自律的語彙拡充を行う機械翻訳システム*

神山 淑朗† 伊藤 晴美†
日本アイ・ビー・エム株式会社 ソフトウェア開発研究所†

1. はじめに

機械翻訳システムは，翻訳サーバーの形態あるいはクライアントソフトウェアの形態で Web ページの翻訳などの用途で広く一般に使用されるようになった．しかし，実世界での運用を始めると，日々生まれる新しい言葉，例えば新製品の名称や今話題の時事用語などへの対応が日を追うごとに不十分になってくることに気付く．放置すれば徐々に翻訳精度が低下してしまい，精度維持のためには絶え間のない辞書のメンテナンスが必要不可欠である．こうした辞書のメンテナンスは，語彙の収集・辞書の作成から翻訳システムへの組み込みまでを含め，多くの人手を要するコストの高い作業である．

近年，電子化された大量のコーパスが利用できるようになったことを背景に，コーパスベースの(半)自動的語彙抽出の研究は多く行われている[4]．しかし，それらは主に既存のコーパスを基に「静的な辞書」の作成工程の一部の省力化に寄与するための手法である．

本稿では，動的な語彙拡充が必要とされる Web 翻訳を行う英日機械翻訳システムにおいて，辞書の語彙拡充作業を人手を介さずに機械翻訳システム自身に行わせる実験を行ったので，その手法を紹介する．

2. 自律的語彙拡充の手法

提案する手法では，機械翻訳システムへの語彙拡充を機械翻訳システム自らが行う．このとき，語彙収集を行うにあたってそのためだけの特別なジョブを走らせるのではなく，あくまでも機械翻訳システムの本来の仕事である翻訳処理の中で行う．これには二つのメリットがある．一つは翻訳処理の過程においては様々な自然言語処理の解析結果が流用できること．もう一つは語彙収集の情報源が，機械的にリンクをたどったものではなく，ユーザーが実際にアクセスして翻訳したいと思う良質なものになるという点である．ある程度データが集まったら，それを集計し，機械翻訳システム自身による翻訳結果で訳語を補えば，既存の辞書には存在しなかった新しい語彙を持つ辞書を作成し，使用することができる．以下では，本手法の具体的な手順について述べる．

2.1 フレーズの自動収集

絶えず拡充が望まれる新語の大部分は名詞句(Noun Phrase: NP)であることから，まずは Web ページの翻訳を行う際に，同時に辞書に未登録の名詞句を見つけてそれらの収集を行う．翻訳を行っているということは，内部では形態素解析の結果などを活用することができるということであり，品詞推定[1]で得られた品詞列などの情報から名詞句と思われるフレーズをわずかな追加処理コストで抽出することができる．

また，Web ページの翻訳を行うと，どの分野別辞書に含まれるパターンに何回ヒットしたかという情報から主題の分野を推定することができる[2]．その情報は収集したフレーズを分野ごとに分類するのに役立つため，抽出したフレーズをそれが出現したページの推定分野とともにデータベース(DB)に保存する．

2.2 フレーズの自動分類

一般に翻訳辞書は分野ごとに適切な訳出を行うために分野別辞書を持っている．そこで，DB に蓄えられたフレーズを，それが出現したページの推定分野の情報を利用して「政治」，「野球」などのカテゴリに分類することが考えられる．このとき，言うまでもなく頻度の高いものほど重要なフレーズと考えられる．また，高頻度であるということはさまざまな文脈から切り出されたものであるため，そのフレーズが本当に意味のある名詞句になっているという精度も期待ができる．しかしながら，多くの分野に広範囲に出現するフレーズは特定の分野に分類することはできない．ここで，フレーズを索引語，分野を文書と考えると，情報検索の分野でよく行われる索引語の重み付け(term weighting)と類似の課題であることがわかる．すなわち，高い頻度で出現するという性質と，特定の狭い範囲に分布するという性質とを合わせ持つように重みを計算すれば，辞書に登録する価値があり，かつ分野に分類可能と考えられるフレーズを得ることができる．

2.3 辞書の自動作成

辞書に登録すべきフレーズが集まり，その分類を行ったところで，それらを辞書に登録する．このとき，訳語はフレーズを翻訳エンジン自身で翻訳して生成する．そもそもその翻訳エンジンの辞書に存在しないフレーズばかりを集めたので，自身で翻訳しても理想的な訳語が得られるとは限らず，ときには質の悪い訳語を辞書に登録してしまう場合もある．しかし，そのような場合はそのフレーズを辞書に登録しなくても元々訳せないので，それ以上品質が劣化する心配は少ないと考えられる．訳語を生成する際に，翻訳エンジンを調整してフレーズを名詞句的に訳出したり，推定分野の辞書の優先度を上げるといった工夫により，全体的には向上を見込める．

2.4 辞書の自動登録

作成された辞書を自動的に翻訳エンジンへ組み込み，次回以降の翻訳に反映させる．次回からはその辞書を使用してフレーズ収集や分野推定も行うことになる．

---

*A Machine Translation System with Self-Extending Lexicon
†Yoshiroh KAMIYAMA, Harumi ITOH
Software Development Laboratory - Yamato (YSL)
IBM Japan, Ltd.

3. 評価実験

提案手法の効果を確認するためにプロトタイプを作成して実験を行った.

