日英対訳パターンを用いた名詞句翻訳

吉岡篤志　徳久 雅人　村上 仁一　池原 悟
鳥取大学 工学部 知能情報工学科
{yosioka,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

重文・複文の日英機械翻訳の手法の１つとして，文型パターンを用いる手法が提案されている [1]．その手法では，パターンの記述要素に，名詞句や動詞句といった品詞に対応した変数が使われており，その変数に代入された日本語表現を翻訳する必要がある．

そのため，名詞句の翻訳に関しては，名詞句パターン辞書が構築されたが，辞書の性能評価にとどまっている [2]．また，その辞書を用いて，英訳生成部を自動化し，bi-gram による訳出選択が行われたが，表現選択において改良の余地があることが示されている [3]．

そこで，本稿では，名詞句パターン辞書で生成された複数の英語表現に対して，日本語と英語の構造に着目した確率モデルを提案し，表現選択を行う．また，日本語と英語の表現が対応する確率（以下，翻訳確率と呼ぶ）をパラメータの１つとした多変量解析による訳出選択も行い，両選択方式を比較する．なお，翻訳精度は人手で評価する．

2　名詞句パターン辞書

本稿は，[2] の名詞句パターン辞書を用いる．パターン辞書は，日英名詞句の組が約 4.5 万件収録された名詞句コーパスに，単語アライメントを適用して作成された．その規模は，23,834 パターン対である．以下に具体例を示す．

- **日本語パターン：** $PRN1$ の $N2$（男女，親）
- **英語パターン：** $PRN1$`^poss` $N2$

パターン辞書を用いた翻訳の手順は，基本的には次の通りである．まず，日本語の名詞句を入力し，適合する日本語パターンを辞書から検索する．次に，対となる英語パターンの変数に訳語を挿入して，英語の名詞句を生成する．上述の例に見られる「$PRN1$」や「$N2$」は代名詞や名詞の変数であり，10 種類が辞書で使われている．「(男女，親)」は意味属性制約であり，変数に代入可能な表現を制限する．「`^poss`」は英語表現を所有格に変形する関数である．

例えば，「彼のお母さん」が上述の日本語パターンに適合すると「$PRN1$=彼」，「$N2$=お母さん」となり，各訳語を英語パターンに挿入して格の変形を行うと「his mother」が訳出できる．

3　英語名詞句生成

英語生成の自動化を行う．具体的には次の 3 点である．

- **訳語挿入：** 変数に代入された日本語表現の英訳語を英語パターンに挿入する．
- **形態素調整：** 付随する関数に従い，英訳語を変形する．
- **訳出選択：** ４章に示す翻訳確率あるいは５章に示す多変量解析を用い，訳出選択を行う．

3.1　訳語挿入

名詞句パターン辞書では，一般名詞の変数は，接辞および純粋な名詞の連続に適合し，一般名詞の変数を除く変数は，変数の示す品詞の単語に適合する．したがって，変数に代入された日本語表現を順に和英辞書を用いて英語表現に変換することで，訳語挿入は実現できる．なお，日本語変数と英語変数が対応関係にあっても，品詞が異なる場合がある．その場合は，和英辞書に登録されている品詞情報を用いて，当該品詞の訳語を挿入する．和英辞書には１つの日本語単語に対して複数の英語単語が対応する．また，１つの英語パターンには複数の変数が存在する．したがって，訳語挿入の結果は，これらの組み合わせを全通り出力したものとなる．

3.2　形態素調整

英語パターンでは，活用形を表す関数「`^poss`」，数を表す関数「`^pl`」の２種類が使用されている ．和英辞書における品詞情報に基づき，関数の示す条件を満たす単語に置き換える．本和英辞書は，主に CD-ROM 版マル

チ辞書（旺文社）における和英・英和データから作成した．本稿で使用している規模は，約 12 万語である．以下に和英辞書の一部を示す．

季節 season, tide

season(1111)(2101), tide(1111)(1130)(2101)

（ ）内は品詞コード（1000 番台は名詞，2000 番台は動詞）

4 確率モデルによる英語表現選択

部分表現列（n 個の単語列）$w_{j1}\ w_{j2}\ w_{j3}\ \ldots\ w_{jn}$ が $w_{e1}\ w_{e2}\ w_{e3}\ \ldots\ w_{en}$ に翻訳される確率を以下の式と仮定することができる．

$$\begin{aligned} &P(w_{e1}w_{e2}w_{e3}...w_{en}|w_{j1}w_{j2}w_{j3}...w_{jn}) \\ \simeq\ &P(w_{e1}|w_{j1})\cdot P(w_{e2}|w_{j2})\cdot P(w_{e3}|w_{j3})\cdot ... \\ &\cdot P(w_{en}|w_{jn})\cdot P(w_{e1}w_{e2}w_{e3},...w_{en}) \end{aligned}$$

