英語固有名詞のカナ表記への変換

3N－5

平石 智宣　延澤 志保　斎藤 博昭　中西 正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

現在の日英機械翻訳システムは大規模な辞書を利用しているものが多いが、人名・地名に代表される固有名詞は無数の組合せが存在し、そのすべてを辞書に逐一登録するのは不可能である。よって、これらの固有名詞の大多数はシステムのなかで未知語として処理される。我々が英語の固有名詞を翻訳する際にはカナ表記で記述するのが普通であり、英語固有名詞のカナ表記への変換のシステムは有益であると考えられる。

2. 先行研究の紹介

代表的なものとして文献[1][2][3]が挙げられる。文献[1]は生成されたカナ表記を発音させ、その音声で評価を行っていて80％以上正答率が報告されていたが、追実験を行い活字での評価を行った結果、正答率は60％弱であった。他2つの文献[2][3]も高い正答率は報告されていない。

3. 本研究の概要

本研究では対応規則を、アルファベット表記とカナ表記の対訳関係にある英語固有名詞データから自動抽出し、変換の際に文字列の共起情報を用いずに、その対応規則と出現頻度情報のみを用いて、高精度な変換を可能にすることを目的とする。

1. 英語固有名詞辞書から対応規則の抽出
2. 変換対象の英語文字列の前処理
3. 変換対象の英語文字列の分割
4. カナ表記へ変換

3.1 英語固有名詞辞書から対応規則の抽出

英語固有名詞辞書の英語文字列を母音字で分割し、対応するカナ文字列も促音・長音・拗音等を前の文字に含めて分割する。そしてそれぞれの分割数が一致した場合、対応規則と判断し、変換テーブルと呼ばれる対応規則を格納するテーブルに登録する。以上の処理を入力データ全体に対して施す。次に上記の処理で対応規則を抽出できずに残ったデータから、変換テーブルの対応規則中にあるものをそれぞれの文字列から取り除く。この操作で生成された新しい文字列に対し、上記の対応規則の抽出方法を適応する。

3.2 変換対象の英語文字列の前処理

この前処理は略語に対して行う。略語の多くは連語の頭文字で形成され音読できない。よって、略語は前処理の段階で例外扱いをしたい。そこで、入力された英語文字列の全ての文字が大文字であるならば、略語であると判断し入力文字列をそのまま出力する。しかし人名の接頭辞として一部分に大文字が混入している場合は、その部分を小文字に変換して通常の処理を行う。

3.3 変換対象の英語文字列の分割

1. 英語文字列を任意の場所で分割し、各部分が変換テーブル中に登録されている対応規則となるような分割方法を発見する
2. 1の操作で複数通りの分割方法が存在する場合は、評価値を計算し最大のものを採用する

評価値は分割された部分の数が少ないものを優先、つまり文字列のマッチングの最長一致のものを選択する。さらにその分割数が同じ場合は、各部分の対応規則の出現頻度の合計を計算し、その値の最大のものを選択する。

3.4 カナ表記への変換

前処理によって分割された英語文字列の各部分を、評価値算出に用いた対応規則を用いてカナ表記へ変換する。本研究の評価値が適切なものであるかを見るため、以下の2種類の出力を行い正答率を比較し考察する。

Translating English Proper Nouns into Japanese Kana
Tomonobu HIRAISHI Shiho NOBESAWA Hiroaki SAITO Masakazu NAKANISHI
{ tomonobu,shiho,hxs,czl } @nak.ics.keio.ac.jp
Department of Computer Science, Faculty of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

1. 第一変換候補のみの出力
2. 第一変換候補 + 評価値残りの上位 3 つの出力

4. 実験結果および評価

4.1 使用するデータ

対応規則の抽出のためのデータは英語固有名詞データ (WWW 上の辞書、データ数 6789) を用いた。辞書 [5] から抜粋した英語人名データ (データ数 559)・英語固有名詞データ (除：人名、データ数 898) を変換の対象とした。前者のデータ中より、後者のデータ中に存在しているものは除外してある。つまり、変換対象のデータは変換テーブルを自動構築するデータに対して完全に未知語であると言える。よってこの実験より、本研究が未知語に対しても有効であることを証明できる。

4.2 ゆれの範囲

日本語のカナ表記にはゆれが存在する [6][7]。本研究では [6][7] を参考に以下の範囲のものを正しい変換と見なす。この範囲は先行研究 [1] と同様である。

1. 完全に一致するもの
2. 1 文字違うもの (1 文字長い・短い場合を含む)
3. 2 文字異なるもの

4.3 対応規則抽出実験結果

本実験に先立つ前に、対応規則を抽出するまでの実験を行い、抽出されたものが妥当なものであるかを確かめた。その際 “ar” などに代表される、二重母音字・三重母音字の多くが抽出できなかった。そこで、先行研究 [1][3] でも問題であった子音字 r 、n に関しての共起情報を対応規則抽出の際に用いた。この共起情報は実際の文字列を指定した細かいものではなく、前後の母音字・子音字の有無による簡易的なものである。

4.4 変換実験結果

実験結果として表 1 のような正答率を得た。データ全体として 70 % を超える正答率になり、先行研究に匹敵する結果を得た。

表 1: 実験結果

データ名	評価値最大	評価値上位 4 つ
英語人名データ	70.48 %	77.10 %
英語固有名詞データ	71.71 %	77.62 %
データ全体	71.24 %	77.42 %

5. 考察

5.1 *n-gram* 等の共起情報の導入

本研究では簡易的な共起情報を対応規則の抽出の段階で用いたが、*2-gram*、*3-gram* などの共起情報を同時に用いることで対応規則の抽出の精度・英語文字列の分割の精度がさらに向上することが予想される。

5.2 評価値算出式の確定

本研究での評価値はマッチングの最長一致と対応規則の出現頻度を評価値の基準にしたが、互いを独立に計算している。最長一致の法則と出現頻度の 2 つの尺度を総合した評価値算出式の確定が全体の正答率の向上につながると予想される。

5.3 その他

その他以下の点に注目することで、正変換率が向上することが予想される。

1. さらに大規模なデータからの対応規則の抽出
2. カナ表記のゆれの修正
3. コスト最小法などの仮名漢字変換に用いられる手法の導入

6. おわりに

本研究の手法を用いることで、英語固有名詞の未知語に対して 70 % を超える正答率を得られた。しかも、完全な未知語のテキストから得られたものであるので、本研究の結果は有益なものであると考えられる。また多数の変換規則や大規模な辞書は用いていないので、他の日英翻訳システムなどの中に組み込むことも十分可能であると考えられる。

参考文献

[1] 住吉 英樹, 相沢 輝昭：英語固有名詞の片カナ変換, 情報処理学会論文誌 vol.35 no.1 pp.35–45, 1994.

[2] 増田 恵子, 梅村 恭司：人名辞書から名前読み付与規則を抽出する試み, 情報処理学会 自然言語処理研究会報告 120-15 pp.97–102, 1997.

[3] 堀内 雄一, 山崎 一生：英単語のアルファベット表記から仮名表記への変換, 情報処理学会 自然言語処理研究会報告 79-1 pp.1–8, 1990.

[4] 宮内 忠信：片カナ表記からの英単語検索システムの実現, 情報処理学会 自然言語処理研究会報告 97-17 pp.119–126, 1993.

[5] NEW COLLEGIATE ENGLISH–JAPANESE DICTIONARY 5th edition, 研究社

[6] 国語審議会：外来語の表記, 答申, 1991. 2.

[7] 国立国語研究所：現代表記のゆれ, 秀英出版, 1983.