Statistical Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables

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

Our statistical machine translation system that uses large Japanese-English parallel sentences and long phrase tables is described. We collected 698,973 Japanese-English parallel sentences, and we used long phrase tables. Also, we utilized general tools for statistical machine translation, such as "Giza++"[1], "moses"[2], and "training-phrase-model.perl"[3]. We used these data and these tools, We challenge the contest for IWSLT07. In which task was the result (0.4321 BLEU) obtained.

1. Introduction

Many machine translation systems have been developed for long time and have over three generations of technology.

The first generation was a rule-based translation method, which was developed over the course of many years. This method had translation rules that were written by hand. Thus, if the input sentence completely matched the rule, the output sentence had the best quality. However, many expressions are used for natural language, this technology had very small coverage. In addition, the main problem are that the cost to write rules was too high and that maintaining the rules was hard.

The second generation involved example-based machine translation method. This method finds a similar sentence from corpus and generates a similar output sentence. The problem with this method is calculating the similarity. Many methods like dynamic program (DP) are available. However, they are very heuristic and intuitive and not based on mathematics.

The third generation was a statistical machine translation method and is very popular now. This method is based on the statistics, and it is very reasonable. Even though, many models for statistical machine translation are available. An early model of statistical machine translation was based on IBM1 $\sim$ 5. This model is based on words and thus a "null word" model is needed. This "null word" model sometimes has very hard and serious problems, especially decoding. Thus, recent statistical machine translation usually use a phrase-based models.

Incidentally, two points are used to evaluate the English sentences of machine translations; one is adequacy, and the other is fluency. We believe adequacy is related to translation model $P(E/J)$ and fluency is related to language model $P(E)$. So, we need to make long phrase tables to achieve high adequacy. Similar languages like English and German may have short phrases. However, languages that differ greatly, like Japanese and English, need long phrases. And long phrase tables mean that a large number of Japanese-English parallel sentences are needed. Also, we believe that word trigram model is enough to express the fluency of English.

We implemented our statistical machine translation using a large number of Japanese-English parallel sentences and long phrase tables. So, our system was similar to a statistical example-based translation system. We believe that these concepts provide the best method for Japanese-English translation.

We collected 698,973 Japanese-English parallel sentences. And we made number of 3,769,988 phrase tables from them. Also, we used general tools for statistic machine translation, such as "Giza++"GIZA++, "moses"[2], and "training-phrase-model.perl"[3]. We used this data and these tools, we challenge the contest for IWSLT07 and obtained the BLEU score of 0.4321.

2. Concepts of our Statistical Machine Translation System

In this section, we describe our concepts for our Japanese-English statistical machine translation system.

2.1. Standard Tools

Many statistical machine translation tools have been developed and published. These tools have high reliability and are widely used. Whenever possible we did not make special tools, instead relying on following tools.

1. GIZA++.2003-09-30.tar.gz [1]
2. moses.2007-05-29.tgz [2]
3. training-release-1.3.tgz(train-phrase-model.perl) [3]

We made only some small tools to build a temporal corpus.

2.2. Large number of Japanese-English Parallel Sentences

We should collect many Japanese-English parallel sentences as possible to achieve high performance. We collected 698,973 parallel sentences from electronic medias. Next, we describe section 3 in more detail. Attention, these parallel sentences have some errors.

2.3. Long Phrase Tables (Adequacy)

We evaluate the adequacy and the fluency of English translated sentences. We believe that adequacy is related to translation model $P(E/J)$. Thus, we need to make long phrase tables to achieve high adequacy.

In similar language like English to German, word position change is very small. In such a case, short phrase table has little trouble. However, in Japanese to English translation, verbs are moved from their original position. So, we need to make long phrase tables.

2.4. Word Trigram Model (Fluency)

We can evaluate English sentences on two points; adequacy and fluency. We believe that fluency is related to language model $P(E)$. Thus we used a normal trigram model and did not use a higher N-gram model. In general, when we use a higher order N-gram, the number of parameters dramatically increase. It occurs that, the reliability for each parameter becomes low.

We believe that a trigram model is the best language model to express fluency.

3. Large J/E Parallel Corpus

We collected large number of Japanese-English parallel sentences from many electronic medias. There are many electronic media like Japanese English dictionaries, English sample sentences and CD-ROMs. There are 8 types in electronic medias. Following gives the details of these types.

