ルールベース翻訳とパターンベース翻訳の融合

長瀬友樹 小玉修司 小屋岡剛一 塩津誠

富士通株式会社

1. はじめに

従来の機械翻訳システムの多くは「ルールベース翻訳」と呼ばれ，原文解析，訳文生成，トランスファーの各過程で，専門の開発者が記述したルールに基づいて翻訳を行う．ルールベース翻訳は言語事象を一般化するのに適しており直訳的な翻訳を少数の規則で記述できるという利点がある反面，例外的な翻訳や意訳を必要とする翻訳のルール化は一般に効率が良いとは言えない．

原文と訳文のペアをパターン化して登録するだけで望みの翻訳結果を得ることができる「パターンベース翻訳」が提案されている [1][5]．パターンベース翻訳では，ルールベースのような難しい規則は必要なく，システムや文法について詳しくない者でも訳質を改善していくことができる．

ルールベース翻訳とパターンベース翻訳を比べたとき，開発効率，保守性の観点からパターンベースに利があることは明らかである．しかし，語彙に依存しない基本的なパターンについては，パターンベースといえども専門家による整備が必要であり[5]，この点はルールベースと変わらない．また，パターンベースで既存のルールベース・システムと同等な処理精度を得るためには膨大な数のパターンが必要であり，処理の高速化が大きな課題となっている [3]．

本論文では，既存のルールベース翻訳とパターンベース翻訳を効率良く融合させる方法について提案する．語彙に依存しない基本的な翻訳は従来のルールベースの枠組みを利用し，語彙に依存した例外的な翻訳は用例パターンによって処理する．パターンベースの利点に加え，既存のルールベースの資産をほとんど変更せずに活用できるというメリットがある．全文で一致したときのみパターン翻訳が働くのではなく，用例パターンが原文の一部にマッチするような場合にも，マッチした部分をパターンベースで，それ以外の部分をルールベースで翻訳する．また，用例パターンが大規模になっても高速処理が可能である．

2. システムの概要

本研究で作成した機械翻訳システムの構成を図1に示す．用例パターンは，構文解析結果の統語構造（解析木）に対してマッチングが試され，マッチングに成功したパターンに基づき中間構造（意味ネットワーク）を変換する．その後は，従来のトランスファー過程，生成過程を経て訳文が出力される．

2.1 用例パターン

用例パターンは，英文とそれに対応する日本文の変数付きパターンのペアとなっている．以下に用例パターンの例を示す．

```
S : It amazed <N1> to learn that <S1>
    <=> S : <S1>と知って<N1>はびっくりした
          # N1=HUM
S : <N1> turned off from <N2> into <N3>
    <=> S : <N1>は<N2>からそれて<N3>に入った
```

"<","<">で囲まれた部分（<N1>,<S1>）は変数部で，他の文字列に置き換え可能な部分を表している．変数部の記号（N,S）は統語カテゴリ（名詞句，文）を表し，その統語カテゴリの部分木が存在する文字列範囲とのみマッチングすることができる．統語カテゴリの右の数字は，英語側変化部と日本語側変化部の対応関係を表している．英文側パターン，日本文側パターンの先頭の記

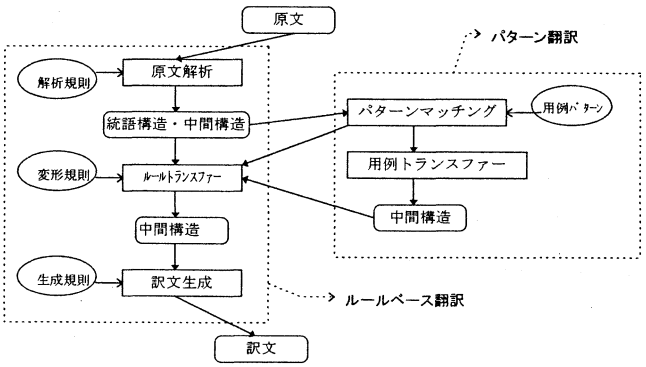


図1 システムの構成

号（S）はそれぞれパターン全体の統語カテゴリを表している．各用例パターンは，意味素性によって変数マッチングに制約を与えることができる．例の中の“N1=HUM”はN1にマッチするフレーズが「人間」の素性を持っているいることを表している．

