少量のパラレルコーパスと
大量のモノリンガルコーパスを用いた統計翻訳の精度調査

藤原勇　村上仁一　徳久雅人　村田真樹
鳥取大学工学部知能情報工学科
{s072046,murakami,tokuhisa,murata}@ike.tottori-u.ac.jp

1　はじめに

近年，機械翻訳において，統計翻訳が主流となっている．統計翻訳では，パラレルコーパスから自動的に翻訳規則を獲得し，翻訳を行う．そのため，統計翻訳における翻訳品質は，パラレルコーパスの量に大きく依存する [1]．しかし，パラレルコーパスの作成には，モノリンガルコーパスに比べて膨大なコストがかかる．

この問題を解決するために，先行研究 [2] では，大量のモノリンガルコーパスを，統計翻訳を用いて翻訳し，パラレルコーパスに付与した．しかし，翻訳精度の向上はほとんど認められなかった．これは，モノリンガルコーパスとその翻訳文すべてを用いたためであると考えられる．

そこで本研究では，モノリンガルコーパスと，精度の高い翻訳文の対を学習データに加えることで，翻訳精度の向上を目指す．また，対訳辞書データを補うため，“英辞郎”[3] を用いる．翻訳対の量が多い英辞郎のデータをパラレルコーパスに付与することで，統計翻訳の精度を向上させる．

2　提案手法

2.1　提案手法の概要

日英パラレルコーパスは，日本語と英語の対訳文のコーパスである．また，日本語学習文と，英語学習文はそれぞれの言語のモノリンガルコーパスである．本研究では，パラレルコーパスを増加させるため，はじめに，統計翻訳を用いて日本語学習文を翻訳する．次に，プログラムを用いて，翻訳文から精度の高い文を抽出する．そして，日本語学習文と抽出した文との対をパラレルコーパスに加える．

2.2　抽出方法

精度の高い文の抽出には，英語学習文から得た N-gram モデルを用いる．日本語学習文の翻訳文において，N-gram の尤度の高い文を抽出する．また，文の長さによる偏りを防ぐため，単語数で正規化を行う．抽出の際の閾値は，英語学習文 10 万文の尤度の平均値とする．