3.1. Electronic Dictironary

Many Japanese-English or English-Japanese electronic dictionaries are on the market. Two kind of formats are available. One is open format and the other is special format.

3.1.1. A: Open Format for Electronic Dictionaries (EPWING)

The EPWING format[33] is Japanese unique electronic dictionary format. This format is mainly built by Fujitsu and used by various publishing and software companies. JIS X4081 specifications were established for this format in 1996. Currently, over 50 kinds of dictionaries of this type are available.

This format is based on JIS code, so extracting raw sentences is very easy. However, difficulties often occur with extracting parallel sentences in these medias. Certain dictionaries have parallel sentence files that are separate from the dictionary. These enable parallel sentences to be extracted easily. However, there are very few cases. Parallel sentences are normally completely embedded in raw dictionary characters. Therefore, some special keys are needed to extract these sentences. For example, the head line of parallel sentence is "～ " in Genius English Japanese Dictionary[32]. Therefore, we had to make many small tools to extract parallel sentences for each electronic media.

3.1.2. B: Special Format in Electronic Dictionaries

Some electronic dictionaries have their own format and need to use special browsers. Extracting parallel sentences from these dictionary is very hard. However, "Random House[13]", a very well known English dictionary, is suitable for this task. Because the format of this dictionary has already been well known in detail.

"ビジネス技術実用英語大辞典 [14]" also has many parallel sentences, so we surveyed this format and extracted parallel sentences from it.

However the survey of all this kinds of dictionaries is difficult. So, we did not used any more this kind of these dictionaries.

3.1.3. C: Books with CDROMs

Some books are now published with CDROM. Thus some tools enables extracting parallel sentences from CDROM. However, each book has small parallel sentences. "英文ビジネスレター文例大辞典"[10]" is a good example.

3.1.4. D: Internet

Some parallel sentences exist on the Internet (Web). These sentences are simple example sentences and are used for educating middle school children. "英語教師用データベース [16]" is one good example. (This site is now closed.)

3.1.5. E: Newspapers

Newspapers are very important linguistic resources. The major newspaper publishing companies publish both Japanese and English newspapers. Among them, the "Yomiuri Shimbun" publishes "The Daily Yomiuri" and we can obtain these CDROMs. However, Japanese articles do not correspond to English articles perfectly. Thus making Japanese to English parallel sentences is very difficult. However, NICT[34] publishes parallel sentences that we can use.

3.1.6. F: Published Parallel Sentences

Electronic Japanese-English parallel sentences have been published, though these medias have a few case. However

they are free to research purpose. A good example is “英文ビジネスレター文例大辞典 [10]”.

3.1.7. G: Unpublished Parallel Sentences

These sentences are the best kind of Japanese-English parallel sentences. Machine translation researchers actively collect them. They are ideal and complete parallel sentences, because they have no errors and represent the best English translation. The one disappointing thing is that they cannot be given to other researcher. A good example is “IPAL English sentences[9]”. We have collected many this kind of parallel sentences. However, we did not use this kind of parallel sentences for IWSLT07.

3.1.8. H: Other

No copyrighted books that have been translated in English are available. The project to create such books is called “Project Sugita Genpaku [35]”. We obtained Japanese texts and translated English texts. Unfortunately we can not obtain a Japanese sentence using an English sentence. Therefore, we could not extract Japanese-English parallel sentence from this corpus. However, this project has been very interesting.

We have heard that many translated patent texts are available in English. Sentences from these texts may enhance our future database.

3.2. Number of Parallel Sentences

3.2.1. Extracted parallel sentences

We collected about 698,973 parallel sentence with 8,439,907 words in English and 10,367,940 words in Japanese. About 70% of these sentences are simple sentences, about 20% of are complex or compound sentences, and the remaining 10% of these sentences are complex and compound sentence and very long sentence. A lot of these sentences have descriptive text. The amount of dialog text is small. So most of these sentences are not included the travel or tourist domain.

Table 1 shows examples of our collected Japanese-English parallel sentences.

Table 2 shows the names of the dictionaries and the type of dictionaries and the number of extracted sentences.

