Statistical Machine Translation without Long Parallel Sentences for Training Data

○Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara

(Tottori University, Japan)

The Strategy of Our Statistic Machine Translation
1) Long Phrase Tables (Adequacy)

Adequacy ~ translation model P(E/C)

long phrase tables = achieve high accuracy

<u>20 words</u>

English to German

Word position change is very small.

→ short phrase table

Chinese to English

Some word are moved from their original position.

→ long phrase tables.

The Strategy of Our Statistic Machine Translation
2) 4-gram Model (Fluency)

- Fluency ~ language model P(E)

  Not use higher N-gram model.
  (the reliability for each parameter becomes low)

  <u>normal 4-gram model</u>

The Strategy of Our Statistic Machine Translation
3) Remove Long Parallel Sentences

Long Parallel Sentences → Wrong Phrase Table
→ Low Blue Score

Much Parallel Sentences → High Translation?

The Strategy of Our Statistic Machine Translation
4) Standard Tools

GIZA++.2003-09-30.tar.gz

moses.2007-05-29.tgz

training-release-1.3.tgz(train-phrase-model.perl)

(Made only some small tools to build a temporal corpus.)

C 1 在 门厅 下面 。我 这 就 给 您 拿 一些 。如果 您 还 有 什么 需要 的 请 告诉 我 。
C 2 不用 担心 那个 。我 要 买 它 你 不 需要 把 它 包 起来 。
C 3 你 可以 改改 吗 ？
C 4 红绿灯 是 红 的 。
C 5 我们 想 要 张 靠 窗户 的 桌子 。

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 .

BTEC-CE , Challenge-CE training-data
(Not change the case,Punctuation procedure)

E1 it's just down the hall i'll bring you some now if there is anything else you need just let me know
E2 no worry about that i'll take it and you need not wrap it up
E3 do you do alterations
E4 the light was red
E5 we want to have a table near the window

C1 在 门厅 下面 。我 这 就 给 您 拿 一些 。如果 您 还 有 什么 需要 的 请 告诉 我 。
C2 不用 担心 那个 。我 要 买 它 你 不 需要 把 它 包 起来 。
C3 你 可以 改改 吗 ？
C4 红绿灯 是 红 的 。
C5 我们 想 要 张 靠 窗户 的 桌子 。

Challenge-EC training-data
(Small case,Punctuation procedure)

Long Phrase Table

train-phrase-model.perl (training-release-1.3.tgz)

Long phrase table:
Max-phrase-length: 20 (default 7)

Other parameters :defaults value.

Example of Phrase Tables (BTEC-CE)

一 个 日语 导游 |||

a Japanese speaking guide |||

0.5 0.00339841 0.333333 0.00723042 2.718

一 个 日语 导游 吗 ？ |||

a Japanese speaking guide ? |||

1 0.000771748 0.5 0.00668676 2.718

一 个 时钟 收音机 谢谢 |||

a clock radio , please |||

1 0.000602041 1 0.0325873 2.718

一 个 明天 十点 开始 的 |||

a tee-off time for ten tomorrow |||

1 0.000547434 1 3.02033e-05 2.718

一 个 明治 神殿 的 护身符 可以 预 知 |||

A charm from Meiji shrine , a written oracle key holder |||

1 3.76995e-05 1 7.77965e-08 2.718 \\

4-gram language model

Best language model for IWSLT2007)

Stanford Research Institute Language Model (SRILM) toolkit

smoothing parameter : " -ukndiscount -interpolate".

19972 parallel sentences

1-gram, 8346 lines
2-gram, 49685 lines
3-gram, 17241 lines
4-gram, 14651 lines

Remove Long Parallel Sentences

English-Chinese Parallel (> 64 char) ： 645 sentences

It's just down the hall . I'll bring you some now . If there is anything else you need , just let me know .
在 门厅 下面 。 我 这 就 给 您 拿 一些 。 如果 您 还 有 什么 需要 的 请 告诉 我 。

I twisted it playing tennis . It felt Okay after the game but then it started turning black-and-blue . Is it serious ?
我 打 网球 时 扭伤 的 。 刚 打 完 后 觉得 没 什么 可是 现在 它 开始 变 得 青 一块 紫 一 块 的 。 它 严重 吗 ？

I'm looking for a nice , quiet grill-type restaurant . Would you point them out on this map ?
我 在 找 一 家 好 点 的 安静 的 烧烤 类型 的 餐馆 。 你 能 在 这 张 地图 上 指出 它们 吗 ？

The pleasure is all mine , Mr . Green . I've heard a lot about you from Mr . Smith .
我 真 高兴 格林 先生 。 我 从 史密斯 先生 那儿 听到 很多 有关 你 的 情况 。

From two hours before the departure . And please come to the counter at least thirty minutes before flight time .
从 起飞 前 两 个 小时 开始 办理 。 请 至少 在 班机 起飞 前 三 十 分钟 来 到 柜台 。

