統計的手法を用いた双方向方言機械翻訳システム

柴田　直由†　　横山　晶一††　　井上　雅史††
†,†† 山形大学大学院理工学研究科情報科学専攻
〒992-8510 山形県米沢市城南4-3-16
E-mail:†tfw98479@st.yamagata-u.ac.jp, ††{yokoyama, mi}@yz.yamagata-u.ac.jp

1. はじめに

方言とは、地域社会で使用されている日常生活語であり、その地域に住む人々との円滑な会話やよりよい人間関係を築くためには必要不可欠とされている[1]。この方言を理解する手段の一つとして、方言機械翻訳システムを用いた方言学習があげられる[2]。言語学では、方言学という学問体系が確立しており[3-5]、音韻やアクセント、語彙、文法など、様々な面から研究が行われている[6]。

方言機械処理の研究目的は以下の通りである。

- 方言は、語彙や文法の面で古い日本語を保持していると言われており[1]、規則的な解明が現代語に対する示唆を与えることが期待される
- 方言は主に話し言葉として伝えられてきたため、方言文法を作成することは、話し言葉文法を作成することに寄与する
- 方言話者とのより良い人間関係を築くための、有効な手段となる可能性がある

以上の考えから、我々は方言を電子的に扱う研究に着手した[6]。その後、過去約10年にわたりルールベース機械翻訳手法(以後 RBMT)を用いて共通語から方言への機械翻訳システムを構築してきた[7]。しかし、詳細なルールと語彙辞書の作成において、時間的コストと方言に関する多量の知識が必要不可欠となる問題がある。山形(村山)方言については方言の詳細な文法や語彙に関する文献がほぼ皆無であり、ルールや語彙辞書の作成に多大な時間をかけてきたが精度は余り上昇していない。特に、逆方向である方言から共通語への機械翻訳に関しては方言文法の構築に問題がある。

以上を踏まえ、本研究では、文法やルールを統計量を用いて算出し翻訳を行う、句ベース統計的機械翻訳手法(以後 PB-SMT)[8]を用いることで、低コストで方言の知識がなくとも十分な精度が出る双方向(共通語から方言および方言から共通語へ)の方言機械翻訳システムの構築を目指した。また、方言自体のリソースがほぼ皆無であることが本研究での大きな問題となるが、ある程度文の誤りを許した共通語-方言対訳コーパスを作成することで、どの程度方言機械翻訳が省力化可能かの実験を行った。ここでは、共通語を日本語共通語、方言を山形村山方言とする。

2. 関連研究

2. 1. 句ベース統計的機械翻訳(PB-SMT)[8]

PB-SMT とは、ソース言語文 **f** が与えられたとき、全ての組み合わせの中から確率が最大になるターゲット言語文$\hat{\mathbf{e}}$を探索し、翻訳を行う手法である。

$$\hat{\mathbf{e}} = \mathrm{argmax}_{\mathbf{e}}\, P(\mathbf{e}|\mathbf{f}) = \mathrm{argmax}_{\mathbf{e}}\, P(\mathbf{f}|\mathbf{e})P(\mathbf{e}) \quad (1)$$

(1)式の $P(\mathbf{f}|\mathbf{e})$は翻訳モデル、$P(\mathbf{e})$は言語モデルを表し、$\hat{\mathbf{e}}$が最大となるような **e** を求めることをデコーディングという。PB-SMT では、**f** は連結された句$\bar{\mathbf{f}}_1 \ldots \bar{\mathbf{f}}_I$としてみなし、句$\bar{\mathbf{f}}_i$ は句$\bar{\mathbf{e}}_i$ に翻訳される。特に、言語モデルでは *N*-gram モデルが一般的に使用され、翻訳モデルではフレーズテーブルと呼ばれる表によって管理される(表1)。

PB-SMT は、自然言語の文法やルールに関する情報を統計量を用いて算出するため、これらの知識がなくとも翻訳が可能であるという利点がある。この手法により、文法やルールの作成コストの削減と、知識量不足のカバーを行うことが可能であると考え、本研究ではこの手法を共通語から方言、方言から共通語の双方向で取り入れた。

