Vol.2017-NL-233 No.2
2017/10/24

事前学習と汎化タグによる方言翻訳の精度向上

長谷川 駿[1,a)]　田中 駿[2,b)]　山本 悠二[3,c)]　高村 大也[1,d)]　奥村 学[1,e)]

概要：標準語の文を方言に翻訳する方言翻訳では，大規模な対訳データを用意することが困難である．小規模な対訳データから学習する方法として，対訳データによる学習の前に単一言語データを用いてAutoEncoderによる事前学習を行う手法が提案されている．しかし，AutoEncoderは入力列と参照列が同一であり，入力列において直前の出力単語の次にある単語さえ分かれば正しく学習できてしまう．そのため注視機構付き系列変換モデルの事前学習としてAutoEncoderを用いると，入力列においてどの単語の次にどの単語があるかが重要な情報として学習されてしまい，文法的な単語の並びが学習されるとは限らない．そこで本研究では，品詞ごとに方言翻訳の性質が違うことに着目した単一言語データによる事前学習を行い，さらに汎化タグを用いてデータを抽象化し研究課題の複雑さを減らすことで方言翻訳モデルの精度向上を目指す．

キーワード：方言翻訳，翻訳，深層学習，事前学習，汎化タグ

1. はじめに

方言は，限られたコミュニティでのみ使用される話し方であり，その地域の住民との円滑な会話や，より良い人間関係を築くためには欠かせない要素である．また，方言には古い日本語の言語的な性質が保持されていると言われており [1]，方言の言語的な性質を解明することは現代語に対する示唆を与えると期待されている．そこで，方言の言語的な性質を解明するため, また，方言を様々なアプリケーションに利用するために標準語の文を方言に翻訳する方言翻訳の研究が行われてきた．本研究では日本語の方言翻訳を対象に研究を行う．

方言翻訳は単一言語内翻訳タスクの一つである．現在，多くの翻訳タスクで高い性能を出している手法は，大規模な対訳データを用いて注視機構付き系列変換モデルを学習する手法である．しかし，日本語の大規模な方言翻訳の対訳データは存在しておらず，人手で作成することも困難であるため，方言翻訳では小規模な対訳データで学習を行わなければならない．

では，そもそもなぜ多くの翻訳タスクで大規模な対訳データを必要としているかを考える．理由として挙げられるのは (1) 多くの単語がある中で単語列から単語列を予測するため，研究課題が煩雑で学習が難しいから， (2) 難しい課題でも学習可能な表現力の高いモデルを用いるため，小規模な対訳データでは過学習してしまうから，である．つまり，これらを軽減することができれば小規模な対訳データからでも翻訳器の学習が可能となる．そこで本研究では，日本語の方言翻訳の特徴に着目して，上記 2 点の問題点を緩和する．本研究で着目する，日本語方言翻訳の一つ目の特徴は，翻訳しても変化しない単語が多数見られることである．二つ目の特徴は，翻訳時に単語の並び替えがほぼ起きないということである．

表 1　方言翻訳の例文

例	標準語	関西弁
1	学生 です	学生 や
2	今日 も 見 た よ	今日 も 見 た で ぇ

まず，表 1 の例 1 では “学生”は変化していない単語であり，“学生”を他の一般名詞に置き換えても多くの場合で正しい翻訳である．もし，同様に多くの翻訳例で一般名詞を他の一般名詞に置き換えても正しい翻訳であるのであれば，全ての一般名詞を一つの汎化タグに置き換えることで方言翻訳に必要な表現力をあまり失うことなくデータを抽象化することができ，研究課題が煩雑になるのを避けることができる．そこで本研究では研究課題が煩雑になるのを

[1] 東京工業大学
[2] 株式会社サイバーエージェント AI Studio AI Lab
[3] 株式会社サイバーエージェント 技術本部 秋葉原ラボ
[a)] hasegawa.s@lr.pi.titech.ac.jp
[b)] tanaka_shun_xa@cyberagent.co.jp
[c)] yamamoto_yuji_xa@cyberagent.co.jp
[d)] takamura@pi.titech.ac.jp
[e)] oku@pi.titech.ac.jp