2.3　提案手法の手順

提案手法の手順を図 1 に示す．

準備　大量のモノリンガルコーパスとして，日本語学習文と英語学習文を準備する．また，統計翻訳に用いる日英パラレルコーパスと英辞郎のデータを準備する．

手順 1　日英パラレルコーパスと英辞郎から，フレーズテーブルを作成する．また，英語学習文から N-gram を作成する．

手順 2　作成したフレーズテーブルと N-gram モデルを用いて，日本語学習文を翻訳する．

手順 3　翻訳文から尤度の高い文を抽出し，日本語学習文と対訳文とする．これを “抽出文対” と呼ぶ．

手順 4　抽出文対を日英パラレルコーパスに付与し，新たなフレーズテーブルを作成する．

手順 5　手順 4 で作成したフレーズテーブルと，手順 1 で作成した N-gram を用いて日本語テスト文を翻訳する．

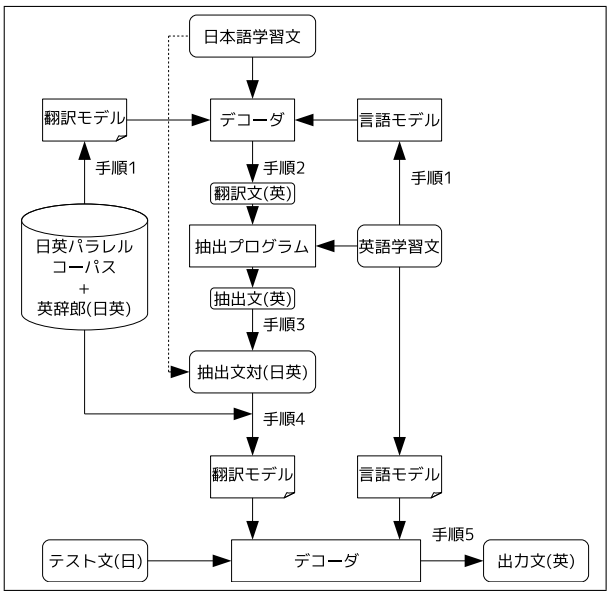

図 1　提案手法の枠組

3 実験環境

3.1 翻訳モデルの学習

翻訳モデルの学習には，“train-factored-phrase-model.perl[4]”を用いる．

3.2 言語モデルの学習

言語モデルの学習には，“SRILM[5]”の “ngram-count”を用いる．本研究では，N-gram モデルに 5-gram を用いる．

3.3 デコーダ

デコーダには，“moses[4]”を用いる．また，moses の各パラメータは “mert-moses.pl[4]”を用いて最適化する．しかし，“ttable-limit”と “distortion-limit”についてはパラメータチューニングでは変更されない．“ttable-limit”とは，1 つの日本語のフレーズに対して考慮する，目的言語のフレーズ数の制限である．また，“distortion-limit”とは，フレーズの並び替えの範囲の制限である．本研究では，“ttable-limit”の値を 60，また “distortion-limit”の値を-1（無制限）とする．

3.4 実験データ

3.4.1 英辞郎

“英辞郎”は EDP(Electronic Dictionary Project) がアップデートし続けている英和・和英データベースである．英辞郎のデータには，通常の英語辞書にない新しい語彙や複雑な言い回しも含まれている．英辞郎のデータを学習データに加えることで対訳辞書データを補完する．本研究では，不適切なデータを除去し，学習データとして用いるためにクリーニングした 1,366,575 文対 [8] を用いる．表 1 に，英辞郎のデータ例を示す．

表 1 クリーニング後の英辞郎データ例

catch on to ||| を理解する
catch on to ||| に気付く

3.4.2 単文コーパス

実験には，辞書の例文より抽出した単文コーパス 181,988 文 [6] から，以下のように用いる．

- 日英パラレルコーパス：50,000 文対
- 英語学習文：100,000 文
- 日本語学習文：100,000 文
- テスト文：10,000 文
- ディベロップメント文：2000 文（日本語学習文の翻訳に 1,000 文，テスト文の翻訳に 1,000 文）

統計翻訳の前処理として，各コーパスの日本語文に対して，“MeCab[7]”を用いて形態素解析を行う．また，英語文に対して “tokenizer.perl[4]”を用いて分かち書きを行う．

3.5 評価方法

出力文の評価において，自動評価法である NIST[9], BLEU[10], および METEOR[11] を用いる．また，人手評価として対比較評価を行う．

4 実験結果

4.1 自動評価

テスト文 10,000 文を入力文として翻訳実験を行い，出力文に対して自動評価を行った．表 2 に，それぞれの手法における自動評価の結果を示す．

表 2 中のベースラインとは，抽出文対（手順 3）をパラレルコーパスへ付与せずに，入力文を翻訳し，評価を行った結果である．また，提案手法において，パラレルコーパスに付与した抽出文対は 50,117 文対であった．

表 2 自動評価結果

	NIST	BLEU	METEOR
ベースライン	4.7198	0.1216	0.4990
提案手法	4.7200	0.1241	0.4999

結果より，提案手法の方がわずかに高い値を示しているが，ベースラインと提案手法には，ほとんど差が認められない．

4.2 対比較評価

4.2.1 評価結果

ベースラインと提案手法の出力文から，それぞれランダムに抽出した 100 文を用いて，人手による対比較評価を行った．評価の基準を以下に示す．また，評価結果を表 3 に示す．

- 提案手法○　：提案手法の方が良い
- 提案手法×　：提案手法の方が悪い
- 差なし　：双方とも意味が分からない，または，意味に差がない
- 同一出力　：完全に同じ文が出力されている