$P(w_{ei}|w_{ji})$ は，単語 w_{ji} が w_{ei} に翻訳される確率，$P(w_{e1}w_{e2}w_{e3}...w_{en})$ は語順の確率である．しかし，一般に日本語と英語の構成単語数が異なるため，上記の式は利用できない．

一方，パターン翻訳では日英パターン対において使用される変数の数が同一である，という点に着目すると，変数単位での翻訳確率と，非変数部すなわちパターン単位での翻訳確率に分けた近似式が考えられる．日本語変数に対応する表現 v_{ji} が英語変数に対応する表現 v_{ei} に翻訳される確率，すなわち，変数単位の翻訳確率を $P(v_{ei}|v_{ji})$ とし，日本語パターン p_j が英語パターン p_e と対応する確率を $P(p_i|p_j)$ とすると，上述の部分表現列の翻訳確率は次の式で近似できる．

$$\begin{aligned} &P(w_{e1}w_{e2}...w_{en}|w_{j1}w_{j2}w_{j3}...w_{jn}) \\ \simeq\ &\{\Pi_i^m P(v_{ei}|v_{ji})\}\cdot P(p_e|p_j)\cdot P(w_{e1}...w_{en}) \\ &(v_j\in\{w_j\}, v_e\in\{w_e\}) \end{aligned}$$

ただし，英語単語 n 個のうち，m 個が変数に対応しているものとする．

例えば，日本語名詞句「もっとも良い季節」を翻訳する場合，「もっとも良い $N1$ ⇔ the most suitable $N1$」というパターン対で作成された「the most suitable season」と，「もっとも $AJ1N2$ ⇔ the $AJ1$ ^est $N2$」というパターン対を用いて作成された訳出候補「the best season」の選択を行う状況を考える．そのためには，2 つのうち，確率の高い訳出候補を選択すればよい．

- $P($ the most suitable season $|$ もっとも良い季節)
- $P($ the best season $|$ もっとも良い季節)

「もっとも良い $N1$」と「the most suitable $N1$」を使って翻訳する場合，$N1$=季節=season のときの翻訳確率は，以下のようになる．

$$\begin{aligned} &P(\text{the most suitable season}|\ \text{もっとも良い季節}) \\ =\ &P(\text{season}|\ \text{季節}) \\ &\cdot P(\text{the most suitable } N1|\ \text{もっとも良い } N1) \\ &\cdot P(\text{the most suitable season}) \end{aligned}$$

また，「もっとも $AJ1N2$」と「the $AJ1$^est $N2$」の英語パターンを使って翻訳する場合，$AJ1$=よい=best, $N2$=季節=season のときの翻訳確率は以下のようになる．

$$\begin{aligned} &P(\text{the best season}|\ \text{もっとも良い季節}) \\ =\ &P(\text{best}|\ \text{良い})\cdot P(\text{season}|\ \text{季節}) \\ &\cdot P(\text{the } AJ1\ N2|\ \text{もっとも } AJ1N2) \\ &\cdot P(\text{the best season}) \end{aligned}$$

5 多変量解析による英語表現選択

5.1 評価関数の作成

[4] では，重文・複文の入力に対して複数の適合パターンが出力される場合に，英訳に翻訳可能な適合パターンの選択手法として多変量解析を用い，最適な適合パターンの選択を実現し，その有効性を示している．

本稿では，入力名詞句に対して複数の英訳語の候補が生成される．その際に，訳出選択が大きな課題となる．そこで，訳出選択手法の一つとして，多変量解析を用いる．

評価関数の作成では，入力名詞句の訳出候補に対して４章で示した翻訳確率を含めた６つのパラメータを使用する．翻訳確率をパラメータの１つとして加えた理由は，多変量解析により翻訳確率の訳出選択よりも品質の高い英訳を得ることを目的とするためである．

