帰納的学習による機械翻訳手法への接続情報の導入

5H－8

西方 弘樹† 荒木 健治†† 宮永 喜一† 栃内香次†
† 北海道大学工学部 †† 北海学園大学工学部

1 はじめに

我々は機械翻訳の一手法として，帰納的学習による学習型機械翻訳手法を提案し，また実験システムの構築を行なっている[1][2].

本手法は，構文及び意味規則に基づいて解析主導で翻訳を行なうものではなく，文例に基づく文の変換によるものであり，入力文との距離が最小である既知の文例に適用される変換ルールを適用して翻訳するという考え方に基づいている．そして，ルール適用をより一般化するために，入力文から変数スロット付きルールを抽出し，辞書中のルールとの比較を繰り返すことによりルールの一般化を行なっている．

我々はこれまで本手法の利点と問題点を明らかにすることを目的とした評価実験を行ない[4]，その結果から，部分的に解析的知識を用いてルール抽出時に分割点に制限を設けることにより誤ったルールの抽出を抑制する，という改良を行なってきた．

本稿では，翻訳部において変数スロットへの誤ったルールの適用によって生じる誤翻訳の減少を目的として，変数スロットに対してルールに関する接続情報を用いることにより，より適切なルールの適用を行なうという改良手法について述べる．

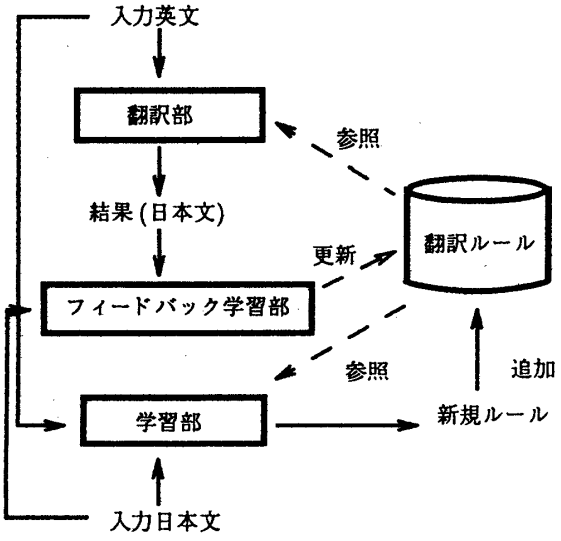

図1：処理のながれ

Use of Conjunctive Relation in Machine Translation by Inductive Learning
Hiroki Saihou, Kenji Araki, Yoshikazu Miyanaga, Koji Tochinai
†Faculty of Engineering, Hokkaido University
††Faculty of Engineering, Hokkai-GakuenUniversity

2 帰納的学習による機械翻訳手法

2.1 処理の流れ

図1に帰納的学習による機械翻訳手法[1]の流れを示す.

2.2 学習処理

2つの翻訳例の共通部分と差異部分とを多段階に抽出することで翻訳ルールを獲得する．新規に入力された翻訳例を既知の翻訳ルールと全て比較して，差異部分が一意に対応づけられるとき，この差異部分と，差異部分を変数として付加した共通部分を翻訳ルールとして抽出する．抽出された翻訳ルールを既知の翻訳ルールと比較して，共通部分・差異部分が抽出されなくなるなるまで抽出を繰り返す．このとき，過剰な抽象化を防ぐために抽象化に閾値を設ける[2].

以下に基本的な例を示す．ここで，(1)(2)の2つの対訳が与えられた場合(3)(4)(5)が翻訳ルールとして抽出される．

I *play tennis* .;私 は テニス を する . (1)

You *play tennis* .;あなた は テニス を する . (2)

⇓

I; 私 (3)

You; あなた (4)

@0 *play tennis* .; @0 は テニス を する . (5)

学習部ではこの他，差異部分だけの抽出や，共通部分だけの抽出，また，異表記ルールの抽出及びそれを用いた有効なルールの抽出，誤ったルールの削除を行なう[3].

2.3 翻訳処理

翻訳部では学習部により抽出された翻訳ルールを用いて翻訳文を作成する．入力文に適用できる最も具象的なルールを始めに適用し，続いて，適用された翻訳ルールの変数部分に適用できる最も具象的なルールを適用していく．複数の翻訳結果が得られた場合，フィードバックデータを基にして優先順位を決定する.

なお，フィードバックデータについては翻訳結果の正誤により，フィードバック学習部において度数の更新処理を行なう.

