スクリプトを用いたハイブリッド翻訳処理

内野 一　古瀬 蔵*1　大山 芳史*2　白井 諭*3

*1ＮＴＴサイバーソリューション研究所
*2ＮＴＴコミュニケーション科学基礎研究所 *3ＡＴＲ音声翻訳通信研究所

1. はじめに

機械翻訳処理においては、定型的な文章をパターンとして捉えて翻訳を行なうテンプレート型の翻訳方式、及び単語の意味を捉えて文章を解析して翻訳を行なう意味解析型の翻訳方式などがある。

意味解析型翻訳方式は、現段階では翻訳精度が十分とはいえず、ユーザの好みに沿った形にカスタマイズを行なうことも難しいが、その対象とする範囲が広いことから、外国のWWWの文章を読む場合のように不完全な訳であっても利用者が概要を把握できるレベルでよいと割り切る場合などに使用されることが多い。

それに対して、テンプレート翻訳方式は、ある程度タスクが限定され業務として最終的に翻訳結果を提供する必要がある場合、たとえば、翻訳者が大量のマニュアルなどを翻訳する場合などには、文書内に同じ表現や、定型的な表現が多く現れることや、ユーザの意図に沿った形での翻訳が行ない易いことから、翻訳できない文に関しては後で人手を加えることを前提とした翻訳支援として使用されることが多い。このような使用形態ではテンプレート翻訳や用例翻訳を中心とし、意味解析型の翻訳は補助とした構成のシステムが有効である。英語の経済ニュースの翻訳では原文の定型性に応じて翻訳テンプレート、経済専用文法、一般文法を切り替えて翻訳することで70%近い翻訳率が達成されている[相沢 96]。

我々は、実際の市況速報記事に対し、テンプレート翻訳[白井 97a]と意味解析型翻訳[池原 91]を並列に走行させ、自己評価の高い翻訳結果を組み合わせるという単純な構成によるシステムの評価実験を行ない、両方式を併用する場合における問題点の抽出を行なった。

本稿では、テンプレート型翻訳と意味解析型を別々に走行させた場合の問題点を延べ、スクリプトルールを用いて、両者を融合させたハイブリッド方式の構成によって、それを解決する手法を提案する。

2. 並列走行型翻訳方式の問題点

2.1. 並列走行型翻訳の適用結果

テンプレート型翻訳と意味解析型翻訳を並行して走行し、各々の翻訳方式ごとに定めた基準に従って自己評価を行ない、評価点の高い方を選択する方式で評価実験を行なった。

◇東証外国部・大引け
【ＮＱＮ】ニューヨーク株の大幅高を映し堅調。売買高は概算20万株。売買の成立した39銘柄（値付き率48.1%）のうち、値上がり18、値下がり3、変わらず1、比較できず17だった。アップル、モトローラが上げ、ＩＢＭが年初来高値に顔合わせした。ボーイングは年初来高値更新。半面、グラクソＷＬ、バークレイズ、ＣＳＨが下げた。
Tokyo Foreign Stocks Cls: Up on rally in N.Y.
Foreign stocks ended higher Friday in line with the overnight run-up on Wall Street. Turnover was estimated at 200,000 shares. Among 39 issues changing hands, 18 increased, three declined and one issue settled flat. No comparison was available for 17 stocks. Boeing marked a year's high. IBM matched its year's high. Other gainers included Apple Computer and Motorola. In contrast, Glaxo Wellcome, Barclays and CS Holding lost ground.

◇東証CB大引け・5年ぶり大商いで続伸
大幅高で14日続伸。株高から電機関連を中心に株価連動銘柄が買われたほか、債券高を材料に利回り銘柄も物色された。売買高は概算2000億円と90年5月16日の2300億円以来、約5年2カ月ぶりの大商い。ただ、利食い売りも目立ち、値上がり銘柄数395に対し値下がりも157とけっこうあった。住友精化(1)、富士通(9)(10)、ＮＥＣ(8)(9)が高かった。一方、東ガス(3)、関西電(3)は軟調。
Tokyo CBs Cls: Up in active trading
Convertible bonds registered their 14th consecutive day of gains Friday on trading centered around electric issues which track their underlying stocks and speculation in high-yielders. Turnover was a whopping 200 billion yen, the first time that trading volume has reached this level in 62 months. Despite the fact that gainers outnumbered decliners by 395 to 157, profit-taking resulted in losses for a considerable number of issues. The QUICK CB Index closed the day 1.55 points higher at 481.10. Sumitomo Seika Chemicals (No. 1), Fujitsu (Nos. 9 & 10) and NEC (Nos. 8 & 9) rose. Meanwhile, Tokyo Gas (No. 3) and Kansai Electric Power (No. 3) were weaker.