Table 1: Example of Extracted Sentences

元気がなくぼんやり見つめていた。 She was listless and had a vacant stare.
星がさっと空を横切って流れた。 A star shot across the sky.
出発が遅れたが時間に間に合って到着した。 I arrived in time in spite of a late start.
何か言いかけたが、思い直してやめた。 He started to say something, then thought better of it.
自分が来た道をじっと振り返っていた。 He stared back the way he had come.
どんよりと生気のない目付きで彼女をじっと見つめた。 He stared at her glassily.
現行の標準におけるセキュリティ・アソシエーションの定義は様々であり、本論文はそれらの定義を明らかにすることを試みる。 There are varying definitions of a security association in current standards and this paper attempts to clarify these definitions.
本論文では、1 次元および 2 次元の静電問題を解くために、リチャードソン外挿を有限差分法と組み合せて用いる。 In this paper, Richardson extrapolation is used in conjunction with the finite difference method to solve both one- and two-dimensional electrostatics problems.

Table 2: Extracted Sentences

	Name of Dictionary	Type	♯sentences
AA	機能試験文集 [8]	D	5,273
AC	アンカー和英辞典	A	39,923
AD	アンカー英和辞典	A	20,701
AE	学研英和辞典	F	3,826
AF	基本語用例辞典	G	24,000
AI	英文ビジネスレター文例大辞典 [10]	A	9,355
AJ	外国人のための日本語例文・問題シリーズ	F	13,830
AK	LDB	F	33
AL	SENSEVAL 対訳コーパス	A	1,096
AM	講談社和英辞典 [11]	A	40,334
AO	小倉書店 英語文型・文例辞典 [12]	F	1,330
AQ	研究社 新編英和活用大辞典 [20]	A	103,064
AR	ランダムハウス英語辞典 [13]	A	39,517
AS	ビジネス技術実用英語大辞典 [14]	B	9,309
AT	コンピュータ用語辞典第 3 版 [15]	A	3,283
AU	佐良木コーパス	A	400
AW	鳥取大学池原研究室 斎藤健太郎コーパス：比較構文	D	143
AX	鳥取大学池原研究室 澤田康子コーパス：因果関係構文	D	334
AY	英語教師用データベース [16]	D	758
AZ	研究社 総合ビジネス英語文例事典	A	952
BA	新実用英語ハンドブック [24]	A	304
BB	研究社 新和英大辞典 [25]	A	27,599
BE	エクシード英和辞典	G	2,030
BF	科学技術日英・英日コーパス辞典科学技術日英・英日コーパス辞典	B	265
BG	日本語文型辞典	G	3,721
BH	旺文社 マルチ辞書 辞ショック	A	58,005
CI	向井京子 英文 E メール文例集 池田書店 [27]	C	1,360
CK	読売新聞 (文対応データ) [18]	E	122,078
CO	NHK やさしいビジネス英語 実用フレーズ辞典	C	7,055
CQ	自然科学系和英大辞典 増補改訂新版 (小倉書店)	A	10,195
CR	ジーニアス英和・和英辞典 [28]	A	5,319
CS	朝日出版社 最新ビジネス英文手紙辞典 CD-ROM 版	A	2,232
CT	株式会社アスク 機械を説明する英語	D	2,447
DA	IWSLT training	D	39,953

4. Experiments with Statistical Machine Translation

4.1. Phrase-Tables

We used the "train-phrase-model.perl[3]" 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.

Using these 698,973 parallel sentences, we obtained 3,769,988 phrase-tables(About 385,773,337 bytes). Table 3 shows examples of phrase-tables.

4.2. Trigram model

We calculated the trigram model using ngram-count in "SRILM"[36], Also we used default parameters. With 698,973 parallel sentences, We obtained as follows. For ngram 1, we had 126200 lines. For ngram 2, we had 1578329 lines. For ngram 3, we had 779718 lines.

4.3. Decoder

We used "moses[2]" as a decoder. In a Japanese to English translation, the position of the verb is significantly changed from its original position . Thus, we set the "distortion weight (weight-d)" to "0.2" and "distortion-limit" to "-1". Also we set the " "weight-t" to "1.0 0.0 0.0 0.0 0.0". Other parameters were set to the defaults. We do not optimize these parameters.

4.4. Other

We used "chasen"[4] for a Japanese tagger. Also we used the punctuation procedure in English sentences. It means that we changed "," and "." to " , " and " . ". Also, we did not handle English case.

