統計的機械翻訳における未登録語のグループ化による翻訳

吉崎　大輔 [*1]　山本　博史 [*2,3]　大熊　英男 [*3]　匂坂　芳典 [*1,3,4]
[*1] 早稲田大学 GITI，[*2] 近畿大学 理工学部，[*3] NiCT，[*4] 早稲田大学 ことばの科学研究所

1. はじめに

現在広く利用されている翻訳技術として挙げられる統計的機械翻訳において，翻訳システム外のデータ（未登録語）の翻訳は翻訳精度向上の課題の一つである．従来の未登録語の翻訳手法としては，未登録語について翻訳システム外から求めた訳語を用いて単語単位で翻訳するという方法が取られてきたが，未登録語とその前後の単語との位置関係に考慮したフレーズ単位での翻訳は困難であった．

上記の問題について，先行研究では，意味的な分類 (以下下位分類) が同じ単語は文中での出現位置が類似することに着目して，入力文中の未登録語を対訳コーパス中で同じ下位分類に属する頻出単語に置き換えるという手法が提案され，未登録語の訳語の位置関係に関する翻訳確率値を用いた翻訳が可能となった [3]．しかし，頻出単語に関する確率値はコーパス中での出現回数に依存するため，使用するコーパスの量と内容によっては頻出単語が正しい翻訳文を導出するのに十分な統計量に達しない恐れがある．

以上の問題点を踏まえ，本研究では，対訳コーパス中において未登録語と同じ下位分類に属する単語を全て一つのグループとして抽象化（グループ化）することで統計量の確保を図り，それによる未登録語を含む文の翻訳手法を提案した．また，提案した手法による翻訳実験から，提案手法による翻訳の妥当性について検証を行った．

2. 学習コーパスの不足による未登録語の出現と翻訳精度の低下の問題

統計的機械翻訳での訳文導出には言語モデル，翻訳モデルの２種類の確率モデルが必要であり，言語モデルは翻訳先言語のコーパスから、翻訳モデルは対訳コーパスから生成される．そのため，コーパスの量の増加に伴い翻訳システムにおける翻訳可能な単語の数及び翻訳結果の正解率は上昇するが，現実問題として与えられるコーパスの量にも限界があり，コーパスの不足により翻訳モデルに学習されない単語（未登録語）が必ず出てくる．入力文に未登録語が含まれている場合，その未登録語が翻訳されないだけでなく，訳文全体に対して誤った翻訳確率を与えてしまい，翻訳精度の低い訳文が生成されてしまう．

統計的機械翻訳システムでは，入力文中の未登録語は翻訳されないままの形で訳文中に挿入されることが多い．そのため，従来の未登録語の翻訳手法としては，未登録語の訳語を対訳辞書等の翻訳システム外のデータから別途求め，それを訳文中の未登録語と変換するという方法が取られてきた．しかし，単に未登録語の訳語のみを無理矢理置き換えただけでは，その前後との位置関係については考慮できていないので，全体から見ると誤った訳文となる可能性がある．

3. 単語の下位分類を利用した未登録語の翻訳

従来の手法での問題点については，未登録語と類似した位置関係を持つ単語を代用して翻訳処理を行えば，信頼性のある確率値による位置関係に考慮した未登録語の翻訳が可能であると推測される．未登録語と類似した位置関係を持つためには，未登録語と同じ品詞で，かつ語彙の意味が同じカテゴリーに分類される（同じ下位分類情報を持っている）ことが条件として挙げられる．

以下の節では，単語の下位分類を利用した未登録語の翻訳手法として，先行研究で提案された手法について解説する．

3.1 未登録語と同じ下位分類の頻出単語を代用した翻訳手法

先行研究では，信頼性の高い翻訳確率値による訳文導出を目的として，未登録語と同じ下位分類で，かつコーパス中での出現頻度が高い単語を代用する手法が提案された [3]．

上記の翻訳手法の処理手順について例文を用いて図（1）に示す．まず入力文中に含まれる未登録語「パキスタン」を抽出し，その下位分類の情報を翻訳システム外の語彙データから求める．「パキスタン」の下位分類が「国名」であるということが分かったら，同じ下位分類である頻出単語「アメ

