5Q-01

言語解析を用いない表現パターンベースの翻訳システム

西雅大† 石田雅子† 田辺利文‡ 乙武北斗‡ 吉村賢治‡
福岡大学大学院† 福岡大学工学部‡

1. はじめに

近年の機械翻訳では機械学習を用いた手法が主流だが,こうした手法では固有名詞を多く含む文は正しく翻訳できないことが多く,また学習データとして膨大な対訳コーパスが必要になる.

カタログの商品紹介文では固有名詞が頻出すること,ならびに固有名詞などの一部の単語を除けばほとんど同じ文体になっていることが特徴的である.こうした言い回しや文体を表現パターンとして取り出し,固有名詞など別の単語に置き換えることができる部分を変数とした表現パターンを翻訳に用いることができれば都合が良い.そこで本稿では,形態素解析や構文解析を必要とせず,表現パターンをテンプレートとして用いて翻訳する手法を示す.

2. テンプレート

今回提案する手法において,テンプレートは辞書のような役割を持つ.言語学の知識を持たないユーザでもテンプレートを作成・追加できるように,テンプレートの仕様は単純なものとなっている.

2.1. 形式

テンプレートは「範疇:和文パターン=英文パターン」の形式で記述する.テンプレートの記述例を以下に示す.

N:葛 = kudzu
N:〈N〉の風味 = the flavor of E(N)
s:〈N〉を味わいください = Please taste E(N)
S:〈s〉。= E(s).

範疇は品詞などの分類よりも大まかなものとなっている.代表的な範疇を表1に示す.

和文パターンの項目には,実際にテンプレートとして使用する日本語文パターンを記述する.パターン中には変数が含まれる場合がある(図1における“〈N〉”や“〈s〉”が変数である).変数にはそこで指定された範疇のパターンを代入することができる.また,一つのテンプレートに同じ範疇の変数が複数含まれる場合は,和文パターン中で出現する順に,“〈N1〉”,“〈N2〉”というように序数を伴う.

表1.代表的な範疇

S	文
s	句点が必要な文
V	句点を付与できない形の動詞句
N	名詞や名詞句
NUM	数字

英文パターンの項目では,和文パターンに対する英文パターンを記述する.変数も和文パターンと対応関係にあり,和文パターンの“〈N1〉”には英文パターン中の“E(N1)”が対応する.“E()”は引数の英訳を返す関数であり,“E(N1)”は変数“N1”のパターンを英訳したものを示す.

2.2. テンプレートの適用可能の定義

テンプレートTと入力文Sについて,Tの変数にSの部分文字列を代入して作られる文字列S_1がSの部分文字列となるとき($|S| \geqq |S_1|$),TはSに適用可能であると表現する.

3. アルゴリズム

今回提案する翻訳システムで行われる処理は主に二つの段階に分けられる.一つ目は入力文に適用可能なテンプレートを抽出する処理,二つ目はそれらのテンプレートを組み合わせて訳文を構築する処理である.

ここでは変数を含むテンプレートを活性テンプレート,含まないものを不活性テンプレートと呼ぶことにする.

3.1. テンプレートの抽出

不活性テンプレートの場合,入力文に適用可能かどうかは単純な文字列の比較により判別できる.

活性テンプレートの場合は変数が含まれるため,単純な文字列の比較ができないが,今回は入力文と変数以外の部分が一致するテンプレートのみを抽出する.

テンプレートが入力文に対し適用可能である場合,リストに追加するにあたって,和文パターンに対して項目(t, i, j)のリストを作成する.tは不活性テンプレートであれば和文パターン全文であり,活性テンプレートであれば変数や文字列に該当する.i,jはそれぞれ入力文中に出現するtの開始位置,終了

Expression Pattern-based Translation System using no Language Analysis
†Masahiro Nishi, Masako Ishida †Grad. Sch. of Electronics & Information Eng, Fukuoka Univ. ‡Toshifumi Tanabe, Hokuto Ototake, Kenji Yoshimura ‡Dept. of Electronics & Information Eng, Fukuoka Univ.

