Statistical Machine Translation using Large Japanese-English Parallel Corpus and Long Phrase Tables

○Jin'ichi Murakami, Masato Tokuhisa, Satoru Ikehara

(Tottori University, Japan)

Our Statistical Machine Translation

Feature:

Large number of Japanese-English parallel sentences
698,973 Japanese-English parallel sentences.

Long phrase tables (20 word)
3,769,988 phrase tables

Standard Tools (moses,GIZA++,etc)

Challenge the contest for IWSLT07:

BLEU score = 0.4321

The Strategy of Our Statistical Machine Translation

Evaluate English sentences → Adequacy & Fluency

We believe :

Adequacy ~ translation model P(E/J)

high adequacy → long phrase

long phrase → a large number of J/E parallel sentences

Fluency ~ language model P(E)

word trigram → enough to express the fluency

Large number of Japanese-English Parallel Sentences

Collect many Japanese-English parallel sentence as possible

・ Japanese English Electronic Dictionaries
・ CD-ROMs
・ Network
・ English sample sentences

.

8 types in electronic medias.

A: Open Format for Electronic Dictionaries

EPWING format:

- Based on JIS code
- Extracting raw sentences is very easy
- Over 50 kind of Dictionaries
- Many parallel sentences
- Not so easy to extracting parallel sentences.
- Parallel sentences are normally completely embedded in raw dictionary characters

Make many small tools
(to extract parallel sentences for each electronic media)

Example: “ アンカー英和””アンカー和英”

B:Special Format in Electronic Dictionaries

- Extracting parallel sentences is very hard

- Random House:
  Format of this dictionary has already been analy

- ビジネス技術実用英語大辞典 :
  We analyzed this format

C: Books with CDROMs

- Some books are published with CDROM
- Enables extracting parallel sentences from CDROM
- (small parallel sentences)

Example: 英文ビジネスレター文例大辞典

D: Internet

- Parallel sentences are in the Internet.
- Simple example sentences
- Educating for middle school children.

Example “ 英語教師用データベース " in ALC

E: Newspapers

・　Publish both Japanese and English newspapers.

・　Japanese articles do not correspond to English articles.

・　(NICT published parallel sentences with some errors)

F: Published Parallel Sentences

・ Very few case.

Example :`` 英文ビジネスレター文例大辞典 ".

G: Unpublished Parallel Sentences

- Best kind of Japanese-English parallel sentences
- No errors and Best English translation.
- Machine translation researchers actively collect
- Cannot be given to other researcher.

Example : ``IPAL English sentences''.

(we did not use these parallel sentences for IWSLT07)

H: Future database(not use)

“Project 杉田玄白　 (Sugita Genpaku)“

Collect no copyrighted books and translate in English

“Patent Text“

About 2,000,000 Japanese English parallel sentences

Total Extracted Parallel Sentences

698,973 parallel sentence

8,439,907 words in English
10,367,940 words in Japanese

Simple sentences:70%
Complex or compound sentences:20%
Very long sentence:10%

Descriptive text :Most Dialog text :Little

Example of Extracted Parallel 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。
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.

Number of extracted parallel sentences (in some parts)

	Name of Dic.	Type	#sentences
AA	機能試験文集	D	5273
AC	アンカー和英辞典	A	39923
AD	アンカー英和辞典	A	20701
AE	学研英和辞典	F	3826
AF	基本語用例辞典	G	24000
AI	英文ビジネスレター文例大辞典	A	9355
AJ	外国人のための日本語例文・問題シリーズ	F	13830
AK	LDB	F	33
AL	SENSEVAL 対訳コーパス	A	1096
AM	講談社和英辞典	A	40334
AO	小倉書店 英語文型・文例辞典	F	1330
AQ	研究社 新編英和活用大辞典	A	103064
AR	ランダムハウス英語辞典	A	39517
AS	ビジネス技術実用英語大辞典	B	9309
AT	コンピュータ用語辞典第3版	A	3283
AU	佐良木コーパス	A	400
AW	鳥取大学池原研究室 齋藤健太郎コーパス：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	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	A	304
BB	研究社 新和英大辞典	A	27599
BE	エクシード英和辞典	G	2030
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	3721
BH	旺文社 マルチ辞書 辞ショック	A	58005
CI	向井京子 英文Eメール文例集 池田書店	C	1360
CK	読売新聞(文対応データ)	E	122078
CO	NHKやさしいビジネス英語 実用フレーズ辞典	C	7055
CQ	自然科学系和英大辞典 増補改訂新版(小倉書店)	A	10195
CR	ジーニアス英和・和英辞典	A	5319
CS	朝日出版社 最新ビジネス英文手紙辞典CD-ROM版	A	2232
CT	株式会社アスク 機械を説明する英語	D	2447

Tagging & Case

"chasen" for Japanese tagger.