表1：フレーズテーブルの例
(共通語:**c**、方言:**d** としたとき)

c フレーズ ||| **d** フレーズ ||| **d-c** 方向の翻訳確率 $P(\mathbf{c}|\mathbf{d})$ **d-c** 方向の単語の翻訳確率の積 **c-d** 方向の翻訳確率 $P(\mathbf{d}|\mathbf{c})$ **c-d** 方向の単語の翻訳確率の積
て 働い て ||| で 働い で ||| 0.0909091 0.357285 1 0.874507
て 出 なけれ ば なら ない ||| て 出 ら ん なね ||| 1 0.000135036 1 0.000347276

2. 2. リソースの乏しい言語の機械翻訳

リソースが乏しい言語の機械翻訳に関して、以下のような様々な手法が提案されているが、次の理由により利用できない。

- ピボット翻訳を用いた(PB-)SMT[9]
  - 方言文を含む対訳コーパスが存在しないため、利用できない

- Word-Net、WordLattice を用いたコーパス等の拡張・生成[10-11]
  - ✧ 方言語彙の Word-Net が存在せず、方言語彙、またはその語彙同士の関係の電子化に多大な時間的コストと知識が必要となり、本研究の趣旨に反する
- コンパラブルコーパス、あるいは非対訳コーパスより疑似翻訳確率を用いた(PB-)SMT[12-13]
  - ✧ 電子化された方言文が存在しない、またはごく微量であるため、利用が難しい
- 音訳、言い換えを用いた(PB-)SMT、ピボット翻訳またはコーパス等の拡張・生成[14-15]
  - ✧ 方言の音素化が難しく、言い換え規則作成に多大な時間的コストがかかり、本研究の趣旨に反する

3. 提案手法

3. 1. 提案手法について

今回、言語資源がない中でも、ある程度文の誤りを許した完全ではない共通語-方言対訳コーパスを作成した。この完全ではない対訳コーパスを用いた PB-SMT を用いても、ある程度十分(RBMT と同様)な翻訳精度を出すことが可能であれば、ルールや文法を人手で書くための時間的コストを減らし、方言に関する人間の知識・文献数が非常に少なくともこれをカバーすることができる。また、他の方言でも同精度の適当な対訳文を用意することである程度十分な精度で翻訳を行うことが可能になる。

3. 2. 共通語-方言対訳コーパスの作成方法

コーパス作成の際、ProjectGutenberg や青空文庫、プロジェクト杉田玄白などの作品について、日本語文と英語文の対訳文対応がつけられた、日英対訳文対応付けデータ[16](以後日英対訳コーパス)を用い、以下の手順で作成した。

① 日英対訳コーパスより、単語数が80単語以内の文(重文、複文を含む)で構成されている日本語102,766文を抽出し、これを共通語文と呼ぶ
② 抽出した共通語文に対して、KyTea(0.4.2)[17]を用いて単語分割を行う
③ ②に対し、柴田ら(2011)[7]の共通語から方言へのルールベース機械翻訳システム(BLEU 精度約63point)を用いて、単語分割済み方言文を生成する
④ ③の方言文と①の共通語文を組み合わせ、これを、ある程度文の誤りを許した共通語-方言対訳コーパスとする

ただし、方言文が正しい文ではない箇所がある不完全な対訳コーパスとなっているため、今回この対訳コーパスに対して、約8.5%(8,773/102,766)の文を確認し5,551文の修正を施した。日英対訳コーパスを使用した理由は、共通語-方言間の翻訳と日-英間翻訳との翻訳精度の違いについて考察可能であると考えたからである。単語分割に KyTea を使用した目的は以下の通りである。

- 本研究は、形態素情報を使わず単純に PB-SMT を使用するため
- 方言単語を分割する際に用いる方言単語分割モデルを、自動的に学習・構築することが可能なため

