モノリンガルデータを増加させた場合の統計翻訳の精度調査

善行佑介 *1 村上仁一 *2 徳久雅人 *2
*1 鳥取大学 工学部 知能情報工学科
*2 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
*1,*2{s102025, murakami, tokuhisa} @ ike.tottori-u.ac.jp

1 はじめに

統計翻訳は，対訳データから学習する翻訳モデルと，モノリンガルデータから学習する言語モデルを用いて，確率的に翻訳をする．

アラビア語-英語 [1] や, 中国語-英語 [2] では数百万文もの多量の対訳データが提供されている．しかし, 日本語-英語で提供されている対訳データの量は少ない．したがって，日英翻訳や英日翻訳において多量の対訳データを用いて，精度の高い翻訳モデルを作成することは困難である．

そこで本研究では，データを収集することが容易な，モノリンガルデータを大量に使用して，言語モデルを学習し，日英翻訳と英日翻訳の翻訳精度の向上を目指す．

2 実験方法

ベースラインは辞書から抜き出した単文 (辞書文) [3] を使用する．翻訳モデルの学習には 100,000 文，言語モデルの学習には 100,000 文を用いる．テスト文として辞書文 10,000 文を使用する．ディベロップメント文には 1000 文を使用する．実験は日英翻訳と英日翻訳を行う．

2.1 翻訳モデルの学習

翻訳モデルを学習するために，GIZA++[4] を用いる．

2.2 言語モデルの学習

言語モデルを学習するために，SRILM[5] の ngram-count を用いる．N-gram モデルに 5-gram の言語モデルを用いる．

2.3 デコーダのパラメータ

デコーダには moses[6] を用いる．

2.4 パラメータのチューニング

本実験では 3 種類のコーパスを扱う (第 2.5 節参照) ．実験条件を揃えるために，各コーパスごとに同じ moses のパラメータを使う．

2.5 実験で用いるモノリンガルデータ

本実験では，対訳データの数を固定して，多量のモノリンガルデータをベースラインに加えて翻訳精度の変化を調べる．実験データとして，テスト文と同分野の辞書文と，別分野の特許翻訳文，Wikipedia から抜き出した文 (Wikipedia 文) の 3 種類を使用する．実験で使用するモノリンガルコーパスの内訳を表 1 に示す．

表 1 実験で使用するモノリンガルコーパス

追加コーパス	英語文	日本語文
辞書文	888,443 文	906,324 文
特許翻訳文	3,507,225 文	3,507,231 文
Wikipedia 文	14,679,468 文	12,893,649 文

2.6 評価方法

本実験の評価方法として，人手評価と自動評価を用いる．

2.6.1 人手評価

人手評価は出力文からランダムに 100 文を取りだし，追加コーパスとベースラインの対比較評価を行う．判断基準の例を以下に示す．

ベースライン○	ベースラインの出力文が追加コーパスありの出力文より翻訳品質が優れている
追加○ (辞書文)	辞書文を追加したコーパスの出力文がベースラインの出力文より翻訳品質が優れている
差なし	2 つの出力文の翻訳品質に明確な差がない
同出力	2 つの出力文が完全に同じ

2.6.2 自動評価

自動評価は，BLEU[7]，METEOR[8]，RIBES[9]，TER[10] を用いる．

3 実験結果

3.1 人手評価

日英翻訳における人手評価の結果を表 2 に，英日翻訳における人手評価の結果を表 3 に示す．

表 2 日英翻訳の人手評価

ベースライン○	追加○ (辞書文)	差なし	同出力
12	16	42	30

ベースライン○	追加○ (特許翻訳文)	差なし	同出力
9	11	35	45

ベースライン○	追加○ (Wikipedia 文)	差なし	同出力
13	6	51	30

表 3　英日翻訳の人手評価

ベースライン○	追加○ (辞書文)	差なし	同出力
7	14	58	21

ベースライン○	追加○ (特許翻訳文)	差なし	同出力
6	8	54	32

ベースライン○	追加○ (Wikipedia 文)	差なし	同出力
10	13	54	23

人手評価の結果より，辞書文を追加した日英翻訳・英日翻訳では，ベースラインと比べて翻訳精度が向上している．特許翻訳文と Wikipedia 文を追加した日英翻訳・英日翻訳では，翻訳精度に変化がなかった．