Decoder : Moses (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
                                        (Cross Entropy)
weight-w     -1
distortion-limit  -1
```

(The position of the verb changed significantly)

Results of Challenge-EC

			bleu	nist	wer	per	gtm	meteor	ter
(case+punc)	primary	ASR.1	0.35	5.67	0.54	0.46	0.86	0.55	49.26
		CRR	0.4	6.19	0.49	0.4	0.85	0.6	43.2
	contrast	ASR.1	0.36	5.92	0.54	0.45	0.86	0.57	48.92
		CRR	0.4	6.48	0.48	0.38	0.86	0.62	42.29
(no-case +no-punc)	primary	ASR.1	0.33	5.62	0.59	0.49	0.84	0.53	52.43
		CRR	0.38	6.17	0.53	0.42	0.83	0.58	45.95
	contrast	ASR.1	0.33	5.9	0.59	0.48	0.84	0.55	52.03
		CRR	0.39	6.47	0.52	0.41	0.84	0.6	45.14

Primary : Standard moses:
19972 English-Chinese parallel sentences
Contrast: Remove Long Parallel Sentences (> 96 char)
19387 English-Chinese parallel sentences

Results of Challenge-CE

			bleu	nist	wer	per	gtm	meteor	ter
(case+punc)	primary	ASR.1	0.23	4.38	0.65	0.58	0.55	0.47	56.17
		CRR	0.27	4.7	0.62	0.55	0.58	0.5	53.61
	contrast	ASR.1	0.21	4.15	0.67	0.6	0.53	0.46	58.06
		CRR	0.26	4.56	0.64	0.57	0.56	0.48	55.37
(no-case +no-punc)	primary	ASR.1	0.26	5.21	0.63	0.53	0.59	0.52	54.55
		CRR	0.3	5.78	0.6	0.5	0.63	0.55	51.72
	contrast	ASR.1	0.24	5.02	0.66	0.55	0.57	0.51	56.42
		CRR	0.29	5.66	0.61	0.52	0.62	0.54	53.2

Primary : Standard moses
19972 Chinese-English parallel sentences
Contrast: Remove Long Parallel Sentences (> 48 char)
19327 Chinese-English parallel sentences

Results of BTEC-CE

			bleu	nist	wer	per	gtm	meteor	ter
(case+punc)	primary	ASR.1	0.29	6.03	0.62	0.54	0.6	0.49	54.18
		CRR	0.33	6.42	0.59	0.5	0.64	0.52	50.45
	contrast	ASR.1	0.28	5.89	0.64	0.55	0.59	0.48	55.46
		CRR	0.32	6.37	0.6	0.51	0.63	0.51	52.1
(no-case no-punc)	primary	ASR.1	0.31	6.89	0.61	0.52	0.62	0.52	53.65
		CRR	0.35	7.43	0.57	0.47	0.67	0.55	49.15
	contrast	ASR.1	0.29	6.72	0.63	0.53	0.62	0.51	55.48
		CRR	0.34	7.35	0.59	0.48	0.66	0.54	51.3

Primary : Standard moses
19972 Chinese-English parallel sentences
Contrast: Remove Long Parallel Sentences (> 48 char)
19327 Chinese-English parallel sentences

Results of IWSLT08

Remove Long Parallel Corpus

→ Effective for Challenge-EC
→ Not effective for Challenge-CE
& BTEC-CE

(Effective for Many Cases)

01in do you or the yasaka-shrine and
outyasaka-shrine 和 和 吗 ？
02in could you tell me where the osaka chinese
out你 能 告诉 我 在 哪 大阪 的 。
03in do or i can find the yasaka-shrine
out我 能 找到 yasaka-shrine 还是 ？
04in where is the yasaka-shrine
outyasaka-shrine 在 哪 ？
05in well they're chips please
out唔 他们 筹码 谢谢 。
06in which way to kiyomizu-temple
outkiyomizu-temple 怎么 走 ？
07in how can i get there
out我 怎么 去 ？
08in how long does it take to kiyomizu-temple
outkiyomizu-temple 要 花 多长 时间 ？
09in here comes our boss now
out我们 现在 在 这里 老板 来 。
10in 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 啊 。很 难 的 人 说话 像 是 很 远 receptive 机器 的 position 的 microphone 因为 到 。

Outputs for Challenge-EC ASR.1

01in do you know where the yasaka shrine is
out你 知道 yasaka shrine 在 哪 ？
02in could you tell me where the yasaka shrine is
out你 能 告诉 我 在 哪 yasaka shrine 吗 ？
03in do you know where i can find the yasaka shrine
out你 知道 我 在 哪 能 找到 yasaka shrine 。
04in where is the yasaka shrine
outyasaka shrine 在 哪 ？
05in world heritage please
out世界 heritage 。
06in which way to kiyomizu temple
outkiyomizu 寺庙 怎么 走 ？
07in how can i get there
out我 怎么 去 ？
08in how long does it take to kiyomizu temple
outkiyomizu 寺庙 要 花 多长 时间 ？
09in here comes the bus now
out现在 巴士 来 了 。