共通語-方言対訳コーパスの作成概要図を以下の図1に、共通語-方言対訳コーパスの例を表2に示す。

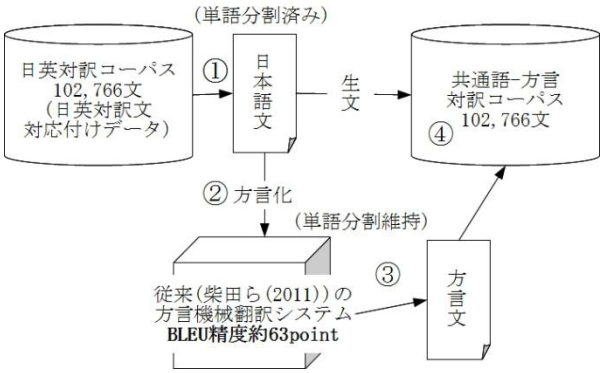

図1：共通語-方言対訳コーパスの作成概要図

表2：共通語-方言対訳コーパスの例

共通語1	荷物　を　持っ　て　くれ　。
方言1	荷物　ば　たがっ　て　けろ　。
・・・〜・・・	
共通語 n	学校　に　行き　たい　。
方言 n	がっこ　さ　えぐ　だい　。

4. 双方向方言機械翻訳システム

4. 1. 共通語→方言

入力された共通語文を、自動単語分割器 KyTea (高性能 SVM モデル(精度約98%))を用いて単語分割し、単語分割された文を PB-SMT 手法を用いて翻訳を行う。システム概要図を図2に示す。

入力 (共通語文)	荷物を持ってくれ。
↓(KyTea)	
単語分割 (共通語文)	荷物　を　持っ　て　くれ　。
↓(Moses)	
翻訳 (方言文)	荷物　ば　たがっ　て　けろ　。

図2：共通語から方言への機械翻訳システムの概要図

4. 2. 方言→共通語

4. 1と同様、入力された方言文を、方言単語分割モデルを学習させた KyTea を用いて単語分割し、単語分割された文を PB-SMT 手法を用いて翻訳を行う。システム概要図を図3に示す。

入力 (方言文)	荷物ばたがってけろ。
↓(KyTea)	
単語分割 (方言文)	荷物　ば　たがっ　て　けろ　。
↓(Moses)	
翻訳 (共通語文)	荷物　を　持っ　て　くれ　。

図3：方言から共通語への機械翻訳システムの概要図

5. 実験

5. 1. 実験環境

翻訳に関して、単語モデル生成に SRILM(1.5.12)、翻訳モデル生成に GIZA++(1.0.5)、デコーダに Moses(2010-04-01)を使用した。実験に用いるシステムの概要図を、図4に示す。この図は、方言から共通語への PB-SMT に関する図であるが、共通語から方言への翻訳もこれとほぼ同じ手順を踏む。相違点は、対訳コーパスから単語分割モデルを学習させる部分と、共通語文と方言文の部分がそれぞれ入れ替わるという部分である。

細かいオプションの設定について、N-gram モデルは 4-gram、スムージングは Kneser-Ney スムージング、distortion-limit は初期の6を用いた。また、moses.ini のパラメータチューニングに関しては、時間的都合と条件の均一化を理由に今回は行っていない。

