Statistical Machine Translation without Long Parallel Sentences for Training Data

Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara

Department of Information and Knowledge Engineering Faculty of Engineering
Tottori University, Japan
murakami@ike.tottori-u.ac.jp

Abstract

In this study, we paid attention to the reliability of phrase table. We have been used the phrase table using Och's method[2]. And this method sometimes generate completely wrong phrase tables. We found that such phrase table caused by long parallel sentences. Therefore, we removed these long parallel sentences from training data. Also, we utilized general tools for statistical machine translation, such as "Giza++"[3], "moses"[4], and "training-phrase-model.perl"[5].

We obtained a BLEU score of 0.4047 (TEXT) and 0.3553(1-BEST) of the Challenge-EC task for our proposed method. On the other hand, we obtained a BLEU score of 0.3975(TEXT) and 0.3482(1-BEST) of the Challenge-EC task for a standard method. This means that our proposed method was effective for the Challenge-EC task. However, it was not effective for the BTECT-CE and Challenge-CE tasks. And our system was not good performance. For example, our system was the 7th place among 8 system for Challenge-EC task.

1. Introduction

Many machine translation systems have been studied for long time and there are three generations of this technology. The first generation was a rule-based translation method and the second generation was an example-based machine translation method. Recently, statistical machine translation method has been very popular. This method is based on statistics.

There are many versions of statistical machine translation models available. An early model of statistical machine translation was based on IBM1 $\sim$ 5[1]. This model is based on individual words, and thus a "null word" model is needed. However, this "null word" model sometimes has very serious problems, especially in decoding. Thus, recent statistical machine translation systems usually use phrase based models.

By the way, two points are used to evaluate English sentences, one is accuracy, and the other is fluency. We believe accuracy is related to translation model $P(English/Chinese)$ and fluency is related to language model $P(English)$. Therefore, long phrase tables are needed for high accuracy. Similar languages like English and German may only require short phrases for accurate translations. However, languages that differ greatly, like Chinese and English, require long phrases for accurate translation. We implemented our statistical machine translation model using long phrase tables, that is similar to a statistical example-based translation system.

Also, we found long parallel sentences for training parallel data are easily result into wrong phrase table, and wrong phrase table makes poor translation results especially for the accuracy. Therefore we removed long parallel sentences from the training parallel data. We used 19972 Chinese-English parallel sentences for BTEC-CE and Challenge-CE task. In the Challenge-EC task, we removed for greater than 48 characters Chinese sentence for training data. so, we used 19387 Chinese-English parallel sentences.

On the other hand, N-gram model is used as a language model for fluency. And, in general, when we use a higher order N-gram, the number of N-gram parameters dramatically increase and the reliability of parameter decreases. So, we chose a 4-gram model. This model was the best language model among N-gram from our experiments at the previous 2007 International Workshop on Spoken Language Translation (IWSLT2007) contest.

We used general tools for statistic machine translation, such as "Giza++"GIZA++, "moses"[4], and "training-phrase-model.perl"[5]. We used these data and these tools, participated in the contest of BTEC-CE, Challenge-CE and Challenge-EC at IWSLT2008. And proposed method was effective for the Challenge-EC task. However, it was not effective for the BTEC-CE and Challenge-CE tasks.

2. Concepts of our Statistical Machine Translation System

In this section, we will describe our concepts behind our Chinese English statistical machine translation system.

2.1. Standard Tools

Many statistical machine translation tools have been developed. These tools have been highly reliable and widely used. So whenever possible we did not make special tools, but instead relied on the following established tools.

1. GIZA++.2003-09-30.tar.gz [3]

2. moses.2007-05-29.tgz [4]
3. training-release-1.3.tgz(train-phrase-model.perl) [5]

We made only a small number of minor tools for building a temporal corpus.