図1　日英の市況速報記事の例（１９９５年７月７日）

実験では、日本経済新聞社のテレコンデータベースから取り出し、[高橋 97]の方法により日英記事を対応付けた市況速報記事１４週分(1995 年６月～９月)のうち、東証外国部、大証、東証ＣＢを対象とした。対象とした記事の例を記事の図１に示す。市況速報記事、特に個別銘柄に関する文においては、定型的な文が極めて多いため、テンプレート翻訳が適用される割合が高く、最終的、東証外国部の記事では文単位で 90%、記事単位で 70%の合格率を得た[白井 97b]。

2.2. 問題点

実験においては、翻訳の正解を判定する際に文単位で翻訳結果が正しいかを判定し、合格率を算出した。しかしながら結果の検証のため、翻訳家による再チェックを行なったところ以下のような問題点があることが判った。

(1)テンプレート翻訳のルールの部において、特定の文の後にのみ適用すべきルールがある。

・「X 社が堅調。」

通常のテンプレート翻訳では “X held firm” と翻訳され、安定しているの意味となるが、「Y 社が買われた。」など上昇を意味する文の後に使われた場合、同様に上昇を意味する翻訳とする。

(2)同じテンプレートルールまたは同じ訳出となるルールを連続して適用する場合、動詞の訳を変えた方が自然となる。

・「X 社が買われた。」「Y 社が上げた。」

文としてみるとどちらの文も “～ was bought.” と翻訳して問題はないが、２文を続けて翻訳する場合、後者を advanced など別の動詞を使用して翻訳したほうが自然となる。

(3)意味解析型翻訳において前の文の解析失敗により、その後の文の翻訳時に不適当な主語などの補完が行われる場合がある。

・直前の文がテンプレート翻訳によって翻訳されているが、意味解析翻訳では係り受けの解析ミスなどを起こしている場合、誤った情報に基づいてその後の文脈処理を行なうため、主語等の補完が正しく行われない。

これらの問題は、市況速報記事に特化された問題点ではなく、テンプレート翻訳方式で文脈を扱えないこと、及び、複数の翻訳方式の利点を生かしていないことが原因である。

3. スクリプト依存型ハイブリッド翻訳処理

テンプレート翻訳処理において文脈処理を可能にするためには、テンプレートとマッチングを行なう範囲を、記事などの大きな単位に拡張しなければならない。

しかしながら、記事全体に適合するようなテンプレートルールを作成してしまうと、その適用範囲が非常に狭いものとなり、現実的ではない。そこで、テンプレートのマッチング範囲は文の単位のままとし、ルールを適用すべき順番を別途スクリプトによって記述する手法を提案する。この手法においては、従来２つの翻訳システムを並走させて各々の結果を得てから翻訳結果を選択していたのに対し、事前にどちらの方式で翻訳手法を行なうかを決定する。それにより、２つの方式をより有効に活用することができる。

3.1. テンプレート翻訳処理

従来のテンプレート翻訳において、図２に示すようにテンプレートルールはルール ID 及びマッチングすべき日本語表現と生成すべき英語表現のペアから構成されている。本手法では基本的には従来と同様のルール形式を取るが、意味解析型翻訳処理との連携を強化するため、いくつかの拡張を行なった。

```
(A370009
(("/売買/商い/") ("/の/が/") ("成立し") ("た")
(1 * "1610") ("銘柄")("の")("うち")("、")
("/値上がり/値上がり銘柄/"){("/が/は/")}(2 * "1610") ("、")
("/値下がり/値下がり銘柄/"){("/が/は/")}(3 * "1610")("、")
("変わら") ("ず") {("/が/は/")}(4 * "1610"){("で")} ("、")
("比較でき") ("ず"){("/が/は/")}  (5 * "1610") {("だっ")("た")}("。"))

("Among the "1" issues changing hands, "2" rose,  "3" fell,
"4" remained unchanged
  and comparisons unavailable for "5"."))

[入力]売買の成立した３４銘柄のうち、値上がり１１、
        値下がり１０、変わらず４、比較できずは９だった。
[出力]Among the 34 issues changing hands, 11 rose,
        10 fell, 4 remained unchanged
          and comparisons unavailable for 9."
```

図２　従来のテンプレートルール例及び翻訳例

3.1.1. 日本文書き換え処理

基本的なルールの構成を

(ID, 日本語パターン, 解析用和文, 出力英文)