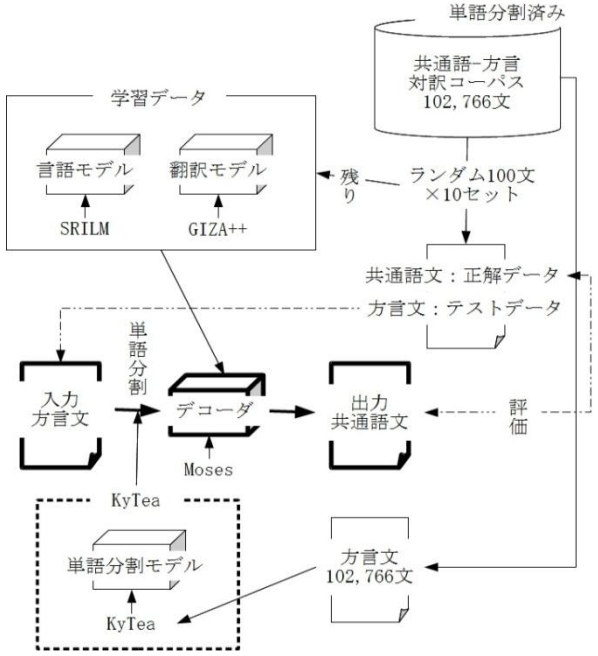

図4：実験に用いるシステム概要図

5. 2. 実験と評価方法

実験は、共通語から方言、方言から共通語の精度をそれぞれ BLEU[18]、RIBES[19]、WER(Word Error Rate)で算出し、方言の単語分割器の分割精度を F 値で算出する。実際は、学習コーパス量を5,000、10,000、50,000、100,000文、確認・修正済みの8,773文と文量を変えながら実験を行い、それぞれ、100文10通り(1,000文)の正解データでもって精度の平均値、分散をとる。いずれも文はランダムで抽出し、特に正解文については学習データよりも先に抽出し固定データとする。また、ベースラインは c2d 方向のみ c2d 方向の RBMT の精度とする。

5. 3. 実験結果と考察

表3,4に、それぞれの方向の PB-SMT の評価結果と、KyTea で学習させた方言単語分割モデルの精度(F 値)を示した。以下考察である。

今回の実験に関して、RBMT を前処理とした SMT[20]手法に近いことを行っているため、c2d 方向においては50,000文以上ではベースラインのスコアを上回る結果となった。確認修正済みの文のみで学習させた場合のスコアは100,000文と比べても高い結果となった。そのため、c2d 方向においては学習量よりも質のよい対訳データを用意するほうがよい、ということがわかる。

d2c 方向では、全体的にスコアが高くなる結果となったが、確認修正済みデータでの結果を見ると c2d 方向のデータのように高いスコアにはならず、学習文数とスコアの推移に沿っているように見える。これは、おそらく対訳コーパス中の方言文の、RBMT で補い切れない残り約40%の文が共通語として存在しているためであると考えられる。つまり、学習の際に共通語 n を方言 n としてではなく共通語 n のままアラインメントがとられ、方言へ翻訳する際、本来方言 n と翻訳されるところが誤って共通語 n のままに翻訳されるため精度が落ち、共通語へ翻訳する際は、偶然にもこの部分がそのまま共通語 n として翻訳されるので精度が上がる、ということである。そのため、確認修正済みの文で学習した場合、d2c 方向のスコアが c2d ほど上昇しなかったと考えられる。

表3：c2d 方向の PB-SMT 評価比較表

PB-SMT(c2d)			
学習文数	**BLEU**	**RIBES (×100)**	**WER(%)**
ベースライン	63.62	94.63	24.69
5,000	46.98	91.55	38.24
10,000	47.69	91.69	37.26
50,000	63.89	94.26	24.17
100,000	63.90	94.22	24.11
8,773 (確認修正済み)	75.62	95.27	18.93

表4：d2c 方向の PB-SMT 精度比較表と
KyTea による方言単語分割モデルの精度比較表

PB-SMT(d2c)				
学習文数	**BLEU**	**RIBES (×100)**	**WER(%)**	**F 値**
5,000	53.86	93.28	36.52	58.80
10,000	55.28	93.69	34.21	61.51
50,000	78.40	97.20	17.20	66.18
100,000	80.89	97.41	15.78	68.89
8773 (確認修正済み)	74.31	96.68	20.44	62.49

6. おわりに

本実験より、方言文が正しい文ではない箇所がある不完全な対訳コーパスを用いても、RBMT で作成したシステムと同精度以上で翻訳することが可能であることが示された。そのため、方言に関する知識がなくとも、ある程度文が間違っているにせよ対訳コーパスを用意することで、共通語-方言間の機械翻訳システムの構築が可能になり、さらにルールや文法を人手で記述するよりも低コストで実現可能であることが確認できた。

また、共通語文と方言文の単語4-gram までの一致度は約36%であるので、この程度、またはそれ以上の似通った他方言であれば上記のことがいえるのではないかと考えられ、さらに、言語資源が少ない他言語の機械翻訳に関しても示唆可能と考えられる。

今後は、今回実験に選んだ山形村山方言ではなく、他方言や他言語、他言語の方言についても本研究が有効であるかを調査することを考えている。

[参考文献]

[1] 飯豊毅一 ほか(編), 方言概説・講座方言学1,国書刊行会, 1998.