避けるため，一般名詞に限らず様々な単語を汎化タグに置き換えることで出来るだけ方言翻訳に必要な表現力を落とさずにデータを抽象化する．複数の汎化タグが同一文に含まれている場合でも，翻訳時に単語の並び替えがほぼ起きないという方言翻訳の特徴から，入力と同じ数，順番で汎化タグが出力されるようにすることで容易に元の単語に復元することができる．

次に，小規模な対訳データでモデルを学習すると過学習しやすい問題の対処について考える．一般的に知られている対処法は，事前学習と呼ばれる，大規模な単一言語データや近いタスクの大規模データを用いてあらかじめモデルを学習しておくことで過学習が起きにくくするというものである．方言翻訳でも，標準語の単一言語データを用いたAutoEncoder[2] による事前学習を行なった後，小規模な対訳データによる学習を行う手法が提案されている [3]．しかし，AutoEncoder は入力列と参照列が同一であり，入力列において直前の出力単語の次にある単語さえ分かれば正しく学習できてしまう．そのため注視機構付き系列変換モデルの事前学習として AutoEncoder を用いると，文法的な文となるためにはどの単語が出力されるべきかではなく，入力列においてどの単語の次にどの単語があるかが重要な情報として学習されてしまい，文法的な単語の並びが学習されるとは限らない．そこで本研究では，表 1 の例 2 の “今日”，“見”のように内容語（名詞，動詞，形容詞）は翻訳時にあまり変化しないことに着目し，内容語のみから元の文を予想するよう方言の単一言語データを用いて事前学習を行う．これにより，内容語（入力からあまり変化しない単語）は入力から情報を得て出力され，内容語以外（変化して出力されるべき単語）は入力での周辺の情報と方言としての文法的な単語の並びに基づいて出力されるよう学習され，対訳データによる学習が容易になると考えられる．

2. 関連研究

本研究で対象とする日本語の方言翻訳では，対訳データを用いて学習した重み付き有限状態トランスデューサを用いる手法 [4]，統計的機械翻訳技術を用いる手法 [5]，AutoEncoder による事前学習を行った後に対訳データによる学習を行う深層学習を用いた手法 [3] 等が提案されている．また，様々な単言語内翻訳タスクのための手法も関連研究としてあげられる．例えば，要約タスク [6] や文の平易化 [7] も単言語内翻訳タスクである．だが，要約タスクや文の平易化では入力と出力の意味が方言要約のように一致せず，文の一部が省略されたり，簡単な表現に置き換えるために抽象化されて出力されるため，方言翻訳とは異なる性質を持った研究課題である．

事前学習手法としては，単一言語データで学習した word2vec[8] をモデル中の単語の分散表現に用いる手法や，モデル全体の事前学習として AutoEncoder を用いる手法 [2] が広く知られている．汎化タグによるデータの抽象化では，表データや数値データから説明文を生成する Data-To-Text と呼ばれる研究課題で，役割の明白な低頻度語を汎化タグに置き換える手法が提案されている．例えばバスケットボールの試合を対象とした場合 [9] には選手名やチーム名を，株価情報を対象とした場合 [10] には統計指標の上がり幅や株価などの数値を汎化タグに置き換えている．

また，要約文生成タスクでは，本研究の着眼点の 1 つに似ている “入力列に含まれる単語が出力されやすい” という性質に着目して入力列の単語を出力するコピー機構 [11] という手法が提案されている．

3. 汎化タグを用いた抽象化

研究課題が煩雑になるのを避けるため，データ中の単語を汎化タグに置き換えデータを抽象化する．理想的には，方言翻訳に必要な表現力を全く失わずにデータを抽象化できると良い．だが実際には，1 節で例に出した一般名詞でさえ変化が全く起こらないというわけではない（例：「鶏肉」は主に西日本で「かしわ」と言い換えられている．）．そのため，変化が起きづらく，かつ周囲の翻訳に影響を与えにくい範囲で汎化タグに置き換えることで，汎化タグに置き換えた単語の翻訳を犠牲にする代わりに全体の翻訳精度を向上させる．本研究では，名詞と動詞を対象に様々な汎化タグを試す．具体的には以下を試す．表 2 には置換例を示した．ただし，汎化タグに置き換えるのは自立語の名詞，動詞のみとする．また，文節内で同じ汎化タグが連続した場合は，まとめて一つの汎化タグとしている．