位置である.例えば「贈り物に最適です。」という入力文が与えられた場合,以下の二つのテンプレート(1)「N：贈り物 = a gift」と(2)「s：〈N〉に最適です = it is perfect for E(N)」があるとき,項目(1)'[(贈り物, 0, 3)]と(2)'[(N1, -1, 3), (に最適です, 3, 7)]ができる.

抽出されたテンプレートが活性テンプレートであれば open,不活性テンプレートであれば close というリストに追加する.上記の例であれば,(1)は close,(2)は open に追加する.

ここで(2)'について,変数N1の開始位置が-1となっているのは,実際に変数に代入されるべき名詞が入力文のどの地点から始まるか特定できないためである.

3.2. 組み合わせ

図1は組み合わせ処理のフローチャートである.

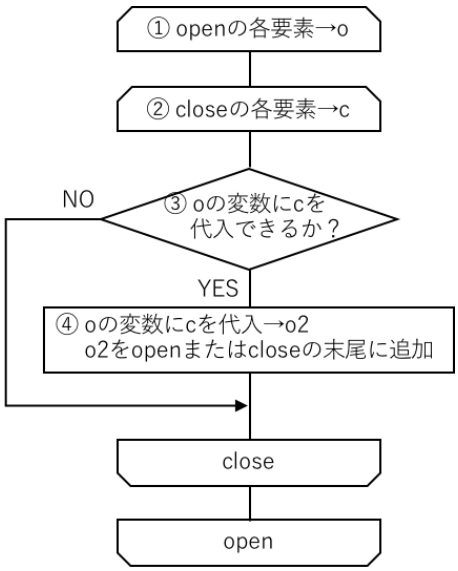

図1.組み合わせ処理のフローチャート

① open から取り出した要素を o とする.
② close から取り出した要素を c とする.
③ o に含まれるいずれかの変数に c を代入できるか調べる.
④ ③で c を代入できる場合,そこに実際に c を代入した要素 o2 を新たに得る.o2 にまだ変数が含まれている場合 open に,含まれていない場合 close に追加する.

open と close の全ての要素を組み合わせて新しい要素が得られなくなるまで,①～④を繰り返す.

この組み合わせ処理によって,和文パターンを適用する中で英文パターンを合成し,英訳を構築していく.

4. 例外処理

特定の言語現象を扱うため, テンプレートではなくプログラムで対処している.

4.1. 名詞の並列構造

入力文が3個以上の名詞からなる並列構造を含む場合,テンプレートを用いて翻訳処理を行うと,構造的な組み合わせ方が膨大な数になり,処理に時間がかかってしまう.そのため,名詞の並列構造に関しては並列構造を1つの名詞にまとめるプログラムで対処する.

4.2. 数字

数字を含む文はアラビア数字と漢数字などの表記揺れがあったり,特殊な単位が用いられることが多い.それら全てのテンプレートを用意するのは効率が悪いため,プログラムで対処する.

4.3. 冠詞

名詞のテンプレートには基本的に冠詞が付与されているが,形容詞が修飾する名詞句のテンプレートを名詞に適用する場合,通常の処理で文を結合すると,冠詞の前に形容詞が付いてしまう.これを回避するために,修飾する名詞に冠詞が付与されていた場合,冠詞と名詞の間に形容詞を挟むという例外処理を行うことをテンプレートに関数表現を用いて記述している.

5. まとめ

本稿では,一般のユーザが作成・追加しやすいように設計したテンプレートの仕様と,テンプレートの組み合わせによって英訳文を作る翻訳システムのアルゴリズムを提案した.

システムの翻訳精度を測るには,現時点ではまだ作成済みのテンプレートの数が足りない.まずは翻訳可能な文に対する翻訳結果の候補数で性能を評価する予定である.

現状の課題について,以下のようなものが挙げられる.

・名詞句の構造について,誤った構造を排除する処理の導入.
・入力文に多少の表記揺れがあっても適切なテンプレートを抽出できるような類似検索の導入.
・翻訳結果が複数ある場合の尤度計算の導入.

謝辞

本研究の一部は（株）AliveCast からの受託研究費「パーソナライゼーションシステム開発」によるものである.