[2] 三間優, “方言学習と方言変換ソフトの改善方法の考察,” 新潟工科大学情報電子工学科卒業論文, 2008.

[3] 藤原与一, 方言学の方法, 大修館書店, 1977.

[4] 北原保雄 監修、江端義夫 編, 方言・朝倉日本語講座10, 朝倉書店, 2002.

[5] 佐々木冠 他, 方言の方法・シリーズ方言学2,岩波書店, 2006.

[6] 横山晶一、安野克彦, “方言の機械処理に関する予備的考察,” 電子情報通信学会技術報告 NLC94-45, pp.39-46, 1995.

[7] 柴田直由、横山晶一、井上雅史, “年代差を考慮した方言翻訳システム,” 自然言語処理研究会報告2011-NL-202(7), pp.1-8, 2011.

[8] Philipp Koehn, “Word-BasedModels”and”Phrase-Based Models,” in Statistical Machine Translation, pp.81-149, Cambridge University Press, 2010.

[9] Hua Wu, Haifeng Wang, “Pivot language approach for phrase-based statistical machine translation,” Mach Translat(2007)21, pp.165-181, 2007.

[10] Istvan Varga and Shoichi Yokoyama “Dictionary generarion for less-frequent language pairs using WordNet,” Literary and Linguistic Computing, Vol.24, No.4, pp.449-466, 2009.

[11]Nguyen Manh Hung、秋葉友良, ”Word Lattice Decoding を利用した対訳コーパスのない言語からの統計的機械翻訳,” 言語処理学会第16回年次大会発表論文集, pp.1006-1009, 2010.

[12]Hiroyuki Kaji、Takashi Tsunakawa、Daisuke Okada, “Using Comparable Corpora to Adapt a Translation Model to Domains,” In Proceedings of the 7th International Conference on Language Resources and Evaluation, pp.2182-2188, May2010.

[13]岡田大輔、綱川隆司、梶博行, ”非パラレルコーパスを用いた統計的機械翻訳の分野適応,” 言語処理学会第16回年次大会発表論文集, pp.1018-1021, 2010.

[14]Amit Kirschenbaum、Shuly Wintner, “Lightly Surpervised Translitetation for Machine Translation,” Proceedings of the 12th conference of the European Chapter of the ACL, pp.433-441, 30 March – 3 Aplil 2009.

[15]Fujita Atsushi、Pierre Isabelle、Roland Kuhn, “Enlarging Paraphrase Collections through Generalization and Instantiation,” Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.631-642, July 2012.

[16]内山将夫, 高橋真弓, 日英対訳文対応付けデータ, http://www2nict.go.jp/univ- com/multi_trans/member/mutiyama/align/index.html , 2003.

[17]Graham Neubig、中田陽介、森信介, “点推定と能動学習を用いた自動単語分割器の分野適応,” 言語処理学会第16回年次大会発表論文集, pp.912-915, 2010.

[18]Papineni Kishore, Salim Roukos, ToddWard, Wei-Jing Zhu, ”BLEU: a method forautomatic evaluation of machinetranslation,” 40th Annual meeting ofthe Association for ComputationalLinguistics, pp. 311-318, 2002.

[19]平尾努、磯崎秀樹、Kevin Duh、須藤克仁、塚田元、永田昌明, “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回年次大会発表論文集 pp.1115-1118, 2011.

[20]福田智大、上村仁一、徳久雅人、池原悟, “ルールベース翻訳を前処理に用いた統計翻訳,” 言語処理学会第16回年次大会発表論文集, pp.672-675.