韓国語コーパスからの外来語自動抽出と言語解析への応用

金 玉錦*1, 藤井 敦*2, 石川 徹也*3

*1 図書館情報大学 yujin@ulis.ac.jp

*2 筑波大学/科技団 CREST fujii@slis.tsukuba.ac.jp

*3 筑波大学 ishikawa@slis.tsukuba.ac.jp

1 はじめに

インターネットの急速な普及により社会の情報化が進み、科学技術用語、外国語、流行語などの新しい言葉がいち早く日常用語の一部としてとり入れられるようになった。新語の急速な出現は、日常言語の計算機処理を目的とした自然言語処理、情報検索、機械翻訳、音声認識などの研究に影響を与えている。これらの言語処理では辞書が不可欠である。しかし、既存の辞書には新語が十分に登録されていないため、未知語問題が発生する。国際交流が盛んになった今日、外国語の日常用語、専門用語、固有名詞なども音訳して使うことが増えてきた。そこで、外来語も未知語になりやすい。

新語を迅速に辞書へ登録して、未知語を極力減らす必要がある。しかし、辞書の人手による編纂には膨大な時間と労力を必要とする。そこで、コンピュータを用いて大規模なコーパスから言語処理に有用な新語を自動抽出し、辞書を更新する研究が提案されている。

韓国語における外来語辞書の自動更新においては、コーパス中の外来語を検出することが容易ではない。カタカナで外来語を表記する日本語とは異なって、韓国語では、一般語の表記に用いられる字種(ハングル)によって外来語も表記するからである。そのため、字種に基づいて外来語を検出することは難しい。例えば、「ワールドカップを見る」という意味の「월드컵을 보다」という韓国語文を例にとる。日本語の場合は外来語である「ワールドカップ」をカタカナで書くため識別しやすいのに対して、韓国語では全文をハングルで表記するので外来語「월드컵(ワールドカップ)」の識別は困難である。

本研究では日本語のカタカナ語を手がかりに韓国語コーパスから外来語用語を自動的に抽出し、日韓対訳辞書を自動的に構築する手法を提案する。さらに、抽出した外来語を形態素解析に応用して、解析精度への効果について分析する。本研究では、韓国語に音訳された西洋言語を外来語と総称する。

以下、2章で外来語辞書の自動作成に関する先行研究について検討し、3章で本研究で提案する手法を説明し、4章で評価実験、5章で結論と残された研究課題について議論する。

2 先行研究の検討

日本語のようにカタカナで外来語を表記する習慣がない韓国語では、外来語の検出手法が重要である。Jeongら[1]は外来語と韓国固有の語との音韻的差異を分析して確率モデルを作成し、外来語であるか韓国語固有語であるかの判断を行い、外来語だけを抽出する手法を提案した。しかし、外来語の音韻的規則に基づいた言語モデルによる言語認識法では外来語判定の精度が低いことが問題である。

外来語は移入元の言葉(原語)と対応関係にあることから、対訳コーパスから対訳辞書を自動的に作成することで、外来語を自動的に抽出する手法[3,9,10,11,12]がある。Lee[3]は、英韓対訳コーパスの形態素解析結果から未知語を対象に言語モデルを利用して外来語判定を行い、さらに音韻比較に基づいて固有名詞の訳語対を抽出した。しかし、対訳コーパスに必ず対訳対が存在する保証はなく、形態素解析結果から未知語だけを外来語判定の対象にするため、形態素解析段階で過分割される外来語は処理できない。また、外来語の判定に利用する言語モデルも精度に影響する。

最近は Web から対訳情報を抽出する手法[7,8]が提案されている。Web から対訳関係にある複数のページや対訳情報を掲載した単一のページを抽出して、対訳辞書を作成する。しかし、Web には低品質な情報が混在していることが多く、言語によっては内容が一致している対訳ページが少ないなどの問題がある。

さらに、外来語の翻字[2,5]や逆翻字[4,6]の観点から、外来語を抽出し、英韓対訳辞書を構築する手法がある。Lee[4]は、コーパス中の外来語候補となる単語を集め、原型(英語)に復元して、復元した原型が英語辞書に登録されているならば外来語と判定し、英韓対訳対と判断する。この手法でも、コーパスの形態素解析結果から未知語だけを対象にするため、適用範囲が狭い。また、単純に文字情報に基づく逆翻字において、翻字の過程で省略される発音の再現が難しいなどの問題がある。例えば、