2.2. Long Phrase Tables (Accuracy)

We have been evaluated English translated sentences both the accuracy and the fluency. We believe that accuracy is related to translation model $P(English/Chinese)$. Thus, we made long phrase tables to achieve higher accuracy. In similar languages like English and German, the difference in word position is small. In such a case, short phrase tables poses little problem. However, in Chinese to English translation, verbs are sometimes moved from their original position. Therefore, we needed to make long phrase tables.

2.3. 4-gram Language Model (Fluency)

We have been evaluated English sentences on two points; accuracy and fluency. We believe that fluency is related to language model $P(English)$. Thus we used a normal 4-gram model and did not use a higher N-gram model. In general, when we used a higher order N-gram, the number of parameters dramatically increases, and the reliability for each parameter decreases. Therefore we chose a 4-gram model. This model is the best language model among N-gram from our previous results at the IWSLT2007 contest.

3. Experiments with Statistical Machine Translation

3.1. Removed long parallel sentences

We used only the IWSLT2008 training corpus. (Chinese-English parallel sentences). So, we used 19972 Chinese-English parallel sentences for the BTEC-CE, the Challenge-CE, and the Challenge-EC task. We refer to this experiments as ”primary”.

On the other hand, in the BTEC-CE and the Challenge-CE task, we removed more than 48 characters Chinese sentences for training parallel data. So, we used 19327 Chinese-English parallel sentences. Also, in the Challenge-EC task, we removed more than 96 character English sentences for training parallel data. So, we used 19387 English-Chinese parallel sentences. We refer to this experiments as ”contrast”.

3.2. Punctuation procedure

We used the English punctuation procedure, it means that we changed ”,” and ”.” to ” , ” and ” . ”. Also, we did not handle English case for BTEC-CE and Challenge-CE task. The table 1 show the Chinese and English training parallel data for the BTEC-CE and the Challenge-CE.

And we used only the lower case in English for the Challenge-EC task. Table 2 shows the Chinese and English training parallel data for the Challenge-EC.

Table 1: BTEC-CE , Challenge-CE training-data

C	1	在 下面。我 就 拿一些。如果 有什需要的 告我。
C	2	不用 担心 那个。 我 要 它 不需要 把 它 包 起来。
C	3	可以 改改 ？
C	4	灯 是 的。
C	5	我 想 要 靠 窗 的 子。
E	1	It’s just down the hall . I’ll bring you some now . If there is anything else you need , just let me know .
E	2	No worry about that . I’ll take it and you need not wrap it up .
E	3	Do you do alterations ?
E	4	The light was red .
E	5	We want to have a table near the window .

Table 2: Challenge-EC training-data

E	1	it’s just down the hall i’ll bring you some now if there is anything else you need just let me know
E	2	no worry about that i’ll take it and you need not wrap it up
E	3	do you do alterations
E	4	the light was red
E	5	we want to have a table near the window
C	1	在 下面。我 就 拿一些。如果 有什需要的 告我。
C	2	不用 担心 那个。 我 要 它 不需要 把 它 包 起来。
C	3	可以 改改 ？
C	4	灯 是 的。
C	5	我 想 要 靠 窗 的 子。

3.3. Phrase Tables

We used the “train-phrase-model.perl[5]” in “training-release-1.3.tgz”. We set the parameter of max-phrase-length to 20 to obtain long phrase tables. Other parameters were set to defaults values. Table 10 shows examples of phrase tables for the BTEC-CE task.

3.4. 4-gram language model

We calculated the 4-gram model using ngram-count in the Stanford Research Institute Language Model (SRILM) toolkit[6], and used the smoothing parameter as ” -ukndiscount -interpolate”. With the 19972 parallel sentences, we obtained the followings. For 1-gram, we had 8346 lines. For 2-gram, we had 49685 lines. For 3-gram, we had 17241 lines. For 4-gram, we had 14651 lines.

