パタン翻訳による翻訳精度向上の可能性と問題点

大倉 清司 †, 潮田 明 †, 富士 秀 †, 小玉 修司 ‡
† (株) 富士通研究所, ‡ 富士通（株）

1　はじめに

機械翻訳システムの研究開発では、翻訳対象分野を限定し、その対象分野に頻出する語句を幅広く登録することによって翻訳精度を効率的に向上させられることがわかっている。本研究では、分野特有の語句に加えて分野特有の表現も登録し、これによる翻訳精度向上の定量評価を行った。特有表現の登録には「パタン翻訳」の枠組みを利用した。評価の結果、分野特有表現をある一定の手順で登録することにより、これまで語句登録だけでは解決できなかった問題点を解決し、翻訳精度を大幅に向上させることができた。

2　パタン翻訳とその問題点

パタン翻訳システム [1,5] とは「（用例）パタン」と呼ばれる、日本語と英語の対応関係を翻訳システムに組み込んで翻訳するものである。本節では、パタン翻訳の概要を説明し、従来の問題点を挙げる。

2.1　パタン

パタン翻訳では様々な種類のパタンが提案されているが、ここでは語彙特有の表現（慣用表現や動詞の格フレーム等）による訳し分けを、原言語側表現と目的言語側表現の対で表したものをパタンと呼ぶ。語彙特有の表現を多く含む対訳文を抽出しておくと効率よくパタンを作成できる [4]。以下にパタンの例を示す：

```
The prices are up by <N1:5> percent.
<==> 価格は<N1:5>%増加する。
```

'<''>' で囲まれた部分を変数部と呼び、その部分は他の文字列に置き換え可能であることを表す。

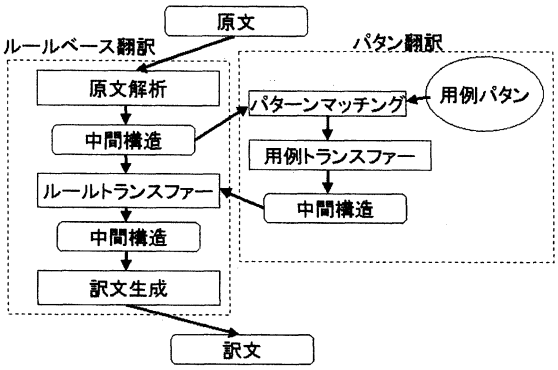

図 1: システム構成

英語の変数部、日本語の変数部を対応づけるために、ラベル（この場合、'N1'）をつけている。上のパタンを登録しておくことにより、例えば

```
The prices were up by 13.2 percent.
```

という文を、

価格は13.2%増加した。

と翻訳させることができる。構造によるマッチングをしているため、パタンが現在形でも、過去形の文にもマッチし正しい訳が出せる。

2.2　パタン翻訳システムの概要

本研究で使用した翻訳システムを図1に示す。用例パタンは、表層レベルではなく、構造的にマッチングが試され、マッチしたパタンに基づき中間構造を変換する。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 [1]。

2.3 従来の問題点

パタン数をただ多くつくるだけでは、文によっては翻訳品質が下がってしまうことがあった [1]。その原因は、

1. 大量につくってもマッチするパタンが少ない
2. 意図しないパタンがマッチする
3. マッチしてパタンが適用されても、長文の翻訳はよくならないことがある

といえる。分野を限定しないと表現は多岐にわたるため、パタンを大量につくってもなかなかマッチしない。

意図しないパタンがマッチする原因は、パタンの原言語側表現に対し、目的言語側が過度に固有な表現であるためである。例えば、

```
<N1:That plant> lives in <N2:Africa>.
<==> <N1: その植物> は <N2: アフリカ> に生息する。
```

というパタンをつくると、

```
Tom lives in Kawasaki.
```

という文の翻訳が

トムは川崎に生息する。

となってしまう。

パタンが適用されても長文の翻訳がよくならないとは、パタンがマッチしたところは局所的によくなっても、他の部分で訳文の意味が通らないところがあると、文全体としての訳質はさほど変わらないということである。