処理対象の用例パターンは，語彙に依存した翻訳知識としている．つまり，用例パターンには，固定部（変数部でない部分）に少なくとも1語の名詞，動詞，形容詞，または副詞が含まれていなければならない．次の例のように固定部に特定の語彙が含まれないパターンは，ここでは扱わない．

S：<N1><V1><N2> <=> S：<N1>は<N2>を<V1>

2.2 用例パターンの単語辞書への格納

用例パターンは，単語辞書中に記述されており，原文の辞書引きのタイミングで検索される．用例パターンを単語辞書へ格納することにしたのは，以下のメリットを考慮したためである．

(1) 高速処理

本システムで使用する用例パターンは，パターン中に少なくとも一語の名詞，動詞，形容詞，または副詞を含んでいなければならない．したがって，ある1つのパターンが適用される確率は極めて低く，パターンが翻訳率の向上に貢献するためには大量のパターンを登録しておく必要がある．用例パターン数が数十万～数百万に達しても実用に耐えられるよう，処理性能には留意する必要がある．

本システムでは，用例パターンを単語辞書中に置き，単語表記をキーにして用例パターンを検索するため，原文に関係ない大部分の用例パターンをあらかじめマッチングの対象から排除することができる．パターンを格納する単語は，原文パターン中の名詞，動詞，形容詞，副詞の中から一語が選ばれる．

パターンを辞書に置いた場合のトランスファー処理時間は，辞書中の用例パターンの総数に比例する．チャート法やearly法のアルゴリズムが文法のサイズの2乗に比例する計算時間がかかることを考えると，用例パターンをパージングの段階で利用する方式に比べ，効率面で有利といえる．

(2)適用パターンの限定

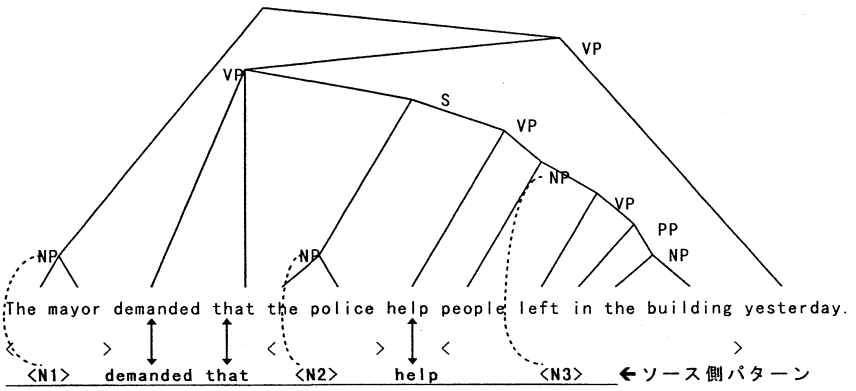
用例パターンは，分野別・文種別の専門用語辞書に分散して格納することも可能である．つまり，専門用語辞書と同じように，翻訳対象の文タイプに合わせて，用例パターンを選択して利用することができる．これにより，特定分野の用例パターンが他分野で誤適用されることがある程度防止できる．

3．用例トランスファー

3.1 パターンマッチング

用例パターンの原文解析木へのマッチングは，以下の手順で試される．

1　英語パターンの固定部がすべて順序どおりに原文に含まれるかチェックする．
2　1のチェックを通ったパターンが複数ある場合は，固定部の長い順にパターンをソートする．3以降の処理は，この固定部の長い順にパターンの適用が試される．