Example " 元気がなくぼんやり見つめていた。”

→ " 元気_が_なく_ぼんやり_見つめ_て_い_た_ .
"

Punctuation procedure in English sentence.
Not change the case.

Example "Pass_the_bread,_please."

→ "Pass_the_bread_,_please_."

Long Phrase Tables (Adequacy)

We believe:

Adequacy ~ translation model P(E/J)

long phrase tables = achieve high accuracy

English to German

Word position is not so moved

→short phrase table

Japanese to English

Verbs are too moved from their original position.

→ long phrase tables.

Long Phrase Table

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

To obtain long phrase table:

The parameter of max-phrase-length: 20
(default 7)

Other parameters :defaults

3,769,988 phrase-tables

Example of Phrase Tables

```
オノ さん ||| Ms. Ono? |||1 0.00194496 0.166667 0.0073622 2.718
オフィス を ||| to the office as ||| 1 0.00253428 1 0.000555509 2.718
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 |||
                1 7.18001e-13 1 9.03199e-08 2.718
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 . |||
                0.5 1.86027e-18 1 2.90433e-10 2.718
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 . |||
                1 9.01527e-17 1 8.61617e-08 2.718
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 |||
                1 1.82164e-26 1 1.71529e-21 2.718 \\
```

Word Trigram Model (Fluency)

We believe:

- Fluency ~ translation model P(E)
  Used a normal trigram model
  Not use higher N-gram model.
  (the reliability for each parameter becomes low)

- (trigram model may be the best language model
  to express fluency?)

Trigram model

ngram-count in "SRILM“

default parameters.

Number of ngram 1 :   126200 lines.
Number of ngram 2 : 1578329 lines.
Number of ngram 3 : 7797188 lines.

```
-1.6732     for three years,
-0.9031     four three .
-0.5052     four three four
-0.9031     four three two
-1.5305     from three .
-1.5305     from three days
-1.3544     from three directions
-1.5305     from three major
-1.5305     from three months
-1.3544     from three o'clock
-1.5305     from three of
-1.5305     from three times
-1.5305     from three to
```

Example of trigrams

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.)

Distortion Weight in Decorder (Moses）

Distortion weight (weight-d) : 0.2
(default : 1.0）

The position of the verb changed significantly

Our Machine Spec.

"open-suse-10.2 AMD64"
"AMD Opteron 185" (2.6GHz)”
“4G byte DDR memory”

“About 30 minute for input 500 sentence”

Results of IWSLT07

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

Long Phrase Table → Effective

Large Parallel Corpus → Not so effective

(Most of parallel sentence is descriptive text,
Domain is very different)

Analysis of output 1: Simple Sentence

・ Good results and no errors.

サイズ は 御 いくつ です か .
→What's your size ?

地下鉄 の 中 で 財布 を 掏 られ まし た .
→ Some pick-pocket stole my wallet on the subway .

もう少し 短く し て 下さい .
→ A little shorter , please .

Analysis of output 2: Long Sentence

- a little difficult. some errors,
- Acceptable translation results?
- <Effect of long phrase table?>

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 .

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 .

Analysis of output 3 : Unknown Word

- Some words are not translated
- Processed as unknown words.

コバヤシ さん が 副 社長 に 昇進 し まし た ．

→Mr. A コバヤシ was promoted to vice-president .

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 .

Analysis of output 4 : Failure

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 ?

Why?????

Consideration:

Statistical Example Based Translation

Our system:

- Long phrase table

  ~ Similar to (statistical) example based translation.

We think :

Statistical example based translation may be best solution for J/E translation.

Conclusions

・ Large Japanese-English parallel corpus from electric medias.

・ Long phrase table.

・ Standard tools

・ Statistical Example Based Translation

・ Good results for simple sentence

・ Better(acceptable) results for long sentence.

・ 0.4321 BLEU score for IWSLT07

Future study

- Optimize parameters
- Unknown word procedure
- Cross Entrory. P(J/E)
- More large database
- Study of closed data
- Minimun devidion method
- Not used parallel sentence
  (If output likelihood is high, use as parallel sentence)

Additional Study (Best result of IWSLT07)

BLEU	NIST	WER	PER	GTM	METER	TER
0.4991	7.9796	0.4317	0.3617	0.7339	0.7147	38.51

Language Model P(E)	5-gram LDC+Newspaper (12,983,208 sentences)
Translation Model P(E/J)	Cross Probability (weight-t=(0.5 0 0.5 0 0) 698,973 sentences
Max Phrase Length	32
Opt. Parameter	No
Unknown Word	No

Acknowledgements

Eiichirou Sumita, Hideo Ookuma,
Hiroshi Yamamoto, Michael Paul,
Hirokazu Yoshioka in ATR

Hiroaki Nagata in NTT