5. Results of Statistical Machine Translation

We obtained the BLEU score of 0.4321 for max-phrase-length set to 20 and using 698,973 Japanese-English parallel sentences. Table 4 shows examples of the results of our statistical machine translation.

Also we studied the comparison experiment of max-phrase-length and the size of parallel sentences. Table 5 shows theses results. In table 5, the 39,953 parallel sentences means IWSLT07 training data only.

Table 5: Results of experiments

BLEU	max-phrase-length	♯ parallel sentences
0.4321	20	698,973
0.4184	7	698,973
0.4315	20	39,953
0.4182	7	39,953

These results shows the following things.

1. If the max-phrase-length is enlarged, the BLEU sore is good.

   It means that the long phrase table is effective.

2. Even if the size of parallel sentences is large, the BLUE sore is not change.

   It means that large parallel sentence is not so effective.

6. Consideration

6.1. Analysis of Outputs

We analyzed the outputs of our statistical machine translation. The results are presented next.

1. Single Sentence

   Single sentences provided very good results. They had few or no errors. Good examples are "B" and "E" in table 4. We think it is the effect of a large number of Japanese-English parallel sentences.

2. Long Sentences

   Long sentences like complex or compound sentence are a little difficult to translate. Even though, these sentence have some errors, we think they provided acceptable translation results. Good examples are "A" and "C" and "H" in table 4. We think this is the effect of the long phrase table.

3. Unknown Words

   Some words were not translated and processed as unknown words. Good example sentences were "F" and "G" in table 4. "コバヤシ |UNK |UNK |UNK" is regarded as unknown words.

   Almost of all these words are a person's name or a place-name. Thus, if we add a procedure to the English alphabet, we will obtain better good results.

4. Wrong Sentences

   Some sentences were completely wrong. We must survey why they occurred. Good example is "J" in table 4.

6.2. Size of parallel sentences

In this study, the size of parallel sentences was not change the BLUE score. We think that this reason is that a lot of papallel sentences have descriptive text. The amount of dialog text is small. So most of these sentences are not included the travel or tourist domain. If we can use a lot of travel or tourist domain parallel sentences, we can obation more high BLEU sore.

6.3. Statistical Example Based Translation

Our system is a very standard statistical machine translation system. We collected a large number of Japanese-English parallel sentences, and we used long phrase tables. Thus, our system is very similar to an example based translation, and we call it a statistical example based translation. We believe statistical example based translation may be best solution for Japanese-English translation.

7. Conclusions

We collected large Japanese-English parallel corpus from electronic medias of 698,973 parallel sentences. And we made long phrase table. Our statistical machine translation system is similar to a statistic example based machine translation system. We utilized standard statistical machine translation tools like "moses"[2] and "GIZA++"[1].

We used this data and these tools, and obtained good results for simple sentences and acceptable results for complex and compound sentences for the contest of IWSLT07. We obtain 0.4321 BLEU score. In analyzing the error results, we found that unknown words decrease the BLEU score. Also we do not optimize parameters of "moses"[2] and other programs.

We will optimize these parameters and will add a procedure of unknown words, that will enable our system to achieve better performance.

8. Additional Study

We obtained the BLEU score of 0.4876 using 5 gram for language model $P(E)$ and cross entorpy for translation model $P(E/J)$ and max-phrase-length set to 20 and 698,973 Japanese-English parallel sentences.

9. Acknowledgements

We thank Eiichirou Sumita at ATR to entry the IWSLT07 and Hiroaki Nagata at NTT for their valuable comments.

10. References

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

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

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

[4] chasen, http://chasen-legacy.sourceforge.jp/

[5] EPWING, http://www.epwing.or.jp/

[6] alc, http://www.alc.co.jp/

[7] 青空文庫, http://www.aozora.gr.jp/guide/nyuumon.html

[8] 機能試験文, http://www.kecl.ntt.co.jp/icl/mtg/resources/index-j.html

[9] IPAL, http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

[10] 英文ビジネスレター文例大辞典, http://www.nikkeish.co.jp/gengo/eibun.htm

[11] 講談社和英辞典, http://cactus.aist-nara.ac.jp/lab/resource/resource-print.html#KODANSHA

[12] 小倉書店 英語文型・文例辞典, http://www.ogurashoten.co.jp/kyozai3.html

[13] ランダムハウス英語辞典, http://ebook.shogakukan.co.jp/scatalog/randam/top/top.htm

[14] ビジネス技術実用英語大辞典, http://www.nichigai.co.jp/newhp/whats/unno3.html ISBN4-8169-8127-6