名詞の汎化タグ

名詞　全ての名詞を “名詞” に置換.
名詞・サ変　サ変接続を “サ変接続” に，それ以外の名詞を “名詞” に置換.
名詞の細分　品詞の細分に置換.

動詞の汎化タグ

動詞　全ての動詞を “動詞” に置換.
動詞の活用　動詞の活用に置換.

ここで注意したいのは，翻訳時に単語の並び替えがほぼ起きないという特徴から，出力時の制約として入力と同じ順番でのみ汎化タグを出力できるようにすることで汎化タグを元の単語に復元できることである．

4. 単一言語データを用いた事前学習

まず，既存の事前学習方法である AutoEncoder を説明し，その後，本研究の提案手法である品詞ごとに方言翻訳の性質が違うことに着目した事前学習を説明する．

表 2 汎化タグの置換例

例文 1					
表層	学会	が	宮古島	で	開催
品詞	名詞	助詞	名詞	助詞	名詞
名詞の細分	一般	-	固有名詞	-	サ変接続
名詞の汎化タグを使用					
名詞	名詞	が	名詞	で	名詞
名詞・サ変	名詞	が	名詞	で	サ変接続
名詞の細分	一般	が	固有名詞	で	サ変接続
例文 2					
表層	走れ	ば	疲れる	よ	
品詞	動詞	助詞	動詞	助詞	
動詞の活用	仮定形	-	基本形	-	
動詞の汎化タグを使用					
動詞	動詞	ば	動詞	よ	
動詞の活用	仮定形	ば	基本形	よ	

4.1 AutoEncoder

AutoEncoder[2] とは，単一言語データを用いて深層学習に基づくモデルの事前学習を行う手法の一種である．具体的には，入力と同じ単語列を予測できるよう学習を行う．1 節でも説明した通り，注視機構付き系列変換モデルで AutoEncoder を行うと，文としての単語の並びが学習されるとは限らない．

4.2 ContToSent: 内容語のみの系列から文を復元

本研究では内容語が翻訳時にあまり変化しないことに着目し，内容語のみの系列から元の文を予想するよう方言の単一言語データを用いて事前学習を行う．関西弁での例を表 3 に示す．この方法により，内容語（入力からあまり変化しない単語）は入力列から情報を得て出力され，内容語以外（変化して出力されるべき単語）は入力列での周辺の情報と方言としての文法的な単語の並びに基づいて出力されるよう学習され，対訳データによる学習が容易になると考えられる．汎化タグによる置き換えを行なっている場合も同様の方法で事前学習を行う．

表 3 各事前学習手法における学習例: 「野球が好きや」の場合

事前学習手法	入力文	目的文
AutoEncoder	野球 が 好き や	野球 が 好き や
ContToSent	野球 好き	野球 が 好き や
ContToSent + 汎化タグ[*1]	名詞 名詞	名詞 が 名詞 や

5. データ

本研究では，代表的な日本語方言として関西弁を取り上げ，標準語から関西弁への翻訳課題を扱う．

[*1] 3 節で述べた “名詞”を汎化タグに用いた場合．“野球”と “好き”が “名詞”に置換されている．

5.1 対訳データ

テレビアニメ「アトム ザ・ビギニング」[*2]の LINE アカウントで配信していた発話データに対して，クラウドソーシングを用いて関西弁への翻訳を行い対訳データを作成した．この発話データは，他者との会話におけるアニメキャラクターの発話であるため，対話における発話と近い性質を持った文である．評価データ，開発データにそれぞれ 100 文ずつ，残りの 1,392 文を訓練データとして用いた．発話データにおける各発話の平均文長を表 4 に示す．