3 変数スロットへの接続情報の利用

本システムの問題点の一つとして，翻訳部において変数スロットへルールを適用する際に，本来適用されるべきでないルールを適用してしまうことによる誤翻訳がある．

これは，本手法ではフィードバックデータとして，親ルール番号，正翻訳度数，誤翻訳度数，使用した異表記ルール番号だけを利用し，またルールの淘汰には正・誤翻訳度数のみを使用しており，変数スロットへの他のルールの接続の強さを考慮していなかったためである．以下に，これによって生じる誤翻訳の例を示す．

$$some\ @1\ ;\ 数本の\ @1 \tag{6}$$

$$trees\ ;\ 木 \tag{7}$$

$$some\ @2\ ;\ いくらかの\ @2 \tag{8}$$

$$milk\ ;\ 牛乳 \tag{9}$$

$$\Downarrow$$

$$some\ milk\ ;\ 数本の\ 牛乳 \tag{10}$$

ルール (6) にルール (7) を適用した場合，(8) に (9) を適用した場合はそれぞれ正しいが，ルール (6) にルール (9) を適用した翻訳例 (10) は誤翻訳となる．

これは，一般化したルールの変数スロットに他のルールを適用する場合,「変数スロットが親ルールにおいてはどの様なものであったか」という情報を用いて，それからあまりにかけ離れたルールの適用は行なわないことによって解決できると考えられる．

これに基づき，本稿では以下に示すように変数スロットへの他のルールの接続に関する情報を利用する手法を提案する．

学習部

- ルール抽出時に，変数スロット数が 1 以上のルールについては，親ルール番号，使用異表記ルール番号の他に，

[スロット番号,接続ルール番号,正翻訳度数,誤翻訳度数]

からなる接続データを抽出する．
- その接続ルールを親ルールとする子ルールについても，同様に接続データとして登録する．

翻訳部

- 変数スロットにルールを適用する場合，まず接続データ内から探す．データ内のものが適用不可の場合には，他のルールを適用する．

フィードバック学習部

- 正翻訳となった場合には，接続データが既に登録されているものについては，その正翻訳度数を増加させる．接続データが登録されていないルールの場合は，新たにデータを登録する．
- 誤翻訳となった場合で，接続データが既に存在するものだけを使用して翻訳文が出来ている場合にはその接続データの誤翻訳度数を増加させる．

また，以上の改良により誤ったルールのフィードバック処理の改善も見込まれる．

従来のシステムのフィードバック部では上で述べたような処理を行なっていなかった．そのために，誤翻訳となる場合に多く見られるものであるが，翻訳文生成の際に使用したルールのうちある少数のものが正しくない場合には，他の大部分のルールに対しても「誤翻訳」としてフィードバック処理してしまうという問題点があった．その他，(10) 式で示したようにルールそのものが間違っているのではなく，組合せだけが間違っている場合にも正しいフィードバック処理が行なわれないという問題があった．

本稿で提案した改良手法によれば，フィードバック処理については変数スロットに対するその他のルールとの接続の強さの情報について処理するため，上記 2 点のような問題は起こらないと考えられる．

4 おわりに

本稿では，帰納的学習による機械翻訳手法への接続情報の導入について提案した．上で述べた手法を用いることにより，変数スロットへのルールの誤適用による誤翻訳が減少できる．また，ルールの淘汰に関しても改良が見込まれる．

今後の課題としては，変数スロット数 0 のルールについて，単語数の多いルールはそのまま同じ単語列が出てくることが出現することは少ないので，そのルールの中で接続に関して代表的となる語を幾つか抽出することにより，ルールの適用を多少一般的にするという改良などが考えられる．

その他，同じ変数スロットに入りやすいルールをまとめることにより，ルールのグルーピングを行なうことも考えられる。

参考文献

[1] 荒木健治，栃内香次：多段階共通パターン抽出法を用いた翻訳例からの帰納的学習による翻訳，情報処理北海道シンポジウム講演論文集，pp.47-49,(1991)

[2] 内山，荒木，宮永，栃内：帰納的学習による機械翻訳手法の評価実験，情報処理学会研究報告,NL93-4,pp.23-30,(1993)

[3] 内山，荒木，宮永，栃内：帰納的学習による機械翻訳手法の改良，電気関係学会北海道支部連合大会講演論文集,p397,(1993)

[4] 西方，荒木，宮永，栃内：解析型および学習型機械翻訳手法の評価，情報処理北海道シンポジウム講演論文集，pp.15-17,(1994)

[5] 西方，荒木，宮永，栃内：帰納的学習による機械翻訳手法への解析的知識の導入，電気関係学会北海道支部連合大会講演論文集,p179,(1994)