3.5. Decoder

We used “Moses[4]” as a decoder. In a Chinese to English translation, the position of the verb is sometimes significantly changed from its original position . Thus, we set the “distortion weight (weight-d)” to “0.2” and “distortion-limit” to “-1”. Table 3 shows the other parameters. Also, we did not optimize these parameters or did not use the reordering model.

Table 3: Parameters of moses.ini

ttable-limit	40	0			
weight-d	0.1				
weight-l	1.0				
weight-t	0.5	0.0	0.5	0.1	0.0
weight-w	-1				
distortion-limit	-1				

4. Results of Statistical Machine Translation (IWSLT 2008 Automatic Evaluation Scores)

Table 11 shows the summary of the results of our statistical machine translation evaluation for the BTEC-CE, Challenge-CE, and Challenge-CE tasks.

In this table, ”primary” means the normal statistical machine translation and ”contrast” means that we removed the long parallel sentences from the training parallel data. Also, ASR.1 means 1-BEST task, and CRR means TEXT task.

Table 4 shows examples of the evaluation results of our statistical machine translation for the BTEC-CE 1-BEST task . Table 5 shows examples of the evaluation results of our statistical machine translation for the BTEC-CE TEXT task. Table 6 shows examples of the evaluation results of our statistical machine translation for the Challenge-CE task for 1-BEST. Table 7 shows examples of the evaluation results of our statistical machine translation for the Challenge-CE task for TEXT. Table 8 shows examples of the evaluation results of our statistical machine translation for the Challenge-EC task for 1-BEST. Table 9 shows examples of the evaluation results of our statistical machine translation for the Challenge-EC task for TEXT.

Table 4: Outputs for BTEC-CE ASR.1

01	in	旅游 用 包括 餐
	out	Does it include dinner , the charge .
02	in	我 是 歌手
	out	I 歌手 .
03	in	怎 才 能 到 那里 去
	out	How about going to get there
04	in	叫 服Ł生 来
	out	服Ł生 call , please .
05	in	个 怎
	out	How about this
06	in	只要 按 里 就 行
	out	Just press right here .
07	in	我 了 成 了
	out	I bought into .
08	in	再 拿 一 条 毛巾 来
	out	Bring a towel , please .
09	in	个 菜 没有 点
	out	The food hasn’t p .
10	in	死 于 日本 的 人 来 系 把 越 大 越 好 不 系网 或者 通 人 介 而 的 工作 肯定 会 有 一 个 良好 的 端
	out	I must be fine , but there is a large , the better leave as in Japan 来 系 or by the job 端 recommend 系网 someone else 死 marketing 人

Table 5: Outputs for BTEC-CE CRR

01	in	旅游 用 包括 餐
	out	Does it include dinner , the charge .
02	in	我 是 歌手
	out	I 歌手 .
03	in	怎 才 能 到 那里 去
	out	How about going to get there
04	in	叫 服Ł生 来
	out	服Ł生 call , please .
05	in	个 怎
	out	How about this
06	in	只 按 里 就 行 了
	out	Just press here just
07	in	我 了 成
	out	I’ll take 成 .
08	in	再 拿 一 条 毛巾 来
	out	Bring a towel , please .
09	in	个 菜 没有 点
	out	The food hasn’t p .
10	in	于 日本 的 人 来 系网 越 大 越 好 通 系网 或者 通 人 介 而 始 的 工作 肯定 会 有 一 个 良好 的 端
	out	And for Japan by the must be fine , 端 recommend a large , the better leave as 系网 or by the job someone else 系网 marketing 人 来

Table 6: Outputs for Challenge-CE ASR.1

01	in	先 明信片
	out	postcard , please .
02	in	局 在 ł儿
	out	Could you tell me the post office Where
03	in	局 在 ł儿
	out	Could you tell me the post office Where
04	in	不起 我 想 ł念 票
	out	Excuse me , but I'd like to buy a stamp
05	in	台 在 ł里
	out	Where is the counter
06	in	想 明信片
	out	I want to postcard
07	in	等 一下
	out	Wait a minute , uh , please .
08	in	局 怎 走
	out	How do I get to the post office
09	in	最 近 的 局 怎 走
	out	How do I get to the nearest post office
10	in	有 ł念 票
	out	Do you have any commemorative stamps