3.1 システム構成

本手法を適用可能な機械翻訳システムの運用形態はいくつか考えられるが, ここではクライアント側で翻訳を行い, サーバー側の DB にデータを蓄積する方式とした. 図 1 にシステム構成を示す.

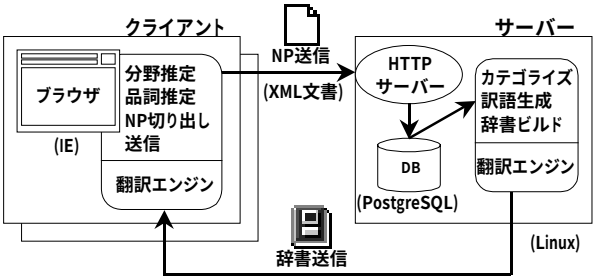

図 1 システム構成

3.2 フレーズ分類のアルゴリズム

筆者らの機械翻訳システムは, 基本辞書の他に主分野, 副分野と呼ぶ二段に階層化された分野別辞書を持っている. 例えば, 主分野辞書である「スポーツ」辞書には各種スポーツで共通の用語が収められており, 副分野辞書である「サッカー」「ゴルフ」等の辞書にはそれぞれのスポーツ固有の用語が収められている. 以下ではこのような階層化された構成の分野別辞書へ, 収集したフレーズを分類するアルゴリズムを示す.

DB に収集されたデータは, 図 2 のようにフレーズ t_i が推定分野 d_j のもとに何回出現したかを表す行列で考えることができる.

$$A = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{cccc} d_1 & d_2 & d_3 & d_4 \end{array} \\ \begin{bmatrix} 10 & 0 & 0 & 1 \\ 1 & 12 & 0 & 1 \\ 3 & 5 & 3 & 2 \end{bmatrix} \end{array}$$

図 2 フレーズ ― 分野行列

個々のフレーズは, 分野を要素としたベクトル空間(vector space)中のベクトルとみなすことができる. ここで分野ごとにその分野だけが要素 1 をもつ単位ベクトルを考えると, それらの単位ベクトルとフレーズのベクトルとの類似度(式(1))で, そのフレーズがどの分野にどの程度特徴的に出現したかの尺度を得られる. フレーズの出現頻度は式(2)により最大頻度で正規化を行い, 両者の積で重み付けを行う.

$$sim(\vec{t}_i, \vec{e}_j) = \frac{\vec{t}_i \cdot \vec{e}_j}{\left|\vec{t}_i\right|\left|\vec{e}_i\right|} \quad (1)$$

$$tf_j^i = K + (1-K)\frac{freq(i,j)}{\max_{i,j} freq(i,j)} \quad (2)$$

はじめに分野集合を全副分野として重み付けを行うと, 特定の副分野に固有に出現し, かつ頻度の高いフレーズがランクの上位に得られる. それらを取り除いた上で, 今度は分野集合を全主分野として同じ計算を行うと, 一つの副分野には偏らないものの特定の主分野に固有のフレーズが上位にランクされる. それらも取り除いた上で頻度が上位のものは基本辞書の候補とし, 残りは利用しない.

3.3 実験結果

約 7000URL の Web ページに対して本手法を適用した. 表 1 に収集されたフレーズとその訳語の一部を示す. これらのフレーズが辞書に登録されることにより, (1)名詞句の認識がより確実になるため構文を大きく取り違える可能性が減り, (2)ヒットした場合は構文解析時のあいまい性が減少するためにその文の翻訳速度が数パーセント程度であるが向上し, (3)分野別辞書の語彙が増すので分野推定の精度が向上する, という効果があることが確認された.

表 1 収集されたフレーズとその訳語(一部)

副分野	フレーズ	訳語
sports/golf	US Masters	US マスターズ
home/cooking	Honey Mustard	蜂蜜マスタード
home/travel	National Tourist Offices	国立ツーリスト・オフィス
sci/medicine	manual vacuum aspirators	手動の真空吸引器
ent/music	breakthrough song	突破歌
主分野	**フレーズ**	**訳語**
computer	enterprise software	企業ソフトウェア
computer	Toon Boom Studio	トゥーン・ブーム・スタジオ
home	Family Fun Experiences	家族楽しみ経験
politics	United Nations Mission	国際連合使命
sports	playoff hunt	プレーオフ狩り

4. まとめ

本稿では, Web ページに現れる辞書に未登録の日々生まれる新しい語彙を, 人手を介さずに機械翻訳システム自らが収集, 分類し, 辞書を作成, 登録する手法を提案した. 訳語を知らないシステムに訳語を生成させるという点で万全ではないが, 一定の効果が確認された. もちろん現実には訳語を与える部分だけを人手で行うことも視野に入れており, その場合は本手法により最小限の手間で最大限の効果を期待できる. 今後はフレーズ切り出しの精度向上や, 人名なのか製品名なのかといった素性の推定などが課題である.

参考文献

[1] 神山淑朗: "機械翻訳システムにおける確率的品詞推定とその応用", 情報処理学会第 63 回全国大会, 2001

[2] 羽鳥洋美, 神山淑朗: "分野判定トリガー情報のフィードバックによる Web 翻訳", 情報処理学会第 63 回全国大会, 2001

[3] 徳永健伸: "情報検索と言語処理", 東京大学出版会, 1999

[4] Ellen Riloff and Jessica Shepherd: "A corpus-based bootstrapping algorithm for semi-automated semantic lexicon construction", *Natural Language Engineering*, 5(2):147-156, 1999