パラメータは具体的には，[4] で用いられているパラメータ３つ（パターン字面適合率，パターン元字面適合率，変数の適合率）および翻訳確率，日英のパターン類似度，日英単語類似度に着目したパラメータ３つ（翻訳確率，句のアライメント確率，日本語と英語間の名詞意味属性距離）を追加した６つである．以下に評価パラメータを示す．

y ： 評価値

x_1： 句のアライメント確率

x_2： パターン字面適合率

x_3： パターン元字面適合率

x_4： 変数の適合率

x_5： 翻訳確率

x_6： 日本語と英語間の名詞意味属性距離

5.1.1 評価値

生成される英訳語の数が多量であり，その英訳語全てに人手評価を行い評価関数を作るのは，とても困難である．本稿では NIST 値 [6] を使用して，訳出候補の品質を評価する．これにより，評価関数を作成する際のコストを削減できる．

NIST 値は，値が高い程，品質の高い英訳を示しており，その値に上限はない．評価値を 0 から 1 の範囲に収めるため，入力名詞句に対する NIST 値の最大値でそれぞれの値を除算したものを評価値とした．

5.1.2 句のアライメント確率

句アライメント確率とは，日本語パターンとそれに対応する英語パターンの使用頻度である．

$$\text{句のアライメント確率} = \frac{\text{対応する英語パターンが同じであった件数}}{\text{パターン辞書に存在した日本語パターンの件数}}$$

具体的には，入力名詞句「もっとも良い季節」の場合，「もっとも良い $N1$ ⇔ the most suitable $N1$」とパターン辞書から抽出される．しかし，日本語パターンが「もっとも良い $N1$」であっても，対応する英語パターンが異なる場合がある．

例えば「もっとも良い $N1$ ⇔ the best $N1$」である．この場合，名詞句パターン辞書に日本語パターンが「もっとも良い $N1$」であったのは 2 件で，そのうち 1 件の対応する英語パターンが「the best $N1$」であった．よって，「もっとも良い $N1$ ⇔ the most suitable $N1$」の句のアライメント確率は，0.5(1/2) となる．

5.1.3 日本語と英語間の名詞意味属性距離

日英単語類似度に着目し，翻訳された名詞変数部分の日本語側意味属性 [5] と英語側意味属性の距離の平均を算出し，それを逆数で示したものを使用する．

例えば，名詞変数が N が「季節」の場合，「季節」は「season」,「tide」に英訳される．そのときの日本語側の意味属性と英語側の意味属性距離の平均とその逆数を表 1 に示す．

表 1 意味属性距離の平均とその逆数 (名詞句変数：季節)

英訳語	season	tide
意味属性距離の平均	5	7
意味属性距離の平均の逆数	0.2(1/5)	0.15(1/7)

この結果から「季節」と「season」が意味的に近いことが分かる．

5.2 評価関数の作成結果

名詞句パターン辞書からランダムに抽出したテスト入力名詞句 300 件を用いて，評価関数を決定した．また，求められた評価関数に対する予測とどのパラメータが評価値に影響を与えるかの要因分析を行い，最適な評価関数を求めた．

「具体的には，全てのパラメータを用いて一度回帰分析を行い，その結果から Ru(説明変数選択基準) を求める．次にパラメータの中で，危険率のもっとも高いパラメータを削除して，再度回帰分析を行い，Ru を求める．これを繰り返し，得られた回帰分析のうち，Ru がもっとも高い評価関数を最適な評価関数とする」([7] より引用)．

結果，全てのパラメータを用いた評価関数がもっとも Ru が高く，最適な評価関数であった．以下に得られた評価関数を示す．

$$\hat{y} = 0.809 + 0.007x_1 + 0.351x_2 - 0.500x_3 - 0.285x_4 + 0.350x_5 - 0.008x_6$$

6 翻訳実験

6.1 実験の方法

翻訳確率，多変量解析の訳出選択で，クローズドテストおよびオープンテストを行う．入力名詞句および正解訳は，名詞句パターン辞書の作成に用いた日英名詞句対よりランダムに抽出した 300 件を用いる．オープンテストの入力には，名詞句パターン辞書から評価関数で用いなかった 300 件をランダムに抽出して使用する．なお，1 つの入力名詞句に対して，平均の訳出候補数は，274 件である．訳出の評価として，訳出順位の上位 10 位まで人手評価を行う．判定規準は以下の通り．

◎： 出力訳と正解訳が，冠詞および数を除き，完全一致している場合

○： 出力訳が，文法的にも意味的にも正しいが，正解訳と異なる場合（例．入力：新しいページ，正解訳：new chapter，出力訳：new page）

×： 出力訳が上記以外

6.2 ベースライン

評価の比較のために，bi-gram による訳出選択方法 [3] の結果を以下に示す．

表 2 bi-gram のオープンテスト

順位	1 位	3 位	5 位	10 位
◎	8%(24)	14%(42)	16%(48)	19%(57)
○	27%(81)	41%(122)	41%(123)	42%(126)
×	65%(195)	45%(136)	43%(129)	39%(117)