表 3 対比較評価

提案手法○	提案手法×	差なし	同一出力
8/100	12/100	50/100	30/100

結果より，人手評価において，提案手法の有効性は認められなかった．

4.2.2 翻訳例

提案手法○，提案手法×，差なしの場合の翻訳例を表 4 から表 9 に示す．

表 4 において，ベースラインに動詞がなく，提案手法には動詞があるため，提案手法○とした．

表 4 提案手法○の翻訳例

入力文	警官 が 交通 整理 を し た 。
正解文	The police kept a clear passage for the traffic .
ベースライン	The policeman the traffic .
提案手法	A policeman is directing the traffic .

表 5 において，ベースラインに主語がなく，提案手法には主語があるため，提案手法○とした．

表 5 提案手法○の翻訳例

入力文	関節 が 痛む 。
正解文	I ache in my joints .
ベースライン	joint aches .
提案手法	I have a pain in joint .

表 6 において，提案手法の翻訳文の意味が入力文と異なるため，不適切である．よって，提案手法×とした．

表 6 提案手法×の翻訳例

入力文	物理 学 の 勉強 に は 数学 の 十分 な 知識 が 必要 で ある 。
正解文	The study of physics demands a good knowledge of mathematics .
ベースライン	The physics is necessary to the study of sufficient knowledge of mathematics .
提案手法	He is necessary to the study of physics sufficient knowledge of mathematics .

表 7 において，時制がベースラインの方が正しいため，提案手法×とした．

表 7 提案手法×の翻訳例

入力文	気温 が ちょっと 上がっ た 。
正解文	The temperature rose a little .
ベースライン	The temperature went up a little .
提案手法	The temperature is going up a little .

表 8 において，どちらの翻訳文も意味を成さないため，差なしとした．

表 8 差なしの翻訳例

入力文	学会 で 研究 を 発表 する 。
正解文	Present one's research at the conference .
ベースライン	read at the meeting .
提案手法	I read .

表 9 において，plenty of time と enough time が入力文に対してどちらでも正しいといえるので，差なしとした．

表 9 差なしの翻訳例

入力文	時間 は まだ 十分 ある 。
正解文	There is still plenty of time left .
ベースライン	There is still plenty of time .
提案手法	There is still enough time .

5 考察

5.1 抽出文対の精度

追加実験 1 として，日本語学習文と，正しい対訳文の対 100,000 文対をパラレルコーパスに加えた場合の結果を表 10 に示す．

表 10 追加実験 1

	NIST	BLEU	METEOR
ベースライン	4.7198	0.1216	0.4990
追加実験 1	5.2830	0.1562	0.5364

結果より，正しい対訳文対を学習データに加えると，評価値は大きく向上する．しかし，提案手法では翻訳精度の向上はほとんど認められなかった．この結果から，抽出文対の精度が不十分であると考えている．抽出文対には表 11 に示すような，誤りのある文が含まれている．誤りのある抽出文対を学習データとして用いた場合に，翻訳精度が下がると考えられる．したがって今後は，より精度の高い文の抽出方法を検討する必要がある．

表 11 抽出文対の例

入力文	金 の 価値 が 上昇 し た 。
抽出文	The value of money .
入力文	こんな 品 が 手 に 入っ た 。
抽出文	I can't work .

5.2 誤りデータの影響

誤りのある文の割合が，学習に及ぼす影響を調査するため，抽出の際の尤度を調整し，追加実験 2 を行った．表 12 に抽出文対数を 10,000 文対，20,000 文対，40,000 文対，80,000 文対，100,000 文対（抽出なし）とした場合の自動評価の結果を示す．抽出文対数が多いほど誤りのある文の割合が高く，抽出文対数が少ないほど誤りのある文の割合が低い．ただし，追加実験 2 において，デコーダのパラメータによる評価結果のばらつきをなくすため，パラメータチューニングは行っていない．

