5Q-02

表現パターンベースの翻訳システムにおける未知語処理

石田雅子† 西雅大† 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

駅や空港に設置してあるデジタルサイネージを利用してお土産などを販売するネットショッピングシステムでは,多様なユーザに対応すべく多言語に対応する必要がある．特にモール型ネットショップでは,商品の出品企業が商品説明文を用意する必要があり，翻訳文も用意しなければならない．商品説明文には商品名などの固有名詞が多く出現し,翻訳ソフトや web 上の翻訳サービスなどの自動翻訳ではそれらが原因で誤った解析が行われ，期待する翻訳結果が出力されないケースが多い．一方で，商品説明文は定型的な表現が多用されるという特徴があることから，本稿ではパターン翻訳を用いた日英翻訳支援システムの開発を試みた．パターン翻訳は，入力文が表現パターンに適合した場合に高い精度の翻訳結果を得ることができる．本システムでは一般のユーザが利用することを想定し，表現テンプレートの作成や追加を簡単に行えるテンプレート構造を採用した.

本稿では，本システムにおいて未知語として多く出現すると予想される商品名等の固有名詞を，文パターンや文章情報を利用することで，形態素解析や構文解析を行わずに検出する手法について報告する.

2 商品説明文の特徴

ここでは通販サイト「よかもん市場」[1] の 7,845 商品の商品説明文 28,554 文を対象にして調査した商品説明文の特徴について述べる.

・体言止めが多い
　おいしさを満喫できる贅沢な[商品名].
　[商品名]は(普段使い/特別な日)に最適な商品.
・名詞並列が多い
　卵,乳,小麦が…/和食,洋食,おかず等に…
・オノマトペを含む修飾句が多い
　チュッと絞って/もっちりとした肌触り
・固有名詞が多い
　商品名,地名,会社名
・定型表現が多い
　発送を持って返させていただきます.
　予めご了承ください.
　…にオススメです.

この中で,「オノマトペを含む修飾句」や「固有名詞」は web 上の翻訳サービスなど一般的な機械翻訳では誤った解析が行われ，翻訳精度を下げる要因となるが，パターン翻訳では入力文と文パターンの対応を取るため未知語として導出することができる.

また定型的な表現が多いことから,表現テンプレートを用いた表現パターンベースの翻訳が適応できると考えられる.

3 表現パターンベースの翻訳システム

表現パターンベースの翻訳において高い精度の翻訳結果を得るためには,大量のテンプレートを必要とする.大量のテンプレートを適切に作成するためには,日英対訳文などの大規模なコーパスだけでなく，テンプレートそのものが複雑な構造を持つため,言語学の知識を持った専門家を要する[2].本研究では，システムに予め基本的な表現テンプレートを準備しておき，必要に応じて一般ユーザが表現テンプレートを作成・追加できるように簡単な構造を持つテンプレートの仕様を定義した.

3.1 テンプレート

本システムでは,テンプレートを範疇記号,和文パターン,英文パターンの 3 つの項目で構成されたものと定義した.範疇とテンプレートの一例を表 1,2 に示す.

表 1.範疇の一例

範疇	意味
S	“。”を付けて文になる表現
V	“。”を付けられない表現
N	名詞や名詞句

表 2.テンプレート記述例

範疇	和文パターン	英文パターン
s	一般的には<s>	in general, E(s)
s	とにかく<s>	anyway E(s)
N	葛	Kudzu
N	<N>の風味	the flavor of E(N)

また,テンプレートは変数を含むもの（活性型）と，含まないもの（不活性型）に分類する．不活性テンプレートは一般的な機械翻訳の単語辞書に相当する.

テンプレートは人手で作成し,それらを集めてテンプレート辞書とした.

Processing of Unknown Words in Expression Pattern-based Translation System

†Masako ISHIDA, Masahiro Nishi †Grad. Sch. of Electronics & Information Eng, Fukuoka Univ. ‡Toshifumi TANABE, Hokuto OTOTAKE, Kenji YOSHIMURA ‡Dept. of Electronics & Information Eng, Fukuoka Univ.

