for language model training as described in Section 7.1. Results in the column with "+Dic+HIT+Europarl" indicate that the translation quality was greatly improved by more than 8 BLEU scores for both the pivot model and the transfer model.

After that, we used the RBMT system to translate the English part of the English-Spanish corpus into Chinese to obtain a synthetic corpus. And the synthetic corpus was added

[11] When we applied the systems to the official test sets of IWSLT 2008, we optimized the parameters on the development sets for tasks with Spanish as the target language.

[12] Available at http://www.statmt.org/europarl/

	Baseline	+Dic+HIT + Europarl	+ RBMT
Pivot model	0.2791	0.3616	0.4136
Transfer model	0.3243	0.4139	0.4423
Trans. selection	-	-	0.4510

Table 8. Results of the pivot task

		(Bleu + Meteor) /2	Bleu	Meteor	Human Eval.
CT_EC	SS	0.5647	0.4818	0.6476	0.3906
	CRR	0.6566	0.5912	0.7219	-
CT_CE	SS	0.5257	0.4166	0.6347	0.4516
	CRR	0.5909	0.4980	0.6837	-
BTEC_ CE	RS	0.5358	0.4474	0.6241	0.4730
	CRR	0.5887	0.5085	0.6688	-
BTEC_ CS	RS	0.3273	0.3218	0.3328	0.4316
	CRR	0.3597	0.3582	0.3611	-
PIVOT _CES	RS	0.3620	0.3657	0.3583	0.4624
	CRR	0.4044	0.4157	0.3931	-

Table 9. Primary evaluation results of IWSLT 2008

to train the CE model. As shown in the column "+RBMT", this model further improved the translation quality by about 5 BLEU scores for the pivot model and about 3 BLEU scores for the transfer model.

Results in Table 8 indicate that the transfer models outperformed the pivot models. The reasons are that (1) the CE model can produce a good English translation for the Chinese input. On this development set, we obtained a BLEU score of 0.6024 for the English translation; (2) the languages English and Spanish are more similar than Chinese and Spanish. The ES model can provide better translation even if the input contains some errors; (3) the pivot model contains much more noise than the transfer model.

To further improve the MT performance, we used the TSAL method described in Section 3.8 to select translations from the outputs of the pivot model and the transfer model, which resulted in about 0.9 BLEU score improvement as shown in Table 8.

8. Primary runs

Table 9 provides both automatic and human evaluation results released by IWSLT 2008 organizers for our primary runs on the 5 tasks that we participated in. The automatic evaluation metrics are BLEU and METEOR[13]. Based on the automatic evaluation results on the full test sets, our systems were ranked the first in all these 5 tasks. For human evaluation, the primary runs of each track were given to at least 3 native-speakers of the target language who were asked to rank translations using the criteria described in [10]. This method ranked each whole sentence translation from best to worst relative to the other choices. The ranking scores are the average number of times that a system was judged better than any other system. Our systems for EC, CS and CES tracks were ranked the best, and those for CE were ranked the second.

9. Conclusions

According to our experimental results and the evaluation results, we draw the following conclusions:

- Adaptation of Chinese word segmentation system to spoken language domain significantly improves the translation quality.
- Refinement of word alignment result by the use of translation dictionary helps improving translation quality.
- Separately handling NE (e.g., person name, location name, date, time, digit) translation results in translation quality improvement.
- The incorporation of additional corpus for language model training helps improving translation quality.
- The hybrid method that combines both the RBMT model and SMT model helps improving translation quality.
- Ensemble of multiple translation models can generate better translations.

10. References

[1] Koehn, P., Hoang, H., Birch, A. Callison-Burch, C. Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL-2007, demonstration session*, pages 177-180.

[2] Stolcke, A. 2002. SRILM -- an Extensible Language Modeling Toolkit. In *Proc. of ICSLP-2002*, pages 901-904.

[3] Och, F. J. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL-2003*, pages 160-167.

[4] Hu, X., Wang, H. and Wu, H. 2007. Using RBMT Systems to Produce Bilingual Corpus for SMT. In *Proc. of EMNLP-CoNLL 2007*, pages 287-295.

[5] Wu, H. and Wang, H. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proc. of ACL-2007*, pages 856-863.

[6] Koehn, P., Och, F. J. and Marcu, D. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT-NAACL 2003*, pages 127-133.

[7] Luo, Z. and Song, R. 2006. Disambiguation in a Modern Chinese General-Purpose Word Segmentation System. *Journal of Computer Research and Development*, 43(6): 1122-1128.

[8] Chang, P., Galley, M. and Manning, C. D. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proc. of the Third Workshop on Statistical Machine Translation*, pages 224-232.

[9] Chen, B., Cattoni, R., Bertoldi, N., Cettolo, M., and Federico, M. 2006. The ITC-irst SMT System for IWSLT 2006. In *Proc. of IWSLT-2006*, pages 53-58.

[10] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. 2007. (Meta-) Evaluation of Machine Translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 136-158.

[13] Available at http://www.cs.cmu.edu/~alavie/METEOR