3.2　自動評価

日英翻訳の自動評価の結果を表 4 に，英日翻訳の自動評価の結果を表 5 に示す．

表 4　日英翻訳の自動評価

コーパスの内容	BLEU	METEOR	RIBES	TER
辞書文				
ベースライン	0.1382	0.4552	0.7105	**69.511**
追加あり	**0.1499**	**0.4641**	**0.7123**	70.004
特許翻訳文				
ベースライン	0.1299	**0.4521**	0.7051	73.437
追加あり	**0.1339**	0.4512	**0.7056**	**70.969**
Wikipedia 文				
ベースライン	0.1310	**0.4540**	**0.7046**	72.719
追加あり	**0.1389**	0.4536	0.7024	**71.893**

表 5　英日翻訳の自動評価

コーパスの内容	BLEU	RIBES	TER
辞書文			
ベースライン	0.1792	0.7686	**65.958**
追加あり	**0.1860**	**0.7690**	66.637
特許翻訳文			
ベースライン	0.1738	**0.7635**	68.097
追加あり	**0.1741**	0.7622	**67.415**
Wikipedia 文			
ベースライン	0.1792	**0.7694**	67.028
追加あり	**0.1815**	0.7644	**66.994**

自動評価の結果から，辞書文を追加した場合は TER 値を除いて，日英翻訳，英日翻訳ともに翻訳精度が向上している．一方特許翻訳文を追加した場合は，日英翻訳では BLEU 値，RIBES 値，TER 値が，英日翻訳では BLEU 値，TER 値の翻訳精度がわずかに向上している．Wikipedia 文を追加した場合は，BLEU 値と TER 値が日英翻訳・英日翻訳ともに翻訳精度がわずかに向上してる．

3.3　実験結果のまとめ

辞書文を追加したときの翻訳精度は，日英翻訳・英日翻訳ともに向上した．特許翻訳文と Wikipedia 文を追加したときの翻訳精度は，日英翻訳・英日翻訳ともにあまり向上しなかった．

4　データ量を変化させた実験

4.1　目的

追加実験として，辞書文と Wikipedia 文を使用し，追加する文を FULL のデータから 1/2，1/4 と減らしていき，モノリンガルデータの量と翻訳精度の関係を調べる．

4.2　実験内容

実験環境は第 2 章と同じである．新たな実験データとして追加するコーパスの 1/2，1/4，1/8 のデータを使用し，辞書文と Wikipedia 文で日英翻訳と英日翻訳を行う．

4.3　実験結果

日英翻訳における辞書文の人手評価の結果を表 6 に，Wikipedia 文の人手評価の結果を表 7 に，英日翻訳における辞書文の人手評価の結果を表 8 に，Wikipedia 文の人手評価の結果を表 9 に示す．日英翻訳の自動評価 BLEU の結果を図 1 に，METEOR の結果を図 2 に，RIBES の結果を図 3 に，TER の結果を図 4 に，英日翻訳の自動評価 BLEU の結果を図 5 に，RIBES の結果を図 6 に，TER の結果を図 7 に示す．