Table 7: Outputs for Challenge-CE CRR

01	in	写 明信片
	out	Please write postcard
02	in	局 在 ł儿
	out	Could you tell me the post office Where
03	in	局 在 ł儿
	out	Could you tell me the post office Where
04	in	不起 我 想 ł念 票
	out	Excuse me , but I'd like to buy a stamp
05	in	台 在 ł里
	out	Where is the counter
06	in	想 明信片
	out	I want to postcard
07	in	等 一下
	out	Wait a minute please .
08	in	局 怎 走
	out	How do I get to the post office
09	in	最 近 的 局 怎 走
	out	How do I get to the nearest post office
10	in	有 ł念 票
	out	Do you have any commemorative stamps

Table 8: Outputs for Challenge-EC ASR.1

01	in	do you or the yasaka-shrine and
	out	yasaka-shrine 和 和 ？
02	in	could you tell me where the osaka chinese
	out	能 告 我 在 ł 大阪 的 。
03	in	do or i can find the yasaka-shrine
	out	我 能 找到 yasaka-shrine 是 ？
04	in	where is the yasaka-shrine
	out	yasaka-shrine 在 ł ？
05	in	well they're chips please
	out	唔 他 。
06	in	which way to kiyomizu-temple
	out	kiyomizu-temple 怎 走 ？
07	in	how can i get there
	out	我 怎 去 ？
08	in	how long does it take to kiyomizu-temple
	out	kiyomizu-temple 要 花 多 ？
09	in	here comes our boss now
	out	我 在 在 里 老板 来 。
10	in	so far people seem quite receptive to talking to the machine of oh it's quite difficult ah because of the position of the microphone
	out	哦 Ł 。 很 的 人 像 是 很 receptive 机器 的 position 的 microphone 因 到 。

Table 9: Outputs for Challenge-EC CRR

01	in	do you know where the yasaka shrine is
	out	知道 yasaka shrine 在 ł ？
02	in	could you tell me where the yasaka shrine is
	out	能 告 我 在 ł yasaka shrine ？
03	in	do you know where i can find the yasaka shrine
	out	知道 我 在 ł 能 找到 yasaka shrine 。
04	in	where is the yasaka shrine
	out	yasaka shrine 在 ł ？
05	in	world heritage please
	out	世界 heritage 。
06	in	which way to kiyomizu temple
	out	kiyomizu 寺 怎 走 ？
07	in	how can i get there
	out	我 怎 去 ？
08	in	how long does it take to kiyomizu temple
	out	kiyomizu 寺 要 花 多 ？
09	in	here comes the bus now
	out	在 巴士 来 了 。
10	in	so far people seem quite receptive talking the machine although it's quite difficult because of the position of a microphone
	out	很 像 是 很 receptive 然 因 个 机器 。 position 的 microphone 的 人 。

We removed these long parallel sentence for training data. Our method was effective in the Challenge-EC task. However, this method was not so effective in the BTEC-CE and Challenge-CE tasks from these results.

5. Discussion

5.1. Removal of long parallel sentences

We sometimes found that poor or wrong phrase tables caused long parallel sentences in training data. So, we removed these long parallel sentences. This method is effective for the Challenge-EC task. However, this method is not so effective for the BTEC-CE and Challenge-CE tasks. This proposed method was effective for IWSLT 2007. So this method may have low reliablilty.

5.2. Unknown Words

Some words were not translated and generated as unknown words. Almost of all these words were a person's name or a place-name. Thus, if we add a procedure for unknown word, we will obtain better better. By the way, to get higher BLEU scores, a well-known trick in the field is to simply delete unknown words form the output. Table12 shows these results. To compare table11, the blue score improved about 1 persent for most experiment conditions.