韓国語の「하이」が英語の「hi」と「high」のどちらに対応するかは、文脈情報がないと判断できない。

3 本研究で提案する手法

3.1 システム構成

日本語と韓国語において、英語からの外来語が音韻的に似ていることから、カタカナ辞書と韓国語コーパスを音韻表記のローマ字表記に変換し、ローマ字化された韓国語コーパスから、カタカナ語と音韻的に類似する語を外来語として抽出し、さらに日韓対訳辞書を構築する。以下、各処理について具体的に説明する。

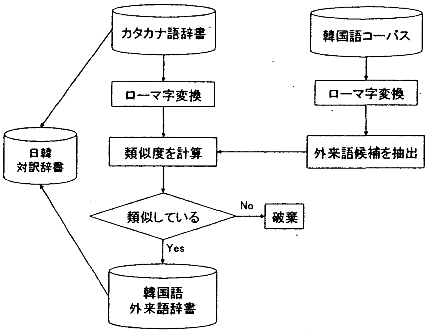

図 1 ：外来語抽出及び日韓対訳辞書作成のシステム構成図

3.1.1 カタカナ語のローマ字変換

同一原語から移入した外来語は音韻的に似ている特徴を持っている。従って、韓国語外来語は移入元である英語に音韻的に類似する。しかし、すべての英単語が外来語になるわけではないので、英語辞書を直接用いることは適切ではない。そこで、本研究では迅速且つ簡単に入手できるカタカナ辞書を用いて、カタカナに音韻的に類似する韓国語を外来語として抽出する。

カタカナ語のローマ字表記方式には、韓国語のローマ字表記に近い新訓令式(第 1 表) を用いる。新訓令式の第 2 表は固有名詞などの特別表記に対応しているので、本研究では標準表記法の第 1 表のみを考慮する。

日本語と韓国語における英語からの移入語の傾向は似ているため、カタカナ語の音節とローマ字との対応規則を 106 通り作成し、利用する。また、「チ」と「ツ」の表記は韓国語の表記に合わせて訓令式とは異なる例外表記を行う。例えば、「チ」は訓令式の表記では「ti」であるのに対して、韓国語で「チ」の発音に対応する「치」は「chi」と表記する。韓国語で「ti」と表記する文字は「티」で、日本語に無い音である。そこで「チ」は「chi」に統一する。同様に、「ツ」は「chu」で統一する。

3.1.2 韓国語のローマ字変換

ハングルは組み合わせ文字である。一音節(一文字)には必ず一つの子音と母音が必要で、最初の子音を「初声」、母音を「中声」と呼ぶ。さらに、「終声」という形で子音(または子音の組み合わせ)が付く場合もある。従って、すべての組み合わせは初声(19)×中声(21)×終声(27＋1)＝11,172 とおりで、韓国語の文字対ローマ字対応表を網羅的に人手で作成することは手間がかかる。そこで、文字コード情報に基づくローマ字変換の規則化を行った。

文字体系の観点から、韓国語の文字コードは字母を単位にコード化した「組み合わせ型」と、音節を単位にコード化した「完成型」に分類できる。韓国で最もよく使われている EUC - KR は「完成型」で、通常よく使う 2,350 文字を字母とは無関係に符号化しているため、2,350 通りの文字対ローマ字の対応が必要になる。それに対して「組み合わせ型」は字母とローマ字の対応を作成すればよい。しかし、「組み合わせ型」は一般的に使われていない文字コードである。一方、Unicode における韓国語コードは「完成型」であるものの、初声、中声、終声と文字コードの間に規則性がある。Unicode における韓国語コードの一部を図 2 に示す。

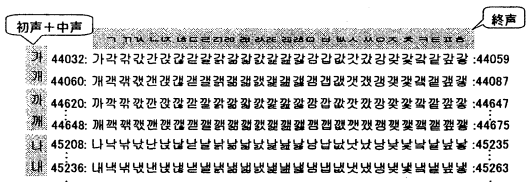

図 2 : Unicode における韓国語コードの一部

図 2 より、初声、中声、終声には以下の規則性がある。

- 中声は一行(28 文字)ごとに変わり、21 行ごと繰り返し循環する。
- 初声は 21 行ごとに変わる。
- 終声は列ごとに変わり、28 列それぞれが異なる終声に対応する。

本研究では Unicode を使い、初声、中声、終声のローマ字対応を 68 通りだけ作成し、韓国語をローマ字に変換する手法を用いた。

ローマ字表記には、音声特徴の再現に適切な「国語ローマ字表記法」を用いた。但し、日本語カタカナ語とマッチングするために、前処理で 15 通りの表記を日本語ローマ字に合わせる。例えば、韓国語の子音「ㄹ」が終声になるときは「l」と表記する。しかし、日本語のローマ字表記には「l」のアルファベットを使わないので、日

本語に合わせて「r」に統一する。韓国語コーパスとローマ字変換後の例を図3に示す。

상품전략면에서는 브랜드 네이밍과 패키지 디자인을 통해 철저히 숙취 제거기능을 강조했다.
sangpumzeonryakmyeneseoneun beuraendeu neiminggwa paekizidizaineur tonghae cheorzeohi sukchwi zegegineungeur gangzohaessda.