3 問題点の解決

小規模な試行実験により、上記問題点を解決するようなパタン作成方法を考案し、この方法に則って実際に分野特定パタンを作成した。この方法について以下に述べる。

1. パタンをつくる対象の分野を限定した
2. 原言語側において意図した通りにマッチするように、変数部の少ないパタンをつくった
3. 長文に対応できるように、長文用のパタンを重視した

分野を限定することにより、頻出する表現をパタンとしてつくることが可能となった。例えば、新聞報道記事は、引用表現が多いので、say や tell、announce などの動詞を中心にパタンを作成することができる。

意図した通りにパタンをマッチさせるには、一般的なパタンをつくるのではなく、その分野に適した、変数部の少ないパタンをつくる必要がある。これにより、マッチすべきではないパタンがマッチして翻訳が低下する、という従来の問題が解決される。

長文用のパタンとは、上に述べた、announce など、節をとる動詞のパタンのことである。頻度が高いため、マッチ率も高く、翻訳品質向上に役立つと考えた。

4 評価実験

4.1 評価方法

本研究では、英日方向の翻訳に限定して実験を行った。

新聞報道記事の英文をトレーニングセット約 15000 文とテストセット約 2500 文に分け、トレーニングセットで頻出する語句（単語・複合語）およびパタンを登録し、テストセットの英文（今回はそのうち 306 文）で評価した。パタン登録によって翻訳のどの部分が向上するかを調べるため、「日本語の自然さ」「語句の訳し方」「構文」の 3 つの観点から評価を行なった。もともとの翻訳システムに対し、語句だけを登録したときと、語句＋パタンを登録したときとで、どのくらい精度が向上したかを評価した。

4.2 評価基準

翻訳文に対して、「日本語の自然さ」「語句の訳し方」「構文」という 3 つの観点から、A,B,C ランクの 3 段階で評価した。

- 日本語の自然さ： 翻訳システムが出力した日本語がどれくらい自然か。A ランク：ほぼ自然な日本語。B ランク：やや不自然な日本語

だが意味が通る。Cランク：日本語として意味が通らない。

- 語句の訳し方： 訳語に着目し、どのくらいうまく単語や複合語などを訳しているか。文の中心となる語句がうまく訳せていれば、文の枝葉的な語句が多少間違えていてもよいとする。逆に、中心要素の訳語が失敗しているとランクを下げる。Aランク：ほぼ問題ない訳語。Bランク：おおまかな線はあっているが、多少難あり。Cランク：訳語が間違えている。

- 構文： 意味的に構文がきちんととれているか。日本語から判断するので、実際に解析で正確に解析されていても、生成で助詞の使い方などがおかしいとマイナスの要因になる。語句の訳し方と同様、文の中心的な構文が違っていれば大きなマイナス要因とする。Aランク：構文がしっかりとれている。Bランク：細かいところは違うが、許容範囲。Cランク：構文がとれていない。

各観点について語句、パタンを登録した結果どのくらいポイントが変化するか評価した。Aランクは2ポイント、Bランクは1ポイント、Cランクは0ポイントとして数え、1文について2ポイントを満点として評価した。

4.3 評価結果

3800語句、141パタンを登録したところ、以下のような結果になった：

結果：
評価セット306文中、
　語句登録のみで訳に変化があったもの： 134文（評価セット中43.8%の文）
　パタンにより訳に変化があったもの： 19文（評価セット中6.2%の文）

語句登録で訳が変わった文の翻訳率：図2
（もとのシステムからの向上度,134文を比較）：
自然さ： +2.2%
語句： +10.0%
構文： +3.0%

パタンで訳が変わった文の翻訳率：図3
（語句登録したシステムからの向上度,19文を比較）：
自然さ： +10.6%
語句： -2.6%
構文： +18.4%

全体の翻訳率：図4
（もとのシステムからの向上度,306文を比較）：
自然さ：
　語句登録： +0.9%
　語句登録＋パタン： +1.6%
語句：
　語句登録： +4.4%
　語句登録＋パタン： +4.2%
構文：
　語句登録： +1.3%
　語句登録＋パタン： +2.4%

図2は、語句登録によって翻訳が変わった文についての変化を表したものである。語句を登録すると、語句の訳し方が向上しているのは当然だが、構文の精度も向上している。これは、複合語などの登録により、名詞句の解析がうまくいっているなどの要因による。

