機械翻訳の前処理の帰納的学習

長島康人　　荒木健治　　栃内香次

北海道大学大学院工学研究科

E-mail:{naga,araki,tochinai}@media.eng.hokudai.ac.jp

1　はじめに

近年の世界規模でのネットワークの普及に伴い，電子化された異言語文書に触れる機会が増大している．それに対して，機械翻訳ソフトは多くのものが市販されているが，それらの翻訳結果には訳語選択や係り受け関係の誤り，不自然な表現などが多く含まれ，いずれも満足のいくものではない[1][2].

機械翻訳精度向上の方法として，入力文に対してあらかじめ処理を施す，前処理がある．前処理は同言語間での変換であり，ある意味では言い換えである．しかし，従来行われてきた言い換えの研究[3][4]には純粋に翻訳精度を向上させるためのものは少ない．また，翻訳精度向上のために行われている前処理に関する研究は，個々の機械翻訳システムに付随して研究されているものが多く[5]，現在，既に多数存在する機械翻訳システムに対して汎用的に適用できるというものではない．これらの問題を解決するために，どのような機械翻訳システムに対しても適用可能な，帰納的学習を用いた汎用的な前処理手法を提案する．

本稿では，まずシステムの概要を説明し，その後実際にシステムを作成して行った評価実験の結果を記述する．

2　処理過程

図1に処理の概要を示す．本稿における実験システムは英日翻訳を対象として作成されている．

2.1　概要

最初に，英語文を形態素解析したものを入力する．ここで，本手法で用いる情報は形態素解析結果のみとする．これは，そもそも機械翻訳における誤りというものは，構文解析や意味解析の解析誤りによるものが多いと考えられるからである．

システムでは，まずルール探索部において，変換ルール辞書中に入力文に対し適用されるルールが存在するか調べる．適用されるルールが存在した場合，ルール適用部において適用ルールを用い入力文を変換する．次に変換結果を機械翻訳システムに入力し，翻訳を行う．この翻訳結果と入力に用いた英語文の翻訳結果から判定を行い，フィードバック処理部において 2.4.1 で述べるルールの正変換度数，誤変換度数の更新を行う．その後，学習処理部において入力英語文の形態素解析結果と人手により作成された正解文から，変換ルールを獲得する．

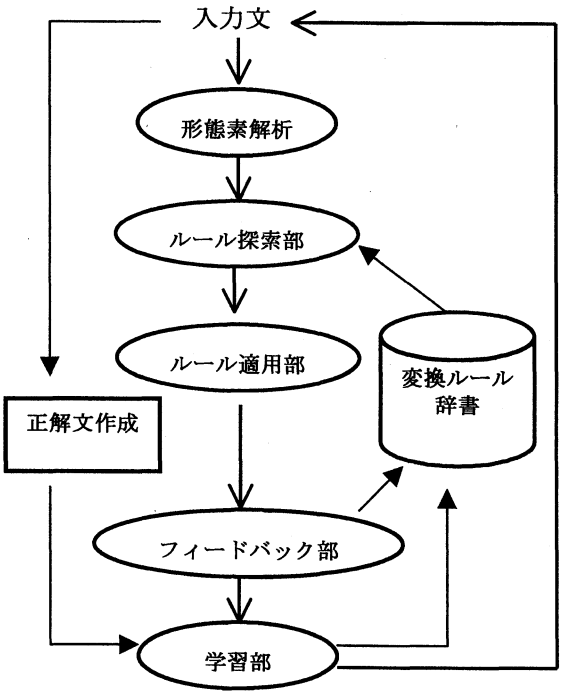

図1　処理の概要

2.2　ルール探索部

ルール探索部では，獲得されたルールが保持されている変換ルール辞書中に適用されるルールが

入力文：Therefore/RB ,/, our/PRP$ proposed/VBN method/NN consists/VBZ of/IN the/DT analysis/NN and/CC learning/NN ./.

正解文：Therefore/RB ,/, the/DT method/VBN we/PRP proposed/VBN consists/VBZ of/IN the/DT analysis/NN and/CC learning/NN ./.

↓

差異部分：“our proposed method “と” the method we proposed “

↓

獲得されるルール： our proposed method ⇒ the method we proposed

差異部分前　Therefore/RB ,/,

差異部分後　consists/VBZ of/IN

RB：副詞，**PRP$**：複数形所有格，**VBN**：過去分詞，**NN**：一般名詞，**VBZ**：一般動詞，**IN**：前置詞，**DT**：冠詞 **CC**：接続詞

図 2：ルール獲得例

存在するか探索する．変換ルールは，各ルールごとに適用条件を保持しており，その適用条件を満たしたルールが適用される．複数のルールが重複した場合，2.4.2 で述べるルール適用正解率の高いルールを適用する．もし，ルール適用正解率が等しい場合には，獲得された時期が最も新しいルールを優先して適用する．