<N1> = The mayor
<N2> = the police
<N3> = people left in the building

図2　パターンマッチング

3　変数部に対応する原文文字列が原文の解析木の中で部分木を構成しており，そのカテゴリが変数部で指定したカテゴリと一致することを確認する．

パターン左端（または右端）が変化部の場合，原文左端（右端）から一語づつ始点（終点）を右（左）にずらしながら部分木の存在を確認し，その統語カテゴリがパターンで指定されたカテゴリと一致した場合，その部分木を構成する文字列が変化部に対応しているとみなす．

4　変数部に対応する文字列候補が変数部の制約条件を満たしていることを確認する．

5　3の条件を満たしたパターンが，先に適用が決まっているパターンの固定部・変化部の境界と交差しないことを確認する．

以上の手続きを辞書引きされたパターン候補すべてについて行い，マッチしたパターンがあった場合にはパターンの固定部および変化部が原文のどの範囲に対応しているかに関する情報を保持しておく．（この情報は次の中間構造の変形で参照される）

図2はパターンマッチングに成功する原文とパターンの例である．<N1>と<N3>はパターンの左端と右端であるので，これら変数部に対応する原文範囲はパターンマッチングの過程で決定される．<N1><N3>ともに名詞句（N）の指定なので，"mayor"を右端とする名詞句の部分木，"people"を左端をする名詞句の部分木をサーチし，それぞれ"The mayor"と"people left in the building"が対応していることがわかる．

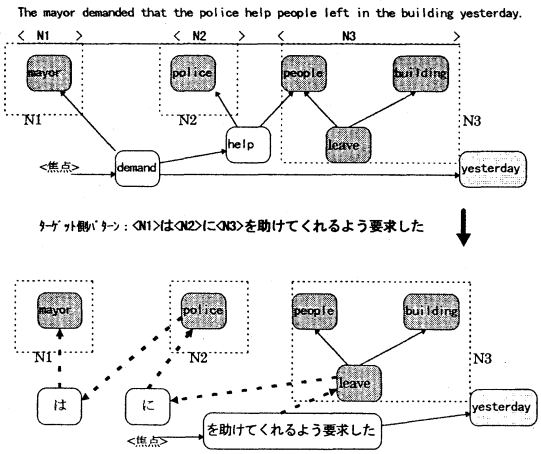


図3　中間構造の変形

3.2 用例パターンに基づく中間構造の変換

ここでは我々が実験で利用している商用システムATLASの中間構造（意味ネットワーク）を用いて説明することにする．

中間構造の変形は次の手順で行う．

1　原文パターンの固定部に対応する意味ノードと，その意味ノードにつながっている意味関係詞をすべて消去する．

2　訳文パターンの固定部に対応する意味ノードを新たに生成する．

3　訳文パターンの最後の意味ノードから最初の意味ノードに向かってカスケード状に意味関係詞で結ぶ．

4　パターン外の意味ノードとパターン最後の意味ノードを関係詞で結ぶ．(関係詞名は原文解析結果と同じにする)

5　パターン最後の意味ノードを中心ノードとする（<焦点>関係詞を立てる）

用例パターンの変数部および用例パターン以外の部分に対応する中間構造は，構文解析後の構造がそのまま残るので，変数部分とパターン外の部分は従来どおり翻訳される．

また，一文で用例パターンが複数適用された場合も，この方法を再帰的に適用すれば矛盾なく処理することが可能である．

4．性能評価

次の3つのケースで翻訳開始から終了までの所要時間を測定した．

①用例パターンなし

②用例パターン 8,000 個

③用例パターン 40,000 個

評価用の用例パターンは市販辞書の例文をもとに人手で作成したものを使用した．表1に英日翻訳で英語文 1,000 文の翻訳処理時間を測定した結果を示す．

②から③で用例パターン文法数が5倍になっているのに対して，処理時間の増加率は3倍弱である．この評価結果を見る限り，数万規模の用例の登録では問題になるほどの性能の劣化は起きないと言える．