表 4 発話データにおける各発話の平均文長

データ	文字数	単語数
訓練データ	13.0	7.5
開発データ	12.5	7.1
評価データ	12.4	7.0

5.2 単一言語データ

方言の場合，単一言語データでも大規模なデータは存在しない．だが，方言は特定の地域で用いられる．そこで，ツイッターのアカウントがどの地域に住んでいる人のものかを，自由記述することができるプロフィールの位置情報欄からルール[*3]を用いて判定し，大阪府のツイートを関西弁の単一言語データ，crude-tw として用いる．しかし，大阪府のツイートには関西弁のツイート以外にも標準語のツイートが多く混ざっている．そこで，対訳データの訓練データを用いて関西弁であるか否かを判別する分類器を学習し，その分類器で関西弁と分類されたツイートのみを crude-tw から抽出することで，より関西弁が高い確率で含まれる単一言語データ，refined-tw も作成する．分類器は，第一・第二著者（非関西弁使用者）が関西弁であるか否かをアノテーションした 1,000 ツイートの開発データにおいて約 7 割の正解率となるようハイパーパラメータを調整している．結果，同様にアノテーションしたテストデータ 1,000 文における精度は 70.1%，再現率は 50.7%であった．表 5 に crude-tw，refined-tw それぞれの文数とアノテーションしたテストデータにおける関西弁の割合を示す．ただし，ツイートを収集する際に，最初の 10 文字が他のツイートと一致するツイートを除去することで bot のツイートを除去し，25 文字以下のツイートを除去することでノイズとなるような非文法的なツイートを除去した．

6. モデル

6.1 注視機構付き系列変換モデル

注視機構付き系列変換モデル [12] とは，リカレントニュー

[*2] http://atom-tb.com/
[*3] 都道府県名のひらがな，漢字，ローマ字で始まるか終わる．
[*4] 6.3 万文しかなかったため，記号に着目したルールで複数文からなるツイートを分割，文数を増やした．

表 5 単一言語データの文数とテストデータにおける関西弁の割合

データ	文数	関西弁の割合
crude-tw	59.6 万	13.5
refined-tw	10.4 万[*4]	70.1

ラルネットワークを用いて文をベクトルに変換（エンコード）し，そのベクトルから文を生成（デコード）するという系列変換モデル [13] に，エンコード途中のベクトルを出力時に参照することのできる注視機構を付与したものである．注視機構は，各単語出力時に入力列の各単語に注目することができる．このモデルは言語間の翻訳をはじめとした多くの生成タスクで高い精度をあげているため，本研究でもこのモデルを用いる．

6.2 コピー機構

要約タスクにおいて入力列の単語が参照文に含まれていることは多々ある．そのため，入力列の単語を直接出力するための機構としてコピー機構が提案された [11]．この手法では 2 つの確率分布を予測する．一つは，注視機構付き系列変換モデル同様の，語彙中の各単語を出力する確率であり，もう一つは入力列の各単語を出力する確率である．これらを組み合わせることで，語彙と入力の双方から単語を出力することができる．2 節でも述べた通り，本研究と近い着眼点で提案されたこの機構を上記の注視機構付き系列変換モデルに導入し比較を行う．

7. 実験

7.1 実験設定

形態素解析には MeCab[*5]を，文節の区切りの解析には CaboCha[*6]を，関西弁と標準語を分類する分類器にはロジスティック回帰を用いた．モデルのエンコーダには片方向の Long Short-Term Memory[14] を用い，単語の分散表現にはあらかじめ crude-tw を用いて学習した word2vec[8] を用いた．語彙には crude-tw で頻度 5 以上の単語を用い，それ以外の単語を未知語とした．モデルの学習には adam[15] を，テスト時の出力にはビームサーチ[*7]を用いた．ただし，モデルの学習の際，単語の分散表現は学習対象外とした．汎化タグを使用する場合は，汎化タグに置き換えた後に word2vec の学習を行い，出力時には入力と同じ順でのみ汎化タグの出力を許すよう制限，さらに全ての汎化タグを出力せずに出力が終了した生成文候補は除外した．また，コピー機構を導入したモデルでは，入力列と参照列の両方に含まれている名詞と動詞をコピー機構によって生成される単語，他の単語を語彙から生成される単語として学習を行った．そのほかのハイパーパラメータとして，モデルの隠れ層の次元を 1,000，単語の分散表現の次元を 300，ビーム幅を 5 とした．