2.3　ルール適用部

ルール適用部では，ルール探索部において適用条件を満たしたルールを用いて，入力文を変換する．変換処理は、基本的に単語の付加，削除，および文の分割からなる．もし，適用条件を満たすルールが存在しなかった場合には入力文をそのまま出力する．

2.4　フィードバック処理部

ルール適用部からの出力を機械翻訳システムに入力し，その翻訳結果に対して判定を行う．判定結果から，適用されたルールの正変換度数，誤変換度数を変化させ，これによりルールの適用条件を変化させる．

2.4.1　正変換度数と誤変換度数

入力された英文の機械翻訳システムによる翻訳結果と，ルール適用後の翻訳結果を比較し，ルールの適用により翻訳結果が改善された場合にはそのルールの正変換度数を＋1，逆に翻訳結果が誤りへ変化した場合には誤変換度数を＋1 とする．また，ルールの適用によって翻訳文の正誤に変化が生じ無かった場合には，正変換度数，誤変換度数は変化しない．

2.4.2　ルール適用正解率と適用条件

ルール適用正解率は，以下の式によって計算される．

$$\text{ルール適用正解率}=\frac{CF}{(CF+EF)}\times 100\ [\%]$$

ここで，**CF** は正変換度数，**EF** は誤変換度数である．このルール適用正解率によって，以下のようにルールの適用条件が決定される．

- ルール適用正解率 80%以上
  差異部分のマッチ
- ルール適用正解率 80%未満 50%以上
  差異部分および前後 2 単語ずつの品詞のマッチ
- ルール適用正解率 50%未満
  差異部分および前後 2 単語ずつの単語のマッチ

2.5　学習処理部

図 2 にルール獲得例を示す．

学習処理部では，入力英語文の形態素解析結果と，人手により作成された正解文を比較し，差異部分を抽出することによりルールを獲得する．ここで，共通部分とは 3 単語以上連続して同一である部分，差異部分とは共通部分にはさまれている

部分と定義する．

ルールは，差異部分，差異部分前後の単語と品詞，正変換度数，誤変換度数から成る．なお，獲得された時点では，正変換度数 1，誤変換度数 0 ではなく，正変換度数，誤変換度数共に 1 とする．これは，本稿で提案するような既存のシステムに対して付加的に処理を行う補助システムの場合は，いかに多くの正解を作るかということよりも，いかに誤りを抑えながら正解を作るかということに重点を置くべきであると考えられるからである．このような観点から，2.4.2 で述べたルール適用正解率が初期状態において中間の適用条件である 80%未満 50%以上となるようにそれぞれの数値を設定した．

3 正解文について

ここで，翻訳結果に含まれる誤りのうち，本稿において前処理の対象とするものを述べ，次に，システムの学習処理部で行われる帰納的学習に用いられる正解文の作成方法を示す．

3.1 前処理の対象

機械翻訳結果における誤りのうち，本稿での前処理の対象を，

1 訳語選択の誤り
2 係り受け関係の誤り
3 ユーザの入力の誤り

以上の 3 点に限定する．ここで，「ユーザの入力の誤り」とは “information” や “knowledge” 等の不可算名詞に対して複数形の “s” を付加しているものや，繰り返し用いられているスペルミスなど，入力時のユーザの意思と反する誤りである．

本稿では，意味が取れる範囲においては，翻訳文の自然性に関しては前処理の対象としない．これは，翻訳文の自然性という問題の改善は，同じ補助システムで考えた場合，前処理よりも後処理で行った方が格段に効果的なことが明白であるからである．これについては別に研究を行っている [6].

3.2 正解文作成のアルゴリズム

正解文は以下の手順により作成される．

Step1：入力英文の翻訳結果に誤りが存在する場合，入力英文に対して人手により正解日本語文を作成．

Step2：システムにより翻訳された翻訳文と **Step1** で作成した日本語文を比較し，誤っている箇所を特定．

Step3：訳語選択の誤りのうち、品詞分類としては誤っていないものを修正．

- **Step3.1**：正解文の中から訳語選択の誤りに対応する部分を，日英翻訳ソフトを用いて翻訳し，その結果を第一候補として適用．
- **Step3.2**：正解文の中から訳語選択の誤りに対応する部分を，日英翻訳ソフトの辞書を用いて探索．複数候補が存在する場合，最も多義性の少ないものから適用．

Step4：訳語選択の誤りのうち，品詞分類まで誤っているものを修正．

- **Step4.1**：**Step3.1** と同様
- **Step4.2**：**Step3.2** と同様

Step5：係り受け関係の誤りを修正

- **Step5.1**：係り受けが誤っている箇所周辺の各句の切れ目ごとに，順次 “,” を付加してみる．
- **Step5.2**：**Step5.1** 同様，“,” の付加の代わりに文の分割を行う．
- **Step5.3**：**Step5.1, 5.2** により文の構造がおかしくなる場合には，単語の付加を行い文構造の回復を試みる．