3.2 テンプレートの適用可能と被覆の定義

テンプレートTと入力文Sに対して，テンプレートTに含まれる変数に入力文Sの部分列S1を代入してできる文字列S2が入力文Sの部分列になるとき，テンプレートTは入力文Sに適用可能という．

また，テンプレートTに含まれる変数を文字列におきかえることをテンプレートに適用すると呼ぶ．

テンプレートTが入力文Sに適用可能なとき，テンプレートTの文字列と一致する入力文S中の文字列を，テンプレートTによって被覆される文字列と呼ぶ．

3.3 翻訳文生成の手順

翻訳文生成の手順を以下に示す．

手順1 テンプレートの抽出

入力文に対して適用可能なテンプレートを，テンプレート辞書から抽出する．適用可能であるテンプレートは，入力文とテンプレートに含まれる文字列との編集距離に基づいて判断する．

適用可能なテンプレートを抽出する際，次に示す構造をもつ項目を作成する．

項目：[範疇，(和文パターン，i，j)，英文パターン]

ここでiとjは，入力文に対してテンプレートの和文パターンが適用可能な文字列の開始位置(i)と終了位置(j)を示している．開始位置iが0の場合は入力文頭を意味する．

作成された項目は，和文パターンが活性である場合はopenのリストに，不活性である場合はcloseのリストに格納する．

手順2 テンプレートの合成

翻訳文は，テンプレートの英文パターンを組み合わせることで生成する．openとcloseから項目をそれぞれ取り出し，closeの項目がopenの項目に適用可能である場合，項目内の和文パターンを適用しつつ，同時に英文パターンの合成を行う．適用された和文パターンを元に新たな項目を生成し，パターンが活性である場合はopenに，不活性である場合にはcloseに追加する．

手順2を適用可能な項目が存在する限り繰り返した結果，close中に項目：(不活性な和文パターン，0，n)が存在した場合，その項目の英文パターンが翻訳結果となる．ここでnは入力文の文字長である．この項目がcloseに存在しない場合，テンプレートの組み合わせでは翻訳が完了しないことを意味する．その場合は，入力文中の文字をできるだけ多く被覆する活性テンプレートをopenから求めて，その変数部分に未知語があると推定する．

4 未知語推定処理

本システムにおける未知語の出現は，名詞をはじめとする不活性テンプレートの不足と，合成に使用する活性テンプレートの不足が要因として考えられる．ユーザがシステムを使用・辞書の拡充を行うことで活性テンプレートは充実するため，未知語として頻出するのは名詞をはじめとする不活性テンプレートであると予想される．

そこで今回は，文や述語に対する十分な数のテンプレートが予め作成されていると仮定し，不活性テンプレートの不足による未知語の推定に注目する．

入力文に未知語が存在することは，次の場合に検出できる．

（ケース1）手順1が終了した時点で，すべての適用可能なテンプレートを組み合わせても被覆されない文字列が入力文中に存在する場合．

（ケース2）手順2が終了した時点で，closeに項目：(不活性な和文パターン，0，n)が存在しない場合．

名詞をはじめとする不活性テンプレートは，ケース1で検出可能であると予想される．そこで，手順1後の未知語推定を以下の手順で行う．

1)適用可能なテンプレート群から，被覆されていない文字列を導出する．

2)1)の文字列情報や，既存のテンプレート，記号や文字種の変化の情報から未知語になりうる部分とカテゴリの推定を行う．

この手順によって，未知語の自動推定が可能であるかを検証する．

5 おわりに

本稿では，一般ユーザの使用を想定した簡単な構造を持つテンプレートを用いた表現パターンベースの翻訳システムの概要と，システムにおける未知語推定の手法について述べた．

今後商品説明文を用いた実験によって，本稿の手法によって不活性テンプレートの不足によって出現する未知語の推定精度の検証を行う予定である．

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．

参考文献

[1] よかもん市場 (https://www.yokamon.jp/)

[2] 池原悟: 非線形な表現構造に着目した重文と複文の日英文型パターン化，自然言語処理，Vol.11，No.3，pp.69-95，2004.