5.3. Size of trainig parallel corpus

In this study, the amount of training parallel corpus was too small. So, there are many unknow words in traslation results. If we used many travel or tourist domain parallel sentences, we would have been able to obtain a higher BLEU sore.

5.4. Analysis of Outputs

We analyzed the outputs of our statistical machine translation. Single sentences provided better results with few or no errors. Long sentences such as complex or compound sentences were difficult to translate. Some sentences seemed completely wrong. We must survey why they occurred in future work.

5.5. Statistical Example Based Translation

Our system is a standard statistical machine translation system and we use long phrase tables. Thus, our system is very similar to an example based translation method , and we call our method a statistical example based translation. We believe statistical example based translation may be the better solution for Chinese-English translation.

By the way, our system is not so good performance compared other system. Our system was the 11th place among 14 system for BTEC-CE task and the 7th place among 8 system for Challenge-EC task and the 11th place among 11 system for Challenge-CE task. So we will improve many points to get better score.

6. Conclusions

We sometimes found such a wrong or poor phrase tables causes long parallel sentences in training data. So, we removed these long parallel sentences. This method was effective for the Challenge-EC task. However, this method was not so effective for the BTEC-CE and Challenge-CE tasks. We used standard statistical machine translation tools, such as "Moses"[4] and "GIZA++"[3] for our statistical machine translation systems. Finally, we obtained a BLEU score of 0.3126(1-BEST) 0.3490 (TEXT) for BTEC-CE, 0.2630 (1-BEST) 0.3034 (TEXT) for Challenge-CE, and 0.3324(1-BEST) 0.3856(TEXT) for Challenge-EC.

Our system was not good performance. For example, our system was the 7th place among 8 system for Challenge-EC task. We did not optimize these parameters or did not use the reordering model. For future experiments, we will optimize these parameters and may be add a structure information, which will enable our system to perform better.

7. Acknowledgements

We thank for the students of Tottori Univercity for their valuable comments.

8. References

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. "The machinematics of machine translation: Patamter estimation", Computational Linguestics, 19(2): pp. 263-311. (1993).

[2] Philipp Koehn, Franz J. Och, and Daniel Marcu. " Statistical phrase-based translation". In Marti Hearst and Mari Ostendorf, editors, HLT-NAACL 2003: Main Proceedings, pages 127.133, Edmonton, Alberta, Canada, May 27 -June 1. Association for Computational Linguistics. (2003).

[3] GIZA++, http://www.fjoch.com/GIZA++.html

[4] moses, http://www.statmt.org/moses/

[5] training-release-1.3.tgz, http://www.statmt.org/wmt06/shared-task/baseline.html

[6] SRILM, The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/

Table 10: Examples of phrase-tables

一个 日 游 |||a Japanese speaking guide |||0.5 0.00339841 0.333333 0.00723042 2.718

一个 日 游 ？ |||a Japanese speaking guide ? |||1 0.000771748 0.5 0.00668676 2.718

一个 |||a clock |||0.5 0.0113267 1 0.293198 2.718

一个 收音机 |||a clock radio |||1 0.00617817 1 0.293198 2.718

一个 收音机 |||a clock radio , please |||1 0.000602041 1 0.0325873 2.718

一个 明天 十点 始 的 |||a tee-off time for ten tomorrow |||1 0.000547434 1 3.02033e-05 2.718

一个 明治 神殿 的 Ł身符 可以 知 —||A charm from Meiji shrine , a written oracle key holder |||1 3.76995e-05 1 7.77965e-08 2.718