表1　パターン数の違いによる性能比較

	翻訳所要時間	率(%)
①なし	840 sec.	100
②8,000 個	876 sec.	104
③40,000 個	925 sec.	110

5. 課題

用例トランスファーを使った場合と使わない場合で翻訳結果の品質を比較したところ，いくつかの文で訳質の低下が起きた．

(1)パターンの誤った適用

パターンの固定部分に含まれる単語が少ない，あるいはパターンが非常に一般的な並びであるために，本来マッチしてはならない文にマッチしてしまい，不必要なパターンを使って翻訳される例が目立った．訳質が低下した文のほとんどがこの原因による．

原文：　We waited for two hours,but he didn't show up.

ﾊﾟﾀｰﾝ：<N1> waited for　<N2> ⇔ <N1>は<N2>を待っていた．

訳文：我々は2時間を待っていたが，彼は現れなかった．

(元の訳：我々は2時間待ったが，彼は現れなかった．)

(2)従来のトランスファーが効かない

本手法は原文・訳文の対応する変数部分は独立して翻訳が可能なことを前提にしているが，変数内の情報が変数外の訳文生成に影響を与える場合がある．この場合，従来のトランスファーがうまく動かず訳質の低下を引き起こすことがある．

原文：　He acknowledged that no major results came out of the task.

ﾊﾟﾀｰﾝ：<N1> came out of　<N2>　⇔ <N1>は<N2>から出てきた．

訳文：彼は主要な結果でなくは協議から出てきたと承認した．

(元の訳：彼はどんな主要な結果も協議から出てこないと承認した．)

(3)複文で語尾が乱れる

現段階のシステムは時制・アスペクト等による訳語の活用語尾変化を考慮していないため，複文に用例パターンがマッチした場合などにしばしば語尾が不自然となる．

原文：　One day, as his face burned with shame, he ran away.

ﾊﾟﾀｰﾝ：<N1> burned with shame ⇔ <N1>は恥ずかしくてほてった

訳文：ある日，彼の顔は恥ずかしくてほてったながら彼は逃走した．

(元の訳：ある日，彼の顔は恥をもって燃えながら彼は逃走した．)

6．終わりに

ルールベース翻訳によるこれまでの資産を十分に活かしつつ，語彙に依存した特殊な翻訳をパターン翻訳で実現する手法について述べた．用例パターンの辞書内格納により大量パターンでも速い処理速度が保たれること，中間構造を対象とした変換で再帰的なパターンの適用が可能なことを示した．将来的には，従来ルール（解析文法，トランスファー）に含まれる語彙依存規則を用例パターンへシフトできれば，ルールがシンプルになり，システム全体の保守性が向上することも期待できる．また，用例パターンは高度な文法的知識がなくても記述できるため，ユーザによるパターンの追加も可能であり，ユーザチューニング機能としての利用価値も高い．今後は，誤適用をなくすための制約条件の強化，固定部の活用語尾処理の実装，対訳コーパスからの用例パターンの自動作成[2][4]などが課題である．

参考文献

[1] Furuse O, Iida H. Cooperation between Transfer and Analysis in Example-Based Framework. In *Proceedings of the 16th International Conference on Computational Linguistics(COLING-92)*, pp645-651, 1992.

[2] Kaji H, Kida Y, Morimoto Y. Learning Translation Templates from Bilingual Text. In *Proceedings of the 16th International Conference on Computational Linguistics(COLING-92)*, pp672-678, 1992.

[3] Takeda, K. Pattern-Based Context-Free Grammars for Machine Translation. In *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics*,pp.144-151, 1996.

[4] 池原悟，白井諭，相沢弘．和語動詞に対する日英対訳例文の収集について．言語処理学会 第2回年次大会論文集，pp.253-260, 1996.

[5] 渡辺日出雄，武田浩一．パターンベース翻訳システム:PalmTree．情全大，第2巻，pp80-81, 1997.