表 12 追加実験 2

抽出文対数	NIST	BLEU	METEOR
ベースライン	3.6217	0.0968	0.4407
10,000	3.5813	0.0975	0.4405
20,000	3.5034	0.0952	0.4362
40,000	3.3314	0.0917	0.4276
80,000	2.9945	0.0826	0.4122
100,000	2.8814	0.0816	0.4062

結果より，学習データに誤りのある文がより多く含まれるほど，評価が下がることが確認できる．また，この結果から，尤度を用いた抽出の有効性も示された．

5.3 モノリンガルコーパスの量

5.2 節において，尤度を高く設定し，抽出文対数を少なくすれば，誤りのある文の割合が減少し，評価結果が良くなることが示された．しかし，パラレルコーパスに付与する抽出文対数が少ないと，学習に与える影響も小さい．したがって，より多くの日本語学習文の翻訳文から抽出を行えば，パラレルコーパスに付与する際に，精度の高い対訳文が増加し，提案手法の翻訳精度が向上すると考えられる．モノリンガルコーパスの収集は比較的容易に行うことができる．よって今後は，提案手法において，より大量のモノリンガルコーパスを用いた場合の，翻訳精度の調査を行う．

5.4 英辞郎の効果

提案手法において，学習データに英辞郎を用いた場合と，用いない場合における出力文中の未知語数を表 13 に示す．また，それぞれの場合における自動評価の結果を表 14 に示す．なお，パラレルコーパスに付与した抽出文対はそれぞれ約 50,000 文対である．

表 13 出力文中の未知語数

	未知語数
提案手法	1520
英辞郎なし	5587

表 14 自動評価結果

	NIST	BLEU	METEOR
提案手法	4.7200	0.1241	0.4999
英辞郎なし	4.3322	0.1125	0.4730

結果より，学習データに英辞郎を加えることで，未知語が減少し，翻訳精度が向上している．したがって，本研究における英辞郎の有効性が確認できる．

6 おわりに

本研究では，モノリンガルコーパスの翻訳文から精度の高い文を抽出し，パラレルコーパスに加えることで，パラレルコーパスを増加させる手法を提案した．実験の結果，翻訳精度の向上がほとんど認められなかった．原因として，抽出された文対のなかに，精度の低い文が含まれていたことが挙げられる．今後は，より精度の高い翻訳文の抽出方法を検討する．また，より大量のモノリンガルコーパスを用いて，翻訳精度を向上させる方法を検討する．

参考文献

[1] 猪澤雅史，村上仁一，徳久雅人，池原悟，“統計翻訳における単文・重文複文の翻訳精度の評価”，言語処理学会第 14 回年次大会, pp.869-872，2008.

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

[3] 英辞郎 http://www.alc.co.jp/.

[4] Moses：moses.2009-04-13.tgz
http://www.statmt.org/moses/.

[5] SRILM(The SRI Language Modeling Toolkit) : srilm.tgz
http://www.speech.sri.com/projects/srilm/.

[6] 西山七絵，村上仁一，徳久雅人，池原悟，“単文文型パターン辞書の構築”，言語処理学会第 11 回年次大会，pp.372-375,2005.

[7] MeCab：mecab-0.97.tar.gz，mecab-ipadic-2.7.0-20070801.tar.gz
http://mecab.sourceforge.net/.

[8] 東江恵介, 村上仁一, 徳久雅人, 池原悟，“日英統計翻訳における英辞郎の効果”, 言語処理学会第 16 回年次大会発表論文集, pp.641-644, 2010.

[9] NIST, Automatic Evalation of Machine Translation Quality Using n-gram Co-Occurence Statistics
http://www.itl.nist.gov/iad/mig/test/mt.

[10] BLEU，NIST Open Machine Translation (OpenMT) Evaluation
http://www.itl.nist.gov/iad/mig/tests/mt/.

[11] The METEOR Automatic Machine Translation Evaluation System
http://www.cs.cmu.edu/~alavie/METEOR/.