6.3 実験結果

翻訳確率と多変量解析の上位 10 位までのクローズドテストおよびオープンテストの評価を以下の表に示す．

表 3 翻訳確率のクローズドテスト

順位	1 位	3 位	5 位	10 位
◎	13%(38)	20%(60)	27%(82)	34%(103)
○	34%(103)	42%(127)	44%(132)	45%(134)
×	53%(159)	38%(113)	29%(86)	21%(63)

表 4 翻訳確率のオープンテスト

順位	1 位	3 位	5 位	10 位
◎	19%(57)	27%(80)	30%(90)	39%(116)
○	21%(64)	28%(85)	31%(92)	31%(95)
×	60%(179)	45%(135)	39%(118)	30%(89)

翻訳確率の訳出選択の場合，上位 10 位までの正解含有率（◎＋○）は，クローズドテストで 79%，オープンテストで 70% であった．これは，上位 10 位までの bi-gram による訳出選択の正解含有率が 61% であったことから，より品質の高い訳出選択であり，翻訳確率は名詞句の訳出選択に有効な手段と言える．

表 5 多変量解析のクローズドテスト

順位	1 位	3 位	5 位	10 位
◎	11%(34)	15%(45)	19%(58)	26%(78)
○	37%(110)	43%(128)	45%(134)	45%(135)
×	52%(156)	42%(127)	36%(108)	29%(87)

表 6 多変量解析のオープンテスト

順位	1 位	3 位	5 位	10 位
◎	12%(36)	19%(56)	23%(68)	28%(85)
○	25%(74)	33%(98)	34%(102)	38%(115)
×	63%(190)	49%(146)	42%(126)	33%(100)

一方，多変量解析による訳出選択は，上位 10 位までの正解含有率は，クローズドテストは 71%，オープンテストは 67% であり，翻訳確率のみの訳出選択よりも良い結果を得られなかった．また，bi-gram のみの訳出選択と比較すると，翻訳確率ほどではないが，よい訳出選択であった．

7 考察

重文・複文の場合の英訳可能な適合パターン選択では，パターン元字面適合率がもっとも有効なパラメータであった．しかし，名詞句は重文・複文に比べて単語数が少ないため，パターン元字面適合率のパラメータはあまり効果がなく，多変量解析の訳出選択では，良い結果がえられなかったと考えられる．

8 おわりに

今回，名詞句パターン辞書を用いて，複数生成される英訳語に対して，本稿で提案した翻訳確率を用いた訳出選択と多変量解析による訳出選択を行った．

実験の結果，多変量解析を用いるより，翻訳確率を単独で用いた訳出選択が良い結果が得られた．また，多変量解析と翻訳確率は，bi-gram のみの訳出選択よりも効果があり，特に翻訳確率のみの訳出選択は，訳出選択に有効な手段であることが示された．

今後，名詞句外の情報 (主名詞と動詞との共起関係) を考慮することや，多変量解析のパラメータの見直しを行い，動詞句翻訳に統合する．

謝辞

本研究は，科学技術振興事業団「JST」の戦略的基礎研究推進事業「CREST」における 研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーによる言語の等価変換と生成技術」の支援によるものである．

参考文献

[1] 池原悟, 阿部さつき, 徳久雅人, 村上仁一 :非線型な表現構造に着目した重文と複文の日英文型パターン化, 自然言語処理, Vol.11, No.3, pp.69-95, 2004.

[2] 神野絵理, 徳久雅人, 村上仁一, 池原悟:文型パターンによる日英翻訳のための名詞句パターン辞書の構築, 言語処理学会第 11 回年次大会, pp.376-379, 2005.

[3] 吉岡篤志, 徳久雅人, 村上仁一, 池原悟:名詞句パターン辞書を用いた日英機械翻訳の試作—bi-gram による訳出選択の場合, 電気・情報関連学会中国支部第 56 回連合大会,pp.305-306,2005

[4] 岡田敏:多変量解析による最適文型パターンの選択方式, 言語処理学会第 11 回年次大会発表論文集,pp.25-28,2005.

[5] 池原, 宮崎, 白井:日本語語彙大系, 岩波書店,1997.

[6] NIST:Automatic Evaluation of Machine Translation Quality Using N-gram Co -Occurrence Statistics,2002.（http://www.nist.gov/speech/）

[7] 実践ワークショップ Excel 徹底活用 多変量解析, 秀和システム, ISBN4-7980-0558-4.