リカ」を入力文中の未登録語「パキスタン」と置き換えて翻訳システムに与える．この時生成された訳文には，未登録語と置き換えた単語「アメリカ」の訳語「America」が含まれている．最後に，翻訳システム外の対訳辞書データから未登録語「パキスタン」の訳語「Pakistan」を導出し，訳文中の「America」と再度置き換える．以上の処理を行うことで，未登録語を含む文の翻訳が可能となる．

しかし，上記の方法では，使用するコーパスの量や内容の違いによっては頻出単語について十分な統計量が得られず，信頼性の高い確率値が算出されないために，未登録語を含む文を正しく翻訳できない恐れがある．

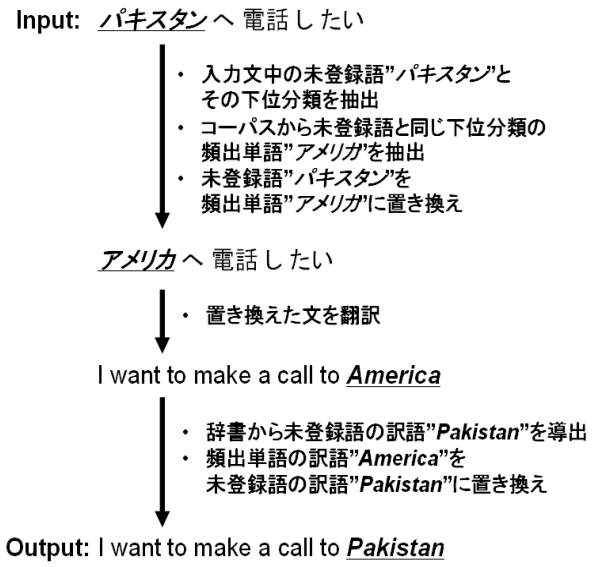

図 1　未登録語と同じ下位分類に属する頻出単語への代用による翻訳処理の例

4. 未登録語と同じ下位分類の単語のグループ化を用いた翻訳手法による統計量の獲得

先行手法での問題点は，未登録語の代用に使う単語を同じ下位分類の頻出単語に限定したことにより生じた．そこで本研究では，未登録語と同じ下位分類の単語を全て一つのグループとして抽象化（グループ化）を行い，未登録語をそのグループで代用する翻訳手法を提案した．単語のグループ化を行うことで，少なくとも頻出単語以上の統計量は確保でき，扱えるコンテキストのパターンも増加する可能性が期待できる．

グループへの代用による翻訳手法の具体的手順を，前章と同じ例文を使用して図（2）に示す．まず未登録語「パキスタン」とその下位分類情報を語彙データから抽出し，その下位分類に該当するグループのラベル「group A」を入力文中の未登録語と置き換えて翻訳する．この時，入力文中のラベル「group A」はそのまま「group A」と翻訳されるものとする．最後に，対訳辞書データから求めた未登録語の訳語を訳文中の「group A」に置き換えることで，未登録語を含む文の翻訳が完成する．

上記の通り基本的な処理は先行手法と同じであるが，提案手法では翻訳処理に各グループのラベルについて学習されているモデルの使用が前提とされている点で異なる．そのため，グループのラベルを含む文を翻訳可能にするには，学習コーパス中の各グループに属する単語をラベルに置き換える処理がモデル学習の前に必要となる．