[*5] http://taku910.github.io/mecab/
[*6] https://taku910.github.io/cabocha/
[*7] 文長等を用いたスコアは用いず，生成確率を用いて文を選択した．

7.2 評価方法

まず，人手評価を行うモデルを決めるため自動評価（ROUGE-2）を行う．多くの単語がそのまま残る方言翻訳では入力をそのまま出力しても約 30.4 のスコアが出てしまうため，極端にスコアの低いモデルを取り除くために自動評価を用いる．今回は事前学習を行わないモデルを取り除いた．

次に，クラウドソーシングによる人手評価を 3 つの指標で行う．指標は，文法性・関西弁らしさ・意味の不変さの 3 つである．文法性・関西弁らしさは 1~3 点（3 が満点），意味の不変さは，変化している（1）・変化していない（2）の 2 択で評価を行う．

7.3 結果

表 6 と表 7 に評価結果[*8]を示す．人手評価は，対訳データのテストデータ 100 文を対象に関西圏に在住の 3 人の評価者によって行なった．表 7 では全てのモデルを人手評価しているため自動評価の結果を割愛している．また，本研究の目的である正しい翻訳がどれだけ行えたかを調べるため，全ての指標が満点であった割合（正しい翻訳率）も併せて記載する．特に記載がない場合は refined-tw が事前学習に用いた単一言語データである．

ベースラインとの比較: 表 6 に結果を示す．ベースラインは，汎化タグを用いずに AutoEncoder による事前学習を行ったモデルである．まず，汎化タグを導入すると正しい翻訳率が増加していることがわかる．また，事前学習を ContToSent に変更すると，意味の不変さがほぼ変わらず，他の指標で評価が向上している．さらに，ContToSent による事前学習と汎化タグによるデータの抽象化の両方を行うことで，全ての指標での評価の向上，さらに正しい翻訳率の大幅な増加が確認できる．よって，提案手法の有効性が確認できる．

事前学習手法と単一言語データの比較: 次に，表 7 の事前学習に用いるデータ，事前学習の手法を変更した 4 つのモデルの結果に注目する．refined-tw で ContToSent を行なった場合に全ての指標でもっとも高い評価を得ていることから，単一言語データの標準語を取り除くこと，そして ContToSent の有効性が確認できる．ここで注目したい点は，事前学習の方法を AutoEncoder から ContToSent に変更すると，refined-tw では文法性と関西弁らしさの両方の評価が上がっているのに対し，crude-tw では文法性が上がって関西弁らしさが下がることである．これは，文としての単語の並びがより学習されるようになった結果，標準

[*8] これらの評価実験は別々に行っている．

語が多く含まれる crude-tw では標準語の単語の並びが学習されてしまい，標準語が出力されてしまったことが原因であると考えられる．

名詞の汎化タグの比較: 表 7 の名詞の汎化タグを比較した結果に注目する．まず，正しい翻訳率では “名詞・サ変” がもっとも高い値となっていて，この手法がもっとも良いことがわかる．関西弁らしさでは “名詞の細分” の方が優っているが，これは “名詞・サ変” と比べて抽象化があまり進んでいないために表現力が高いためであると考えられる．しかし，同様の理由により過学習が進みやすくなり，文法性や意味の不変さが低くなってしまったと考えられる．“名詞” は 3 つの汎化タグのうちで最も低い正しい翻訳率になっている．これは，方言翻訳において動詞に近い役割であるサ変接続とその他の名詞の区別が重要であることを示している．ここでもう一つ言及しておくべき点は，汎化タグを使用した全てのモデルの正しい翻訳率が，汎化タグを使用しない場合より高いか同じであることである．ここでも汎化タグの有効性が確認できる．