のように4要素からなるものとする。ここで日本語パターンは入力された和文とマッチングを行なうパターン、解析用和文は入力された和文を書き換えて意味解析型翻訳で解析させるための文である。たとえば、入力の日本語内には存在するが、英語に翻訳する必要のない文章などは、(ID,日本語パターン,NIL,NIL) と記述することにより処理をスキップすることが可能となる。また、定型的なパターンで翻訳も決まっているが、表現が複雑な文などに対して、解析用和文には簡略化した文を記述しておけば、意味解析型翻訳における解析ミスの可能性を減らし、精度を向上することが可能である。

3.1.2. 省略要素補完処理

従来のテンプレート翻訳においては、意味解析型翻訳との連携が取れなかったため、英語に翻訳する際に、前の文章から主語を補完する必要のある文は対象外としていた。これらの文章に対してもテンプレート翻訳を適用できるよう、意味解析型翻訳処理の解析によって求められた省略要素を取得できるようにテンプレートルールの拡張を行なった。これにより、日本語パターン中の動詞を指定して、その動詞に対する補完された主語の翻訳を取得することが可能となった。

3.1.3. ルール優先度及び特別ルールの設定

同じ日本語表現に適合する複数のルールを使用することができるようルール間に優先度を設定した。また、後述するスクリプトルールにおいて、直接指定された時にのみ適用可能なルール群を設け、特定条件下でのみ適用することが可能なようにした。

3.2. 意味解析型翻訳サーバー

テンプレート翻訳の機能拡張に伴い、並走させていた意味解析型翻訳をサーバー化し、機能の拡張、及び、インターフェースの整備を行なった。意味解析型翻訳サーバーは図3のように構成される。

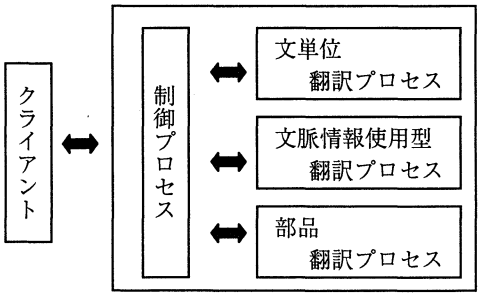

図3　意味解析型翻訳サーバーの構成

クライアントからの要求に対しては、すべて制御プロセスが受け付け、要求に応じて適切なプロセスに振り分けて翻訳及び辞書等の環境設定を行なう。以下、各プロセスの機能について述べる。

(1) 文単位翻訳プロセス

文脈情報を使用せずに一文単位の翻訳処理を行なうプロセスである。スクリプト翻訳を行なう前の前処理プログラムにおいて切り出された括弧内の文などの翻訳を行なう。

(2) 文脈情報使用型翻訳プロセス

文脈情報を使用して翻訳を行なう[中岩93]プロセス。通常の意味解析型翻訳処理はこのプロセスが行なう。また、主語推定の要求があった場合、現在貯えている文脈情報から与えられた動詞に対する主語を翻訳して、結果を返す。

(3) 部品翻訳プロセス

名詞句や複合名詞を翻訳して返すプロセスである。使用している意味解析型翻訳においては、通常時には文が与えられたと解釈し、だ文の推定などを行なって文の形式で翻訳を行なっているためその機能を使用しないようにしたプロセスである。

3.3. スクリプトによる翻訳処理

3.3.1. スクリプトルールの記述方法

スクリプトルールの記述子は以下の3つである。

ルールID	適用すべきルールID
ＡＬＴ	意味解析型翻訳(ALT-J/E)を指定
DC (Don't Care)	翻訳方式の指定を行なわない、ただし特殊ルール群に属するテンプレートルールは使用不可

これらの記述子は0回以上の繰り返しの指定を行なうことができる。また、ルールIDに関しては、複数のIDをorで結んで使用することができる。
たとえば、ID1というテンプレートルールが第1文にしか使用できないものである場合、ID1を特殊ルールに属するように設定し、スクリプトルールとして

（ID1　DC*）

と記述することにより第1文にだけ適用することが可能になる。

スクリプトルールを当初から完全に用意することは困難であるが、DCの記述子を使用して、スクリプトルール　（DC*）を用意しておけば、文のレベルで書かれたテンプレートルールが順に適用されるため、並走型のシステムと同等の性能は確保される。翻訳を行なっていくうちに判明した条件をスク

