

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

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)

1 はじめに

重文・複文の日英機械翻訳の1つの手法として、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、名詞句や動詞句などの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。名詞句の翻訳については、名詞句パターンを用いる手法が提案され、名詞句パターン辞書が構築された [2]。しかし、[2]では、辞書の潜在的な性能を評価するにとどまり、英訳の生成部（複数出力される適合パターンの選択および訳語選択）の自動化が課題として残されていた。一方、近年注目されている統計翻訳では、デコーディングの過程で、目的言語の言語モデルに基づく訳出候補の選択が行われている [3]。そこで、本稿では英語生成を自動化するために、パターン翻訳の過程において、目的言語の言語モデルに bi-gram を用いた訳出選択を行うことにする。そして、その翻訳精度を評価する。

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。本パターン辞書は、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 英語名詞句の生成

[2]の実験では、英語生成が手作業で行われていた。本稿では、英語生成の自動化を試みる。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：数多く出力される英語名詞句の候補を順位付ける。

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、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)

3.3 訳出選択

訳出候補は、英単語列の形式で列挙される。順位付けは、隣接する単語の共起確率を用いる。各訳出候補の共起確率を bi-gram で計算し、確率値の降順で順位を定める。出現頻度は、重文・複文の日英対訳コーパス(約 12 万文対を収録)を用いて計算する。

4 翻訳実験

4.1 実験の目的

本方式による訳出の評価を行う。評価には、次の観点に注目する。

パターン翻訳に期待される点は、パターンの適合時に品質の良い訳出が得られることである。したがって、パターンの適合する割合（カバー率）、および、適合時の正答率が問われる。さらに、訳出選択の評価として、訳出順位の上位 10 位までの正答率の分布を求める。

1 つの日本語名詞句の英語訳は複数通りが考えられる。訳出の人手による評価は、意味的な等価性の解釈が考慮できるが、コストが高く評価者に依存しやすい。NIST スコア [4] など機械による評価は、コストが安いですが、意味的な解釈は考慮できず、また、リファレンスが少ない場合に信憑性が低い。本稿では今後の利用を検討するために、両者の評価値を求める。

なお、名詞句訳語は、文全体とのバランスで変形されることがある。したがって、正解の英訳と出力の英訳において、冠詞の有無、および、単複数の違いは、問わないこととする。

4.2 実験の方法

クローズドテストおよびクロスバリデーションを行う。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の名詞句対よりランダムに抽出した 357 件を用いる。出力訳の評価は、人手評価および NIST スコア [4] による。その際の判定基準は以下の通りとする。

：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（NIST スコアの算出に用いるリファレンスにおいて、正解訳から冠詞および数の表現を除去）

：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、または、出力訳が無い場合
カバー率および正答率は以下の式で求める。

$$\text{カバー率} = \frac{\text{1 つ以上の出力訳を得た入力名詞句の数}}{\text{入力名詞句の数}}$$

$$\text{正答率} = \frac{\text{評価() の出力訳を得た入力名詞句の数}}{\text{1 つ以上の出力訳を得た入力名詞句の数}}$$

4.3 実験結果

カバー率および正答率を表 1 に示す。クローズドテストにおいてカバー率が 100% でない理由は、訳語挿入時の和英辞書引きに失敗したためである。クロスバリデーションにおいて、カバー率は低いが生正答率がクロー

ズドテストに近い。これは、パターン翻訳で期待される点に依っている。

次に、クロスバリデーションにおける正答率の分布を表 2 に示す。出力訳の個数は、48,409 個であった。上位 10 位までにおいて、評価() の訳出を含めると 60% を占めるが、評価() の割合が低い。すなわち、147 個のうち 67 個 (46%) しか上位 10 位までに含まれていない。訳出選択に改善の余地が残っていることが分かる。

また表 2 には、各位の NIST スコアの平均値を示した。NIST スコアの平均値は、候補数当りの人手評価における評価() の個数と相関がある（相関係数=0.979）。今後、訳出選択の手法を改良する際、平均 NIST スコアの全体的な向上から改良の程度を伺うことに役立つ。

表 1: カバー率と正答率

	カバー率	正答率
クローズドテスト	92%(327/357)	89%(291/327)
クロスバリデーション	49%(174/357)	85%(147/174)

表 2: 上位候補の評価

順位	1 位	3 位	5 位	10 位
候補数	174	516	842	1,617
	16%(29)	9%(50)	6%(58)	4%(67)
	50%(88)	57%(298)	58%(492)	56%(912)
x	32%(57)	31%(168)	33%(292)	38%(638)
NIST 平均	0.45	0.39	0.34	0.29

() 内は個数を表す。

5 おわりに

本稿は、名詞句パターン辞書を用いた日英機械翻訳の実現に向けて、訳出選択に bi-gram を用いて英語生成部を試作した。出力訳を評価した結果、パターン適合時の最大正答率は 85% であった。訳出選択の評価として、上位 10 位までの精度を調べたところ、意味的・文法的に正しい出力訳の含有率は 60% であった。しかし、正解訳と完全に一致する出力訳の含有率が低かったため、今後、訳出選択の改良が必要であることがわかった。さらに、人手による訳出の評価はコストが高いことが問題であるが、本実験により、NIST スコアの平均値が人手評価の値と相関があることが確認できた。よって、今後の訳出選択の開発段階においては、NIST スコアを用いることで人手評価のコストが削減できる。

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] Knight Kevin: A Statistical MT Tutorial Workbook, 1999. <http://www.clsp.jhu.edu/ws99/projects/mt/mt-workbook.htm>
- [4] NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, 2002. <http://www.nist.gov/speech/>