図3：韓国語コーパスのローマ字変換例

3.1.3 韓国語外来語候補の抽出

辞書に登録されていない外来語が形態素解析で過分割されるのを防ぐため、形態素解析を行わずに外来語候補を抽出する。まず、コーパスを文節単位に分割し(韓国語は文節単位に分かち書きする)、自立語に連接する付属語を削除する。また、既存の辞書に登録されている語は取り除く。さらに、外来語表記法に基づき、外来語表記に使わないハングル文字を含む単語を取り除く。以上の操作によって外来語候補を絞り込む。

3.1.4 カタカナ語と韓国語の類似度計算

DP マッチングを用いてローマ字変換されたカタカナ語と韓国語の音韻的な類似度を計算する。DP マッチングとは、二つの文字列の類似度を挿入・置換・削除による最小差異数で測定する方法である。外来語の音訳において、子音は基本的な要素で言語による違いがあまりない。それに対して、母音は各国の言語仕様や外来語表記慣習によって表記に揺れが生じることから、本研究では日本語と韓国語における音訳の違いを考慮し、子音をより重要な要素として考慮する。そこで、類似度を式(1)で計算する。

$$\text{類似度} = 1 - \frac{2\times(w\times\text{子音差異数} + \text{母音差異数})}{w\times\text{子音数} + \text{母音数}} \quad (1)$$

類似度の範囲は0〜1である。式(1)で、wは子音の重要度を制御する定数であり、本研究では経験的にw=2としている。特定の閾値以上の類似度を持つ日韓単語対を対訳として最終的に抽出する。すなわち、韓国語における外来語の抽出と日韓対訳辞書作成を同一の枠組で行うことができる。

3.2 本手法の特長

本手法の特長の一つは、高価な対訳コーパスを必要としないで、日韓対訳辞書を作成する点にある。本研究は入力データとして、韓国語単言語コーパスと、字種情報によって迅速且つ簡単に入手できるカタカナ語辞書のみを必要とする。また、外来語判定の前処理として形態素解析を使用しないことで、辞書に登録されていない外来語が形態素解析で過分割されることを回避できる。

4 評価実験

本手法を評価するために、カタカナ語辞書としてクロスランゲージ社(旧ノヴァ社)の機械翻訳用専門用語辞書と日外アソシエーツのコンピュータ用語大辞典(第二版)から収集したカタカナ語(異なり 111,166 語)を用いた。韓国語コーパスとして、韓国経済新聞社 1994 年の新聞記事コーパス（66,146 記事）から無作為に抽出した 50 記事を用いた。

固有名詞の外来語は、対応するカタカナ語の正解判定が難しく、またカタカナ辞書に登録されていない場合が多かったため、普通名詞による外来語のみ(50 記事中 59 件)を評価の対象とした。

実験の結果、類似度の閾値を 0.5 に設定し、類似度順位を上位 10 まで考慮したとき、外来語は 41 件(72%)抽出できた。精度と再現率の変動は図4の通りである。精度と再現率は式(2)を用いて計算した。

$$\text{精度} = \frac{\text{システムが出力した正しい外来語数}}{\text{システムが出力した外来語総数}}$$
$$\text{再現率} = \frac{\text{システムが出力した正しい外来語数}}{\text{コーパスに現れた全外来語数}} \quad (2)$$

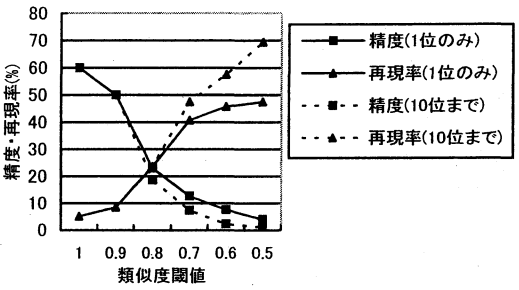

図4：外来語抽出の精度と再現率

図4に示す通り、類似度の閾値を高く設定すると精度が上がり、再現率が下がった。逆に類似度の閾値を下げて網羅性を高くすると、再現率は向上するものの精度は低下した。

抽出に失敗した主な原因は、対応するカタカナ語が辞書に登録されていなかったこと(3 件)、正解の外来語とカタカナ語の類似度が閾値の 0.5 より低かったこと(3 件)、外来語が複合語になっていたこと(12 件)などが挙げられる。例えば、「어학프로그램(語学プログラム)」は