表 6　辞書文の日英翻訳の人手評価

ベースライン ○	+98,555 文 (+1/8 FULL) ○	差なし	同出力
7	10	36	47

ベースライン ○	+197,110 文 (+1/4 FULL) ○	差なし	同出力
7	12	46	35

ベースライン ○	+394,221 文 (+1/2 FULL) ○	差なし	同出力
12	15	46	27

ベースライン ○	+788,443 文 (+FULL) ○	差なし	同出力
12	16	42	30

表 7　Wikipedia 文の日英翻訳の人手評価

ベースライン ○	+1,822,433 文 (+1/8 FULL) ○	差なし	同出力
7	5	49	39

ベースライン ○	+3,644,867 文 (+1/4 FULL) ○	差なし	同出力
8	7	49	36

ベースライン ○	+7,289,734 文 (+1/2 FULL) ○	差なし	同出力
9	7	51	33

ベースライン ○	+14,579,468 文 (+FULL) ○	差なし	同出力
13	6	51	30

人手評価の表において，“+FULL”とは，追加できる最大のモノリンガルデータを，ベースラインのデータに追加した文である．“+1/2 FULL”とは，追加できる最大のモノリンガルデータの 1/2 を，ベースラインのデータ

に追加した文である．“+1/4 FULL”と“+1/8 FULL”も“+1/2 FULL”と同様である．

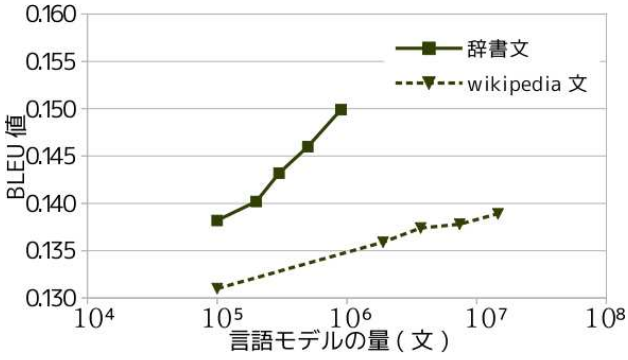

図 1　日英翻訳の BLEU 値の変化

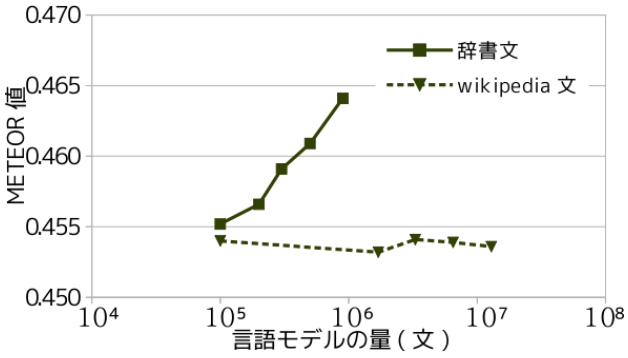

図 2　日英翻訳の METEOR 値の変化

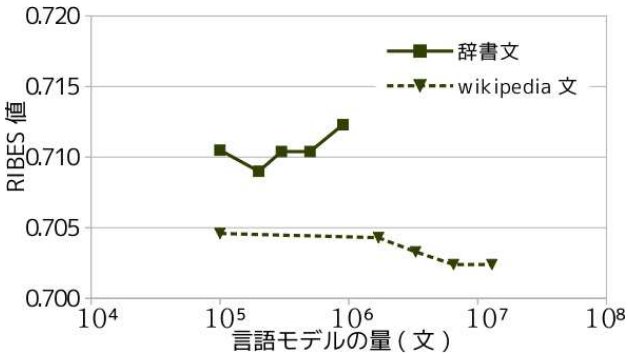

図 3　日英翻訳の RIBES 値の変化