図3は、語句を登録したシステムに対して、パタンの登録が翻訳にどう影響するかを示したものである。パタンによって翻訳が変わった文について、3つの観点から比較した。パタン登録により、自然さ、構文の精度が向上している。特に構文の向上が著しい。これは、助詞など、語句登録ではなおせない部分の修正ができたためだと考えられる。語句の登録だけだと、翻訳精度はそこそこは上がるが限界があり、表現レベルの登録ができると、大幅な精度向上が期待される、という研究がある[2]。パタンはまさに「表現」の登録であり、その結果として「自然さ」「構文」の観点において精度が大幅に向上した。

語句の観点ではパタンの登録により精度が落ちている。これはパタンの副作用である。パタン翻訳により、語句の訳し分けがうまくいかなくなったのではなく、パタン作成時に意図していなかった文がテストセットにあったためである。

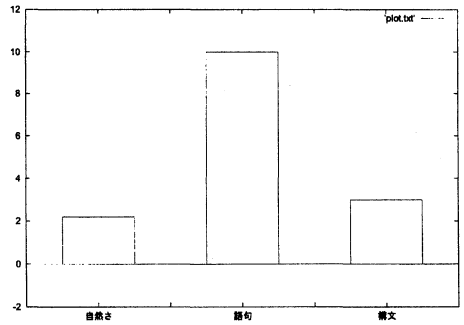

図 2: 語句登録による翻訳精度の向上度（もとのシステムとの比較）

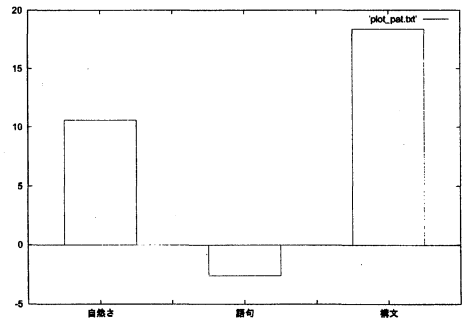

図 3: パタン登録による翻訳精度の向上度（語句を登録したシステムとの比較）

テストセット全体の翻訳として、語句を登録したシステムと、語句＋パタンを登録したシステムを比較したのが図 4 である。パタンにより、語句登録だけでは向上できないところをカバーしている。

4.4 さらなる精度向上のために

語句やパタンである程度、訳語、構文がよくなっても、ベースとなる構文解析が違っていると全体として翻訳がよくならない。

例えば、並列構造を含むような複雑な構造の文では、全体の構文解析を誤る可能性が高い。このため、パタンがヒットして訳が部分的によくなっても、文全体の意味が通らないことには変わりなく、翻訳精度向上には結びつかない。

構文はパタンによりある程度は対処できるが、その他に構文解析自体の精度を向上させることが重要

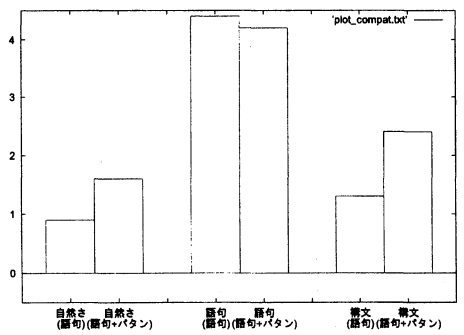

図 4: 語句＋パタンによる翻訳精度の向上度（もとのシステムとの比較）

である。

5 まとめ

分野を限定し、その分野専用の語句およびパタンを登録することにより、翻訳精度を向上させることが可能であることがわかった。特に、パタンは、語句登録だけでは実現できない部分の精度向上に貢献することがわかった。構文解析の精度を向上させれば、さらなる精度向上につながるだろう。

参考文献

[1] 長瀬友樹, 小玉修司, 小屋岡剛一, 塩津誠. ルールベース翻訳とパターンベース翻訳の融合. 言語処理学会 第 4 回年次大会論文集,pp.496-499,1998.

[2] 富士秀. 英日機械翻訳の訳質に関する評価実験. 言語処理学会 第 3 回年次大会論文集,pp.27-30,1997.

[3] 日本電子工業振興協会. 機械翻訳システム評価基準, 1995.

[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. 用例ベース処理を用いたパターンベース翻訳システム. 言語処理学会 第 4 回年次大会論文集,pp.488-491,1998.