動詞の汎化タグの比較: 表 7 の動詞の汎化タグを比較した結果に注目する．結論から言うと，汎化タグを使用しない場合が，多くの指標でもっとも高い評価を得ていた．これは，動詞が周囲の翻訳，特に後方の翻訳に与える影響が大きく，活用まで細かくしても周囲の翻訳に必要な情報が足りていなかったことが原因であると思われる．

コピー機構との比較: 表 7 のコピー機構との比較を行った結果に注目する．まず，コピー機構を用いずに提案手法を用いたモデルが全ての指標でもっとも高い評価を得ている．これはコピー機構を追加したことによってモデルが肥大化し，小規模な学習データからでは学習できなかったことが原因だと考えられる．また，コピー機構を導入した 2 つのモデルを比較すると，ベースラインよりも提案手法の場合の方が正しい翻訳率が高いことから，提案手法がコピー機構を導入したモデルでも有効であることがわかる．

出力例: 最後に，ベースラインとなる汎化タグを用いずに AutoEncoder を行った手法と，提案手法となる汎化タグを用いつつ ContToSent を行った手法の出力例を表 8 に提示する．例 1 はどちらのモデルでも良い出力がされている．例 2 では提案手法はうまく翻訳できているがベースラインの場合，“通話” が “し” になってしまっている．提案手法では “通話” が名詞のサ変接続であるため汎化タグに置き換わり，うまく生成できたと考えられる．例 3 でも，提案手法はうまく翻訳できているがベースラインでは非文法的な文が生成されてしまっている．例 4 では，ベースラインでは非文法的な文が生成されていている．一方，提案手法では文法的で関西弁らしい文が生成されているが，参照文のように直接的な翻訳ではなく意訳に近い翻訳がされてしまっているため文脈によっては間違った翻訳である．

8. まとめ

本研究では，小規模な対訳データから方言翻訳を行うため，文としての単語の並びがより学習されやすい事前学習手法（ContToSent）を提案した．さらに，タスクの難易度を下げるため，汎化タグによるデータの抽象化を提案した．これらの有効性が関西弁で人手評価により確認された．

今後は，対訳データ，単一言語データ量を変化させた場合の結果や，他の方言での結果を調査したいと考えている．また，汎化タグに置き換えた単語を後から方言に置き換える手法の開発も行いたいと考えている．

謝辞 対訳データに用いた発話データを提供していただいたアトム ザ・ビギニング製作委員会に深く感謝いたします．

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] Hinton, G. and Salakhutdinov, R.: Reducing the Dimensionality of Data with Neural Networks, *Science*, Vol. 313, No. 5786, pp. 504 – 507 (2006).
[3] 濱田晃一，藤川和樹，小林颯介，菊池悠太，海野裕也，土田正明：対話返答生成における個性の追加反映，情報処理学会研究報告，Vol. 2017-NL-232, No. 12, pp. 1–7 (2017).
[4] 平山直樹， 森信介， 奥乃博：方言対訳コーパスを用いた日本語方言変換システム，第 75 回全国大会講演論文集，Vol. 2013, No. 1, pp. 519–520 (2013).
[5] 柴田直由，横山晶一，井上雅史：統計的手法を用いた双方向方言機械翻訳システム，言語処理学会第 17 回年次大会発表論文集，pp. 126–129 (2013).
[6] Rush, A. M., Chopra, S. and Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization, *CoRR*, Vol. abs/1509.00685 (2015).
[7] Wubben, S., van den Bosch, A. and Krahmer, E.: Sentence Simplification by Monolingual Machine Translation, *ACL*, pp. 1015–1024 (2012).
[8] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781 (2013).
[9] Wiseman, S., Shieber, S. M. and Rush, A. M.: Challenges in Data-to-Document Generation, *CoRR*, Vol. abs/1707.08052 (2017).
[10] Murakami, S., Watanabe, A., Miyazawa, A., Goshima, K., Yanase, T., Takamura, H. and Miyao, Y.: Learning to Generate Market Comments from Stock Prices, *ACL*, pp. 1374–1384 (2017).
[11] Gu, J., Lu, Z., Li, H. and Li, V. O. K.: Incorporating Copying Mechanism in Sequence-to-Sequence Learning, *CoRR*, Vol. abs/1603.06393 (2016).
[12] Luong, T., Pham, H. and Manning, C. D.: Effective Approaches to Attention-based Neural Machine Translation, *EMNLP*, pp. 1412–1421 (2015).
[13] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks., *NIPS*, pp. 3104–3112 (2014).
[14] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, pp. 1735–1780 (1997).
[15] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (2014).