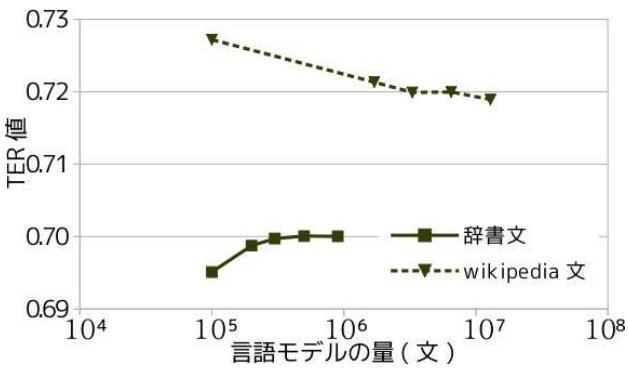

図 4　日英翻訳の TER 値の変化

表 8　辞書文の英日翻訳の人手評価

ベースライン ○	+100,790 文 (+1/8 FULL) ○	差なし	同出力
5	5	54	36
ベースライン ○	+201,581 文 (+1/4 FULL) ○	差なし	同出力
6	9	56	29
ベースライン ○	+403,162 文 (+1/2 FULL) ○	差なし	同出力
5	11	58	26
ベースライン ○	+806,324 文 (+FULL) ○	差なし	同出力
7	14	58	21

表 9　Wikipedia 文の英日翻訳の人手評価

ベースライン ○	+1,699,206 文 (+1/8 FULL) ○	差なし	同出力
9	12	47	32
ベースライン ○	+3,198,412 文 (+1/4 FULL) ○	差なし	同出力
8	13	50	29
ベースライン ○	+6,396,824 文 (+1/2 FULL) ○	差なし	同出力
10	10	54	26
ベースライン ○	+12,793,649 文 (+FULL) ○	差なし	同出力
10	13	54	23

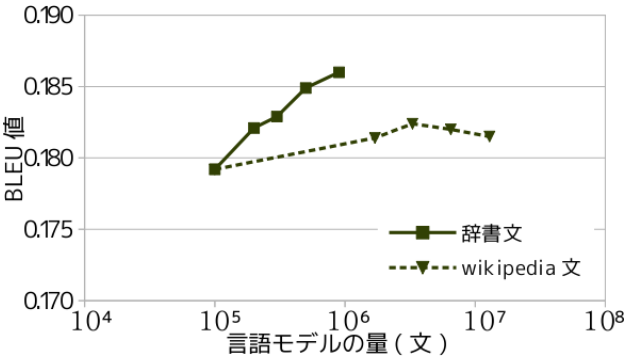

図 5　英日翻訳の BLEU 値の変化

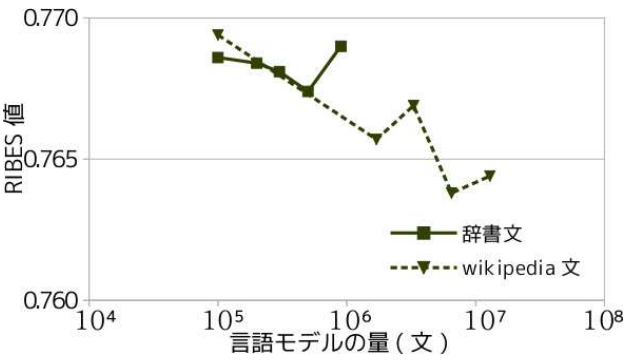

図 6　英日翻訳の RIBES 値の変化

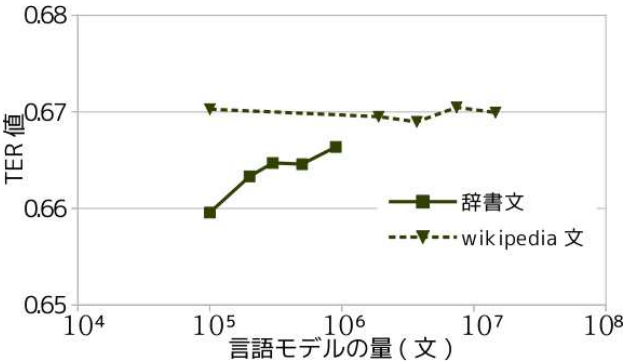

図 7　英日翻訳の TER 値の変化

実験結果より，人手評価では辞書文の翻訳精度が日英・英日翻訳ともに，比例して向上している．自動評価でも TER 値を除けば，辞書文の翻訳精度が日英・英日翻訳ともに，比例して向上している．