10in so far people seem quite receptive talking the machine although it's quite difficult because of the position of a microphone
out很 难 说话 像 是 很 远 receptive 虽然 因为 这个 机器 。position 的 microphone 的 人 。

Outputs for Challenge-EC CRR

01 in 旅游 费用 包括 晚餐 吗
out Does it include dinner , the charge .
02 in 我 是 歌手
out I 歌手 .
03 in 怎样 才 能 到 那里 去 呢
out How about going to get there
04 in 请 叫 服务生 来
out 服务生 call , please .
05 in 这个 怎么样
out How about this
06 in 只要 按 这里 就 行
out Just press right here .
07 in 我 买 了 成 了
out I bought into .
08 in 请 再 拿 一 条 毛巾 来
out Bring a towel , please .
09 in 这个 菜 没有 点
out The food hasn't p .
10 in 死 于 日本 的 营销 人员 来说 关系 把 越 大 越 好 不过 关系网 或者 通过 别人 介绍 而 开设 的 工作 肯定 会 有 一 个 良好 的 开端
out I must be fine , but there is a large , the better leave as in Japan 来说 关系 or by the 开设 job 开端 recommend 关系网 someone else 死 marketing 人员

Outputs for BTEC-CE ASR.1

01in 旅游 费用 包括 晚餐 吗
outDoes it include dinner , the charge .
02in 我 是 歌手
outI 歌手 .
03in 怎样 才 能 到 那里 去 呢
outHow about going to get there
04in 请 叫 服务生 来
out服务生 call , please .
05in 这个 怎么样
outHow about this
06in 只 按 这里 就 行 了
outJust press here just
07in 我 买 了 成药
outI'll take 成药 .
08in 请 再 拿 一 条 毛巾 来
outBring a towel , please .
09in 这个 菜 没有 点
outThe food hasn't p .
10in 对于 日本 的 营销 人员 来说 关系网 越 大 越 好 通过 关系网 或者 通过 别人 介绍 而 开始 的 工作 肯定 会 有 一 个 良好 的 开端
outAnd for Japan by the must be fine , 开端 recommend a large , the better leave as 关系网 or by the job someone else 关系网 marketing 人员 来说

Outputs for BTEC-CE CRR

01in 请 先 明信片
outpostcard , please .
02in 请问 邮局 在 哪儿
outCould you tell me the post office Where
03in 请问 邮局 在 哪儿
outCould you tell me the post office Where
04in 对不起 我 想 买 纪念 邮票
outExcuse me , but I'd like to buy a stamp
05in 柜台 在 哪里
outWhere is the counter
06in 想 买 明信片
outI want to postcard
07in 嗯 请 等 一下
outWait a minute , uh , please .
08in 邮局 怎么 走
outHow do I get to the post office
09in 最 近 的 邮局 怎么 走
outHow do I get to the nearest post office
10in 有 纪念 邮票 吗
outDo you have any commemorative stamps

Outputs for Challenge-CE ASR.1

01in 请 写 明信片
outPlease write postcard
02in 请问 邮局 在 哪儿
outCould you tell me the post office Where
03in 请问 邮局 在 哪儿
outCould you tell me the post office Where
04in 对不起 我 想 买 纪念 邮票
outExcuse me , but I'd like to buy a stamp
05in 柜台 在 哪里
outWhere is the counter
06in 想 买 明信片
outI want to postcard
07in 请 等 一下
outWait a minute please .
08in 邮局 怎么 走
outHow do I get to the post office
09in 最 近 的 邮局 怎么 走
outHow do I get to the nearest post office
10in 有 纪念 邮票 吗
outDo you have any commemorative stamps

Outputs for Challenge-CE CRR

Consideration:Remove Unknown Words

Results of BTEC-CE

			bleu	nist	wer	per	gtm	meteor
(case+punc)	primary	CRR	0.34	6	0.58	0.5	0.66	0.54
	contrast	CRR	0.33	5.93	0.59	0.51	0.65	0.53
(no-case	primary	CRR	0.37	7.15	0.55	0.46	0.69	0.57
+no-punc)	contrast	CRR	0.36	7.08	0.57	0.47	0.68	0.56

Results of Challenge-EC

(case+punc)	primary	CRR	0.41	6.13	0.48	0.39	0.87	0.6
	contrast	CRR	0.41	6.48	0.47	0.38	0.87	0.62
(no-case	primary	CRR	0.4	6.04	0.52	0.42	0.85	0.58
+no-punc)	contrast	CRR	0.39	6.41	0.51	0.41	0.86	0.6

Primary : Standard moses

Contrast: Remove Long Parallel Sentences

Future study

- Optimize parameters
- Unknown word procedure
- More large database
- Not used parallel sentence
  (If output likelihood is high,
  use as parallel sentence)

Conclusions

- Remove Long Parallel Sentences
- Long phrase table
- Standard tools
- Statistical Example Based Translation

- Good results for Change-EC
- 0.4047 BLEU score for IWSLT08