Table 11: Results of experiments

TASK	(case+punc) / (no-case+ no-punc)	primary / contrast	ASR.1 / CRR	bleu	nist	wer	per	gtm	meteor	ter
BTEC-CE	(case+punc)	primary	ASR.1	0.2911	6.0333	0.6208	0.5423	0.6013	0.4876	54.1800
			CRR	0.3266	6.4215	0.5873	0.5037	0.6418	0.5189	50.4520
		contrast	ASR.1	0.2757	5.8867	0.6360	0.5528	0.5924	0.4809	55.4560
			CRR	0.3185	6.3735	0.6017	0.5141	0.6333	0.5126	52.0960
	(no-case +no-punc)	primary	ASR.1	0.3126	6.8900	0.6129	0.5164	0.6220	0.5162	53.6510
			CRR	0.3490	7.4316	0.5703	0.4695	0.6677	0.5527	49.1490
		contrast	ASR.1	0.2932	6.7155	0.6338	0.5268	0.6151	0.5077	55.4810
			CRR	0.3366	7.3459	0.5928	0.4846	0.6600	0.5445	51.2980
CT-CE	(case+punc)	primary	ASR.1	0.2331	4.3752	0.6493	0.5767	0.5479	0.4737	56.1700
			CRR	0.2653	4.7015	0.6226	0.5530	0.5750	0.4953	53.6100
		contrast	ASR.1	0.2140	4.1510	0.6708	0.5961	0.5278	0.4580	58.0620
			CRR	0.2573	4.5648	0.6385	0.5663	0.5613	0.4828	55.3700
	(no-case +no-punc)	primary	ASR.1	0.2630	5.2135	0.6340	0.5333	0.5885	0.5219	54.5490
			CRR	0.3034	5.7842	0.5975	0.5011	0.6272	0.5492	51.7170
		contrast	ASR.1	0.2425	5.0223	0.6553	0.5530	0.5738	0.5072	56.4160
			CRR	0.2897	5.6630	0.6105	0.5161	0.6156	0.5379	53.1980
CT-EC	(case+punc)	primary	ASR.1	0.3482	5.6672	0.5441	0.4557	0.8558	0.5501	49.2630
			CRR	0.3975	6.1945	0.4885	0.3957	0.8521	0.5953	43.2030
		contrast	ASR.1	0.3553	5.9159	0.5422	0.4481	0.8594	0.5666	48.9160
			CRR	0.4047	6.4814	0.4785	0.3804	0.8612	0.6151	42.2910
	(no-case +no-punc)	primary	ASR.1	0.3288	5.6212	0.5852	0.4904	0.8368	0.5324	52.4250
			CRR	0.3827	6.1654	0.5258	0.4233	0.8343	0.5822	45.9450
		contrast	ASR.1	0.3324	5.8971	0.5850	0.4827	0.8415	0.5493	52.0320
			CRR	0.3856	6.4656	0.5162	0.4095	0.8448	0.6011	45.1350

Table 12: Delete unknown words

TASK	(case+punc) / (no-case+ no-punc)	primary / contrast	ASR.1 / CRR	bleu	nist	wer	per	gtm	meteor	ter
BTEC-CE	(case+punc)	primary	CRR	0.3408	6.0026	0.5761	0.4984	0.6607	0.5350	-
		contrast	CRR	0.3329	5.9331	0.5883	0.5065	0.6536	0.5289	-
	(no-case +no-punc)	primary	CRR	0.3705	7.1515	0.5529	0.4576	0.6904	0.5712	-
		contrast	CRR	0.3592	7.0826	0.5721	0.4694	0.6844	0.5630	-
CT-EC	(case+punc)	primary	CRR	0.4085	6.1261	0.4829	0.3922	0.8666	0.5973	-
		contrast	CRR	0.4132	6.4773	0.4718	0.3765	0.8724	0.6171	-
	(no-case +no-punc)	primary	CRR	0.3969	6.0371	0.5184	0.4197	0.8505	0.5842	-
		contrast	CRR	0.3945	6.4120	0.5103	0.4058	0.8573	0.6029	-