[15] コンピュータ用語辞典第 3 版, http://www.nichigai.co.jp/newhp/whats/computer3.html ISBN4-8169-8126-8

[16] 英語教師用データベース, http://home.alc.co.jp/db/owa/engt_structure?stg=4

[17] 科学技術日英・英日コーパス辞典, http://pub.maruzen.co.jp/cd_others/ko-pas/ ISBN4-621-04991-7

[18] 読売新聞記事, http://www.ndk.co.jp/yomiuri/

[19] 斉藤和英大辞典, ISBN4-8169-8078-4

[20] 研究社 新編英和活用大辞典, ISBN4-7674-3573-0

[23] 研究社ビジネス英語スーパーパック, ISBN4-7674-3590-0

[24] 新実用英語ハンドブック, ISBN4-469-74233-3

[25] 研究社 新和英大辞典, ISBN 4-7674-7200-8

[27] 向井京子 英文 E メール文例集 池田書店 2002, ISBN4-262-16896-4

[28] CD-ROM 版 ジーニアス英和・和英辞典 大修館書店, ISBN4-469-79057-5

[29] Epwing 版 リーダーズ+プラス V2, ISBN4-7674-3563-3

[30] 新グローバル＆ニューセンチュリー英和・和英辞典, ISBN4-385-61400-8

[31] ＣＤ－科学技術４５万語対訳辞典　英和・和英, ISBN4-8169-8128-4

[32] ジーニアス英和大辞典 Genius English-Japanese Dictionary ISBN4-469-04131-9

[33] http://www.epwing.or.jp/

[34] http://www.nict.go.jp/

[35] プロジェクト杉田玄白 http://www.genpaku.org/index.html

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

Table 3: Examples of phrase-tables

オノ さん |||Ms. Ono?
オフィス を |||to the office as
オハイオ 州 に 変革 の 風 が 吹い て いる の を 感じる |||feel the winds of change blowing in Ohio
オペレータ の 保護 の ため に 連動 安全 扉 を 備え て いる . |||features an interlocked safety door for operator protection .
オペレーティングシステム の 中 の スケジューラ によって 主 記憶 装置 へ ロード さ れる . |||loaded into main memory by the operating system's scheduler .
オリジナル 信号 を , 小さく まとめ た 形 で 表現 し ながら も 許容 できる 程度 の 歪み で |||the original signal with an acceptable level of distortion while representing it in compact form

Table 4: Examples of Outputs

A	静か で 素敵 な すき焼き の 御店 を 探し て い ます . 地図 で 指し て もらえ ます か . I'm looking for a lovely stores in the calm and sukiyaki . Can I go on the map .
B	サイズ は 御 いくつ です か . What's your size ?
C	月曜日 の 朝 九 時 発 の 一 七 二 便 に 変更 し たい の です が ご 面倒 を 掛け て すみません . I'd like to change to flight one seven two nine o'clock on Monday morning I'm sorry to give you trouble .
D	地下鉄 の 中 で 財布 を 掏 られ ました . Some pick-pocket stole my wallet on the subway .
E	もう少し 短く し て 下さい . A little shorter , please .
F	コバヤシ さん が 副 社長 に 昇進 し ました . Mr. A コバヤシ |UNK |UNK |UNK I was promoted to vice-president .
G	明日 の トスカ の 予約 を 御 願い し たい の です が . I'd like to make an appointment for tomorrow トスカ |UNK |UNK |UNK .
H	分かり ました こちら が 搭乗 券 に なり ます . 本日 御 客 様 の フライト は イー 二 十 五 の 搭乗 口 から 出発 し ます . 出発 の 三 十 分 前 迄 に ゲート に い て 下さい . Okay . Here's your boarding pass , flight depart from 30 minutes before departure gate for twenty-five cents for today is E . Please stay at the gate .
I	来週 末 十 五 日 と 十 六 日 に ダブル の 部屋 を 予約 し たい の です . I'd like to reserve a room for next weekend on Monday and Tuesday to double .
J	日本人 の 八 十 パーセント 近く が 都市 部 に 住ん で い ます . Nearly 80 percent of Japanese cities, you live ?