複合語の一部が外来語でないために抽出できなかった。但し、「비디오테이프(ビデオテープ)」のようにすべて外来語からなる複合語は抽出できた。

次に、本手法で実験対象の 50 記事から自動抽出した韓国語外来語辞書(667 語)を形態素解析(オムロンソフト社 SuperMorpho-K) の辞書に追加して、50 記事を解析した結果を表 1 に示す。解析精度を式(3)に示す。

$$\text{解析精度} = \frac{\text{分割・品詞付与の両方が正しい形態素数}}{\text{システムが出力した形態素総数}} \quad (3)$$

表 1：形態素解析の実験結果

総記事数(総文数)	50(1,213)
総形態素数(総外来語数)	2,987(124)
外来語解析誤り(普通外来語)	30(5)
解析精度	94.6%
外来語解析精度	75.8%
辞書追加後の解析精度	94.8%
辞書追加後の外来語解析精度	79.8%

表 1 に示す通り、外来語の解析精度は 75.8%から 79.8%に 4 ポイント向上し、外来語辞書の追加が外来語の解析に有効であったことを示している。また、外来語辞書を追加する以前に比べて全体の解析精度はほとんど変化せず、辞書の追加によって外来語以外の解析に対する副作用がなかったことを示している。

5 おわりに

本研究ではカタカナ語を手がかりにして韓国語における外来語を抽出し、日韓対訳辞書を作成する手法を提案した。さらに、形態素解析に応用して抽出した辞書の有効性を実証した。

残された研究課題として、抽出失敗の主要な原因であった複合語への対処がある。また、日本語と韓国語における外来語表記の特徴を分析する必要がある。例えば、音訳の際に日本語では長音を使うのに対して、韓国語では長音を使わない。このような両言語体系の違いに基づいて類似度計算法を改善する必要がある。

参考文献

[1] Kil-Soon Jeong, Yun-Hyung Kwon, Sung-Hyun Myaeng. Construction of Equivalence Classes of Foreign Words Through Automatic Identification and Extraction. Proceedings of the Natural Language Processing Pacific Rim Symposium, pp. 335-340. 1997.

[2] Byung-Ju Kang, Jae-Sung Lee, Key-Sun Choi. The Phonetic Similarity Measure for the Korean Transliterations of Foreign Words. Journal of Korean Information Science Society, Vol. 26, No. 10, pp. 1237 –1246, 1999.

[3] Jae-Sung Lee. Automatic Construction of a Transliteration Dictionary from Bilingual Corpus.제 11 회 한글 및 한국어 정보처리 학술대회 전북(第 11 回ハングル&韓国語情報処理学術発表論文集 全北), pp. 142-149, 1999.

[4] Jae-Sung Lee. 다국어 정보검색을 위한 영-한 음차 표기 및 복원 모델 (多言語情報検索のための英韓外来語翻字及び復元モデル). 한국과학기술원 박사학위논문(韓国科学技術院 博士論文). 1999.

[5] Jae-Sung Lee, Key-Sun Choi. Automatic Foreign Word Transliteration Model for Information Retrieval. Proceedings 'of the 4th Conference of Journal of Korean Society for Information Management, Seoul, Korea, pp. 17-24, 1997.

[6] Sung-Hyun Myaeng, Kil-Soon Jeong. Back-Transliteration of Foreign Words for Information Retrieval. Information Processing and Management, Vol. 35, No. 4, pp. 523 – 540, 1999.

[7] Masaaki Nagata, Teruka Saito, Kenji Suzuki. Using the Web as a Bilingual Dictionary. Proceedings of the ACL-EACL Workshop on Data-Driven Machine Translation, pp. 95-102, 2001.

[8] Philip Resnik. Mining the Web for Bilingual Texts. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 527-534, 1999.

[9] 熊野明, 平川秀樹. 言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成. 情報処理学会研究報告, 94-NL-100, pp. 89-96, 1994.

[10] 高尾哲康, 富士秀. 対訳テキストコーパスからの対訳語の自動抽出. 情報処理学会研究報告, 96-NL-115, pp. 51-58, 1996.

[11] 松尾義博,白井論. 発音情報を用いた訳語対の自動抽出. 情報処理学会研究報告, 96-NL-116, pp. 101-106, 1996.

[12] 山本由紀雄, 坂本仁. 対訳コーパスを用いた専門用語対訳辞書の作成. 情報処理学会研究報告,93-NL-94, pp. 85-92, 1993.