また，分野の違う Wikipedia 文では BLEU 値が日英・英日翻訳ともに，比例して向上している．しかし，その他の自動評価の結果では，ばらつきがある．

5　考察

5.1　分野の依存性

今回の実験では，日英翻訳・英日翻訳で翻訳精度において，同分野のモノリンガルデータを増やすと翻訳精度は向上する．しかし，別分野のモノリンガルデータを追加すると，翻訳精度はあまり向上しなかった．よって，統計翻訳における言語モデルの分野依存性は高い．

5.2　データ量と翻訳精度の関係

辞書文のデータを増やすと人手・自動評価において日英翻訳・英日翻訳の翻訳精度が，増加量と比例して向上している．よって，同分野のモノリンガルデータを収集することは有用であると考えられる．

5.3　人手評価の問題点

今回の実験での人手評価は，出力文 100 文に対して対比較を行ったが，翻訳品質に差のある文は 20 文ほどしかなかった．20 文で翻訳精度の差を評価するには信頼性が低い．よって評価する文数を増やす必要がある．

5.4　自動評価の問題点

統計翻訳における自動評価では，様々な問題が報告されている．松本は，単文を用いて自動評価と人手評価を行い，評価結果に差異が生じたことを報告している [11]．本実験でも自動評価の結果と人手評価の結果に差があった．今後，自動評価に対する方法を調べる必要がある．

6　関連研究

Brants らはアラビア語英語間の統計翻訳において，多量のモノリンガルデータを使用し，統計翻訳を行った [12]．4 種類のモノリンガルデータ，2 つの翻訳手法を用いて，BLEU 値が分野に関係なく向上することを報告した．また，Schwenk は仏英翻訳で，多量のモノリンガルデータを使用し，統計翻訳を行い，BLEU 値が上昇したことを報告している [13]．

本実験でも日英翻訳・英日翻訳での BLEU 値が同分野と別分野で向上している．しかし，他の自動評価の値はばらつきがある．上記の関連研究は，BLEU 値が向上したことのみを報告しており，他の自動評価の値は示されない．これは，BLEU の評価の問題点の 1 つであると考えている．

7　おわりに

本研究では英語-日本語間の単文の統計翻訳において，対訳データと比べて，データを収集することが容易な，モノリンガルデータを増やす実験を行った．

実験の結果，テスト文と同分野のモノリンガルデータを増やすと，日英翻訳・英日翻訳の翻訳精度が向上した．一方別分野のモノリンガルデータでは，日英翻訳・英日翻訳において翻訳精度があまり向上しなかった．

今後は重文複文コーパスや，統計翻訳の別手法での実験を行いたい．

参考文献

[1] Li et al.: “Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures”, LREC 2012, pp.1848-1855, 2012.

[2] Ruiqiang Zhang and Eiichiro Sumita:“Boosting Statistical Machine Translation by Lemmatization and Linear Interpolation”, Proceedings of the ACL 2007, pp.181-184, 2007.

[3] 村上仁一，藤波進:“日本語と英語の対訳文対の収集と著作権の考察”, 第一回コーパス日本語学会ワークショップ, pp.119-130，2012.

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

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

[6] moses, `http://www.statmt.org/moses/`

[7] Papineni et al.: BLEU,NIST: “a method for automatic evaluation of machine translation”, 40th Annual meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[8] Banerjee et al.: METEOR: “An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments”, Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), pp.65–72, 2005.

[9] 平尾 他: RIBES: “順位相関に基づく翻訳の自動評価法”, 言語処理学会第 17 年次大会発表論文集, pp.1111–1114, 2011.

[10] Richard Schwartz, Linnea Micciulla, John Makhoul. “A Study of Translation Edit Rate with Targeted Human Annotation”, AMTA, pp.223-231, 2006.

[11] 松本拓也, 村上仁一, 徳久雅人 “機械翻訳における人手評価と自動評価の考察”, NLP-2012，pp.505-508, 2012

[12] Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean, “Large Language Models in Machine Translation”, EMNLP-2007, pp.858-867, 2007.

[13] Holger Schwenk, “Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation”, Proceedings of IWSLT 2008, pp.182-189, 2008.