リプトルールに加えていくことで徐々にテンプレート翻訳処理の性能を向上していくことができる。

3.3.2. スクリプトルールの適用

スクリプト翻訳を行なう際の全体の流れ、及びルールの適用方法は以下のとおりである。

(1) 入力された日本文記事に対して形態素解析を行なう。

(2) 各文の形態素解析結果とテンプレートルールのマッチングを行ない、ルール適用候補のリストを作成する。

文の数をN文とするとルール候補リストは、次の構造を持つ。

((第1文とマッチするルールIDリスト)
(第2文とマッチするルールIDリスト)
:
(第N文とマッチするルールIDリスト))

(3) ルール適用候補リストとスクリプトルールのマッチングを行ない、スクリプトルールを決定する。スクリプトルールのマッチングにおいては、それぞれのX番目の要素を比較し以下のように得点を与え、最高点をとったスクリプトを採用する。

a)スクリプトIDがルールIDと一致　3点
b)スクリプトがALT　IDリストが空　2点
c)スクリプトがDC　IDリストがある　1点
d)スクリプトがDC　IDリストが空　0点

これ以外の組み合わせがあった場合、当該スクリプトは採用されない。

スクリプトルールが決定すると、同時に使用すべきテンプレートルールが決まる。複数のテンプレートが残った場合は、テンプレートルールの優先度により決定する。

(4) 条件a)または条件c)によってスクリプトとルールIDリストがマッチした文はテンプレート翻訳によって翻訳し、それ以外は意味解析型翻訳によって翻訳を行なう。

(5) 文のすべてに対してテンプレートルールが決定し、かつ、テンプレートルール内で主語の補完を要求していない場合、すべての文をテンプレート翻訳処理で翻訳する。

(6) (5)以外の場合、意味解析型翻訳を必要とする文までを書き換え規則に従って構成し、意味解析型翻訳で翻訳を行なう。ただし、テンプレートで翻訳することが決定している文は、テンプレート翻訳の結果を採用する。

4. 具体例

「X社が堅調」の訳出を、直前に「Y社が買われた」という文があった時のみ変更する。

テンプレートルール
(1　「X社が堅調」 NIL "X held firm.")
(2　「Y社が買われた」NIL "Y was bought.")
特別テンプレートルール
(3　「X社が堅調」 NIL "X was gained.")
とし、スクリプトルールとして
(2 3) (DC*)
の2つを用意する。

「Y社が売られた」「X社が堅調」といった順番で入力があった場合、ルール候補適用リストは
(() (1 3)) となる。最初のスクリプトルールは第1文がID指定であり、候補リストの対応するIDリストは空であるため採用されない。2番目の(DC*)のスクリプトは採用されるが、特別ルールであるルール3はDCとマッチしないため取り除かれ、最終的なリストは(( )(1))となり、「X社が堅調」はルール1を使って "X held firm." と翻訳される。

「Y社が買われた」「X社が堅調」の順番で文が入力されれば、スクリプトルール(2 3)の得点が6、(DC*) の得点が2となり、ルール2、ルール3が適用され、"X was gained." と翻訳される。

5. おわりに

本稿では、テンプレート翻訳と意味理解型翻訳を併用した際の問題点を述べ、それに対する解決手法を提案した。今後は、用例翻訳を含めた形でのハイブリッド翻訳方式、及びより効果的な融合方法についての検討を進めていく。

参考文献

[相沢 96]相沢, 加藤, 鎌田:外電経済ニュースの英日機械翻訳, 情報処理学会論文誌, Vol.37, No.6, pp.1041-1048
[白井 97a]白井, 細野 ,野沢 ,木村 ,阿部 ,内野: 市況速報記事に対するテンプレート型日英翻訳の効果, 情報処理学会第55回全国大会, 5J-6, Vol.2
[池原 91]S.Ikehara, S.Shirai, A.Yokoo & H.Nakaiwa: Toward an MT system without pre-editing –Effects of new method in ALT-J/E --, In Proc. of *MT SUMMIT '91*, pp. 101-106
[白井 97b]白井, 松島, 井上, 松尾, 矢部, 内野: 市況速報記事を対象とした日英翻訳システムの構成, 情報処理学会第55回全国大会, 5J-5, Vol.2
[高橋 97]高橋, 白井, 大山, 渡邊, 上田: 日英新聞記事の記事対応コーパス自動作成, 言語処理学会第3回年次大会, D1-4, pp.127-130
[中岩 93]中岩, 池原:日英翻訳システムにおける用言意味属性を用いたゼロ代名詞照応解析, 情報処理学会論文誌, Vol. 34,No. 8, pp.1705-1715