*9 明らかに自動評価結果が低いため人手評価を行わない．

表 6 評価結果 1：関西弁におけるベースラインとの比較

事前学習	汎化タグ	人手評価				自動評価
		文法性	関西弁らしさ	意味の不変さ	正しい翻訳率	
なし	なし			-[*9]		2.8
AutoEncoder	-	2.22	2.06	1.55	28.6	**31.8**
AutoEncoder	名詞・サ変	2.28	1.92	**1.62**	29.0	29.3
ContToSent	-	2.56	2.34	1.53	34.6	28.7
ContToSent	名詞・サ変	**2.63**	**2.43**	1.60	**40.3**	31.2
参照文		2.95	2.83	1.94	83.6	30.4

表 7 評価結果 2：関西弁における特定の視点での比較

事前学習[*7]	汎化タグ	コピー機構	人手評価			
			文法性	関西弁らしさ	意味の不変さ	正しい翻訳率
事前学習手法と使用する単一言語データの比較						
AutoEncoder (crude-tw)	名詞・サ変	-	2.36	2.15	1.59	31.0
ContToSent (crude-tw)	名詞・サ変	-	2.37	2.04	1.46	25.0
AutoEncoder	名詞・サ変	-	2.27	1.87	1.61	26.6
ContToSent	名詞・サ変	-	**2.53**	**2.36**	**1.65**	**39.0**
名詞の汎化タグの比較						
ContToSent	-	-	2.49	2.35	1.53	33.0
ContToSent	名詞	-	**2.53**	2.27	1.53	33.0
ContToSent	名詞・サ変	-	**2.53**	2.36	**1.65**	**39.0**
ContToSent	名詞の細分	-	2.46	**2.44**	1.54	37.0
動詞の汎化タグの比較						
ContToSent	名詞・サ変	-	**2.53**	2.36	**1.65**	**39.0**
ContToSent	名詞・サ変+動詞	-	2.23	2.16	1.49	31.6
ContToSent	名詞・サ変+動詞の活用	-	2.46	**2.38**	1.55	37.6
コピー機構との比較						
ContToSent	名詞・サ変	-	**2.53**	**2.36**	**1.65**	**39.0**
AutoEncoder	-	あり	2.08	2.06	1.43	19.6
ContToSent	名詞・サ変	あり	2.34	2.30	1.42	27.6

表 8 各モデルの出力例

例 1	標準語	いろんな映画の監督をされている人だと聞いています
	参照文	いろんな映画の監督をしてはる人やって聞いてます
	AutoEncoder	いろんな映画の監督をされている人やと聞いてるわ
	ContToSent+汎化タグ	いろんな映画の監督をされてる人やと聞いたで
例 2	標準語	すみません、「通話」はできません
	参照文	ごめん、「通話」はできへんわ
	AutoEncoder	すまん、「し」はできへんわ
	ContToSent+汎化タグ	すまん、「通話」はできへん
例 3	標準語	お話であれば聞けますよ
	参照文	話やったら聞けるで
	AutoEncoder	話あれたら聞けるで
	ContToSent+汎化タグ	お話やったら聞けるで
例 4	標準語	ボクのシステムの時間と現実の時間が一致していないので教えることができません…
	参照文	あたしのシステムの時間と現実の時間が一致してへんから教えられへん
	AutoEncoder	ボクのシステムの時間とこうの時間がってしてへんからそんなことができへん…
	ContToSent+汎化タグ	ボクのシステムの時間と現実の時間が一致してないと答えることができひんねん