Step6：**Step3** により修正できなかったものがあった場合，**Step4, 5** によって修正可能となることがあるため，再び **Step3** を行う．

以上の操作を 1 文ごとに繰り返し行い，正解文を作成する．

4 実験

本手法の有効性を示すために，実際に実験システムを作成して実験を行った．

4.1 実験方法

実験には自然言語処理に関する 2 編の論文 **A, B**，それぞれ 189 文と 168 文，計 357 文を用いた．1 文あたりの平均単語数は 18.36 単語であった．システムの入力のための形態素解析には「**Brill tagger**」[7]を，機械翻訳システムには商用のシステムを用いた．初期条件を一定とするため，初期状態の変換ルール辞書は空の状態から行った．

表 1：実験前後のテキストの状態

		処理前	処理後
正解文数		151 文	194 文
翻訳正解率		42.30%	54.34%
翻訳不可能な文		10 文	7 文
誤り箇所総数		246 箇所	161 箇所
誤りの内訳	訳語選択	114 箇所	69 箇所
	係り受け	93 箇所	71 箇所
	ユーザの入力	39 箇所	21 箇所

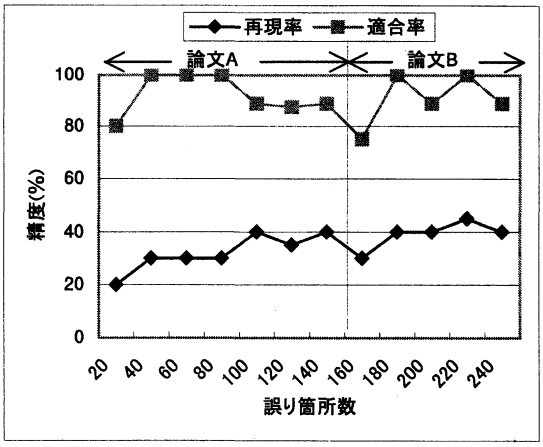

図 3：再現率・適合率の変化

4.2 実験結果

図 3 に実験結果を示す．図 3 は誤り箇所 20 箇所ごとに適合率と再現率を算出したものである．また，表 1 に実験の処理前と処理後のテキストの変化を示す．

5 考察

図 3 の実験結果のグラフから，初期状態で変換ルール辞書は空であるため精度が低いが，帰納的学習により徐々に向上していくことが確認できる．誤り箇所数 150 近辺で精度の低下が見られるが，これは入力テキストの種類の変わり目による影響である．しかし，その後学習を続けることにより精度の回復が見られる．

実験結果を誤りの種類ごとに見てみると，訳語選択の誤りとユーザの入力の誤りは比較的よく改善されている．これは，これらの種類の誤りは差異部分の長さがあまり長くないため，ルールの適用回数が多かったことが要因である．これに対して，係り受け関係の誤りは差異部分の長さが長いため，ルールの適用回数が少ない．また，ルールが適用された場合にも，変換部分前後の状況により翻訳誤りとなるケースも見られた．このことから，係り受け関係の改善には差異部分付近の情報だけではなく，文全体から幅広く情報を得ることが必要であると考えられる．

6 おわりに

本稿では，機械翻訳システムへの入力の前に，入力文に処理を施すことにより翻訳精度を向上させる手法を提案し，実際に実験システムを作成して評価実験を行った．その結果は，246 箇所あった翻訳誤りのうち 85 箇所の改善に成功し，システムへの入力により翻訳結果が誤りとなったものは入力文 357 文に対して，わずかに 8 箇所と非常に良好なものであった．また，入力テキストの種類が変化しても，再び精度が回復することからシステムの適応能力も確認できた．

本手法では特定の機械翻訳システムに固有の情報は一切使用していない．このことから他の機械翻訳システムにも汎用的に用いることが可能であると考えられる．今後は，本手法の汎用性を実証しつつ，さらに精度の向上を図る予定である．

参考文献

[1]Yamauchi Satoshi：A Method of Evaluation of the Quality of Translated Text, MT Summit Ⅶ, pp564-567, Sept.1999

[2]成田一：翻訳ソフトの性能評価，情報処理学会研究報告，自然言語処理 研究報告 No.125-14,pp123-130,1998

[3]佐藤理史：論文表題を言い換える，情報処理学会論文誌，Vol40，No7，pp2937-2945(1999)

[4] 張玉潔，尾関和彦：分類木を用いた日本語長文の自動分割，言語処理学会第 4 回年次大会発表論文集，pp390-393(1998)

[5]田中穂積 監修：“自然言語処理—基礎と応用—”電子情報通信学会編，1999

[6]尾崎正行，荒木健治，栃内香次：帰納的学習を用いた自然な日本語文生成手法の評価，情報処理学会研究報告，自然言語処理 研究報告 No.141-10,pp57-62,2001

[7]Eric Brill：A Corpus-Based Approach to Language Learning, University of Pennsylvania, 1993