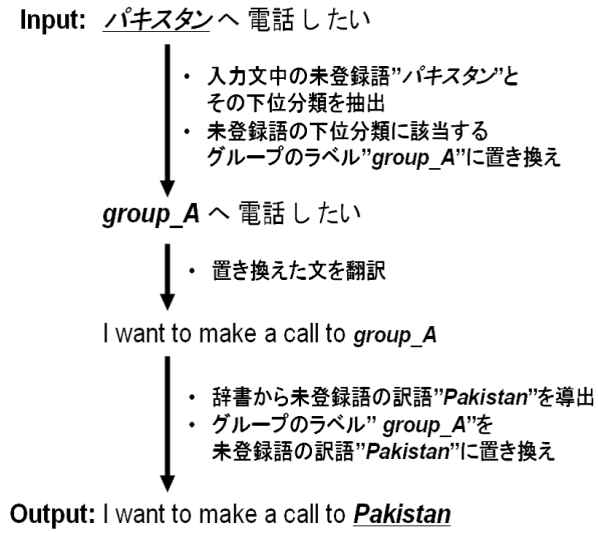

図 2　未登録語の下位分類が属するグループへの代用による翻訳処理の例

5. 提案手法による未登録語を含む文の翻訳の検証

前章にて提案した手法による未登録語を含む文の翻訳（日英，英日）の妥当性について検証する．ただし，提案手法は先行手法の改良案として提案されたため，先行手法による翻訳よりも精度が高いかを検証の判断基準とし，同じデータを用いた際の提案手法及び先行手法の翻訳精度を比較した．また，今回の比較では翻訳可能な未登録語の数の違いによる翻訳精度の差を出さないようにしたい．そこで，使用するコーパスの分野に応じて未登録語になりやすい下位分類を仮定し，その下位分類を基にコーパ

ス内の単語のグループ化及び語彙データと対訳辞書データの作成を行い，どの手法においても決まった未登録語が必ず翻訳されるように予め設定した．

また，上記の方法で提案手法を検証するためには先行研究での手法の妥当性についても確認する必要がある．そこで，上記２つの手法と共に，本稿でも取り上げた「訳文中の未登録語をその訳語に変換する」という単純な翻訳手法を他の手法と同じデータで行い，その翻訳結果との比較も行った．ただし，他２つの手法と同様に決まった未登録語が必ず翻訳されるようにするために，実際には翻訳モデル中に未登録語とその訳語の対応関係を追加するという処理を翻訳前に施し，上記の手法を擬似的に行った. [3]

以降の節では以上３種類の手法の翻訳精度の比較検証について解説するが，便宜上，訳文中の未登録語をその訳語に変換することによる翻訳手法を「従来法１」，同じ下位分類に属する頻出単語の代用による未登録語の翻訳手法を「従来法２」と呼ぶことにする．

5.1 翻訳に使用する各種データの用意

各モデルの学習に使用する対訳コーパスには，NiCT/ATR の旅行会話文データベース「BTEC」（日本語文，英文共に約 46 万文）を使用した．両言語のコーパス中の各単語の下位分類は，英語は ATR 独自のタガーで，日本語は形態素解析ソフト「茶筅」で定義されているため，コーパス内の単語のグループ及び語彙データと対訳辞書データの作成は，それらの下位分類を基に行った [4]．今回使用するコーパスは旅行会話を扱っているため，種類の多さと増加しやすさの点から,「場所」に関する下位分類を持つ固有名詞が未登録語になりやすい単語であると仮定した．ただし,「場所」という分類だけでは，文中での各固有名詞の出現位置にばらつきがあるため，信頼性のある確率値が得られない恐れがある．そこで，今回は文中での位置関係について主観的評価を行い,「場所」という下位分類を更に「宿泊施設」,「国名」,「地名」,「組織名」,「飲食店」の５種類に細分化し，それらを基にコーパス中の単語のグループ化等を行った．各グループとそれに属する固有名詞の種類については図（3）に示す.

テストセット文及びその正解文には，学習コーパスと同じく NiCT/ATR の旅行会話文 447 文を使用した．これらの文は全て必ず１つは図（3）で示した５種類のグループのいずれかに属する単語が含まれている．また，今回の比較の前提条件として，決まった未登録語が必ず翻訳されるようにしなければならない．そこで，テストセットの各文から前節で挙げたグループのいずれかに属する単語を抽出し，その単語を語彙データ及び対訳辞書データに含まれる単語に変換した．実際の翻訳処理には変換後のテストセットを使用するため，このテストセット文による各手法での翻訳はクローズテストとして扱う.

各モデルの学習については，言語モデルの生成ツール SRILM (注1)，及び翻訳モデルの生成ツール GIZA++(注2)を使用し，言語モデルについては 3-gram までの生起確率について学習を行う [5]．また，デコーダには ATR で作成された Cleopatra と呼ばれる，既存のデコーダ Pharaoh(注3)と互換性を持つデコーダを使用した [3].

グループ	ラベル	各グループに属する固有名詞の種類
宿泊施設	ACM	宿泊施設
国名	CNT	国名
地名	PLC	地名, 観光名所, 地形
組織名	ORG	会社名, 組織名
飲食店	RST	レストラン名, 店名

図 3　主観評価により作成したグループの一覧

5.2 評価指標 BLEU による翻訳精度の客観評価

各翻訳手法により生成された翻訳結果を客観的に評価するため，BLEU という客観評価指標を使用した．BLEU は，翻訳文と正解文における単語と語順の一致率から評価し，評価値が高ければその訳文の翻訳精度は高いと評価できる [2], [6]．その評価値は式（2）より算出される.

$$BLEU = BP \times exp(\sum \frac{1}{N} P_n) \qquad (1)$$

$$P_n = \frac{\sum_i \text{翻訳文 } i \text{ と正解文 } i \text{ で一致した全 } n-gram \text{ 数}}{\sum_i \text{翻訳文 } i \text{ 中の全 } n-gram \text{ 数}}$$

また，式（2）中の BP は，訳文の長さが正解文

(注1)：http://www-speech.sri.com/projects/srilm/
(注2)：http://fjoch.com/GIZA++.html
(注3)：http://www.isi.edu/licensed-sw/pharaoh/

より短い場合に BLEU の評価値が不当に高い値になるのを防ぐためのペナルティである．

5.3 翻訳精度の比較から見た提案手法の妥当性の検証

評価指標 BLEU を用いた各手法の翻訳精度を比較結果を図（4）に示す．表中の各数値は，日英翻訳及び英日翻訳における各手法での BLEU のスコアであり，前節の通りこの値が大きければそれだけ翻訳精度が高いことを示している．

従来法１での翻訳精度が他の手法に比べて低いのは，前述の通り未登録語に対して訳語を与える処理しか行っていないため，訳文中での出現位置までは制御できなかったことが原因と考えられる．このことに加え，従来法２と従来法１での評価値の差から従来法２の優位性が示され，従来法２による翻訳の妥当性が確認できた．

また，提案手法と従来法２との間での評価値の差は比較的大きく，日英翻訳で 5.73，英日翻訳では 4.5 もの差があった．このことから，提案手法の優位性がはっきりと示された．以上の結果より，提案手法による未登録語を含む文の翻訳が妥当であることが示された．

(BLEU score)

	提案手法	従来法2（下位分類が同じ頻出単語の代用）	従来法1（訳文中の未登録語の訳語への変換）
日英翻訳	41.03	35.31	30.53
英日翻訳	39.74	35.24	31.26

図 4　BLEU での評価による各手法の翻訳精度の比較結果

6. ま と め

統計的機械翻訳での未登録語の問題について，文中の未登録語をその下位分類に応じたグループで代用する翻訳手法を提案し，信頼性の高い翻訳確率値による未登録語を含む文の翻訳を図った．提案した手法について，先行研究での頻出単語の代用による翻訳手法，及び「訳文中の未登録語をその訳語に変換する」という従来の単純な翻訳手法との翻訳精度の比較をした結果，提案手法による翻訳が最も精度が高く，提案方法による未登録語の翻訳の妥当性が確認できた．今後は提案手法による翻訳の精度向上を目的として，客観的評価によるグループ化処理の自動化が課題となる．

謝辞　本研究は独立行政法人情報通信研究機構で行った研究を基にまとめたものである．研究の機会を与えていただき、ご指導いただきました情報通信研究機構 隅田英一郎グループリーダーに感謝する．

文　献

[1] 北研二，“言語と計算４ 確率的言語モデル,” pp.197-198, 財団法人 東京大学出版会，Nov. 1999.

[2] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, ”BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp.311-318, July. 1993.

[3] 大熊英男，山本博史，隅田英一郎, “フレーズベース SMT への対訳辞書の導入,” 言語処理学会第 13 年次大会，pp.380-383，March. 2007.

[4] 浅原正幸，松本祐治，“「茶筅」/「南瓜」を用いた形態素解析・係り受け解析　-「茶器」によるコーパス管理・検索-,” 自然言語処理技術講習会資料, 自然言語処理技術講習会，pp.13-35，Sept. 2007.

[5] Andreas Stolcke, ”SRILM-AN EXTENSIBLE LANGUAGE MODELING TOOLKIT,” Speech Technology and Research Laboratory SRI International, Menlo Park, CA, U.S.A.

[6] 坂田浩亮，新保仁，松本裕治，“コーパスを用いた言語習得度の推定,” 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座，p.114，Sept. 2007.