7M-6 日本人英語の音声認識を実現する英日音訳法の検討

齋藤邦子† 篠原章夫‡ 永田昌明† 小原永†
†NTT サイバースペース研究所 ‡NTT サイバーソリューション研究所

1. はじめに

WWW の発展によりインターネットユーザーは年々増加している。キーボードの扱いに慣れていないユーザーにとって、音声で制御できるブラウザは非常に魅力的である。我々は 1997 年、日本語音声を入力とする音声制御ブラウザ VCWEb を開発した[1]。しかし WWW には国境がないため、日本語サイトが外国語サイト、特に英語サイトへリンクされている場面は非常に多い。英語サイトを音声で制御するには英語音声認識ソフトが必要であるが、従来の英語音声認識ソフトは英語を母国語とする人の発音を音響モデルとしているために、日本人の発音する英語を殆ど認識しない。そこで、本研究では英語を日本語へ音訳する英日音訳アルゴリズムと日本語音声認識ソフトを組み合わせ、日本人英語を認識させる手法を検討した。

2. VCWeb

VCWeb は、リンクから音声コマンドを自動生成するプロキシサーバーと、ユーザーの発話を認識して音声コマンドを実行するクライアント PC の 2 部から構成される。音声コマンドは JavaScript で記述され、その生成過程は以下の 4 ステップからなる。

1) HTML テキストからアンカー部を抽出する。
2) アンカー部を形態素解析し名詞と読みを得る。
3) これらの名詞で音声コマンドを生成する。
4) 音声コマンドを元 HTML テキストに追加する。

ユーザーが発話したリンクの単語を認識すると、リンク先の HTLM テキストに対してこの処理を自動的かつ高速に行う。形態素解析エンジンには JTAG、音声認識ソフトには VoiceRex を利用した[2,3]。

第 2 ステップに英日音訳処理を追加し、英語サイトの HTML テキストについてはリンクの英単語を日本語表記へ変換する。すると日本語音声認識ソフトでも英語を日本語として認識することが可能となる。

3. 音訳アルゴリズム

英日音訳については幾つか従来研究がある[4,5]。最も単純な手法は変換規則を全て辞書に登録するものであるが非常にコストが高い。そこで変換規則を統計的学習により生成する手法がある。

A Transliteration Algorithm to Recognize English pronunciation of Japanese.
Kuniko Saito, Akio Shinohara, Masaaki Nagata, Hisashi Ohara
NTT Cyber Space Laboratories
NTT Cyber Solution Laboratories

Knight らは正書法表記を一度発音表記に変換した後に英語から日本語へと変換する統計的音訳モデルを提案した[4]。この手法は複数の変換ステップを経るため非常に複雑である。本研究では音訳モデルを単純化するために英語表記から日本語表記へ直接変換し、変換規則の学習には最小限の人手の介入を許して変換規則の精度を改善する手法を検討した。

日本人が英語を読む時には前後の環境を考えながら文字列単位で変換する。例えば「chance」は「cha/チャ n/ン ce/ス」と、「chase」は「cha/チェイ se/ス」と読む。これは、英文字列 $E=e_1e_2\cdots e_n$ と日本語文字列 $J=j_1j_2\cdots j_n$ の同時確率 $P(E,J)$ を最大化する組み $(\hat{E},\hat{J})$ を求める問題に帰着する。

$$(\hat{E},\hat{J})=\arg\max_{E,J} P(E,J) \quad (1)$$

本研究では $P(E,J)$ の計算に文字列単位の bigram 確率モデルを導入した。ここで、e_i, j_i は 1 文字を表すのではなく、文字列（例えば、cha,チャ）である。また *bos,eos* は語頭・語末を表す特殊記号である。

$$P(E,J)=P(e_1,j_1\mid bos)\times\prod_{i=2}^{n}P(e_i,j_i\mid e_{i-1},j_{i-1})P(eos\mid e_n,j_n) \quad (2)$$

このモデルを学習するためには予め英単語とその日本語読みが文字列単位で対応付けられた大量のデータが必要である。単語単位で対応したデータは外来語辞書として入手できるが、それらが文字列単位で対応しているデータは無い。そこで単語単位の対応データから文字列単位に対応付けるアルゴリズムを検討した。

4. 対応付けアルゴリズム

この対応付けアルゴリズムでの入力は英単語と日本語読みの対であり、出力は文字列単位の対である。

change-チェンジ → cha-チェ n-ン ge-ジ

本研究では文字列単位の対応コストを評価するために編集距離の概念を導入した。

英単語を $X=x_1x_2\cdots x_n$、その日本語読みを $Y=y_1y_2\cdots y_m$ とする。ここで x_i, y_i は 1 文字を表す。要素 $x_1\cdots x_i$ と要素 $y_1\cdots y_j$ の距離を $d(i,j)$ とする。要素 $x_{i-s+1}\cdots x_i$ と要素 $y_{j-t+1}\cdots y_j$ が $s:t$ で対応している時、その編集距離を $cost(x_{i-s+1}\cdots x_i, y_{j-t+1}\cdots y_j)$ とすると次式が成立する。

$$d(i,j)=d(i-s,j-t)+cost(x_{i-s+1}\cdots x_i, y_{j-t+1}\cdots y_j)$$
$$\text{ただし}\quad d(0,0)=0 \quad (3)$$

編集距離の値は予め設定した対応規則を用いて以下のように求める。

$y_{j-t+1}\cdots y_j$ が $x_{i-s+1}\cdots x_i$ の対応規則にある時

$$cost(x_{i-s+1}\cdots x_i, y_{j-t+1}\cdots y_j)=0$$

それ以外の時

$$cost(x_{i-s+1}\cdots x_i, y_{j-t+1}\cdots y_j)=1 \qquad (4)$$

各地点 (i,j) で全ての $s:t$ 対応について編集距離 $cost(x_{i-s+1}\cdots x_i, y_{j-t+1}\cdots y_j)$ を計算し、距離 $d(i,j)$ を求める。その時、距離 $d(i,j)$ の最小値と直前の地点 $(i-s, j-t)$ を記憶しておく。これを語頭から語末まで繰り返し、語末まで至ったら記憶された直前の地点を順に辿り、最小距離となる対応付けを求める。図1に対応付けの例を示す。矢印が要素間の対応を表し、数字はその編集距離である。

対応規則は予め用意しておかなければならない。高精度の対応付けを実現するには、十分な量の対応規則を用意するのが望ましい。図1から分かるように、一度対応付けを行うと、処理結果から新しい対応規則が得られる（cha-チェ）。そこで本研究では、人手で初期対応規則を設定し、対応付け⇒結果人手修正⇒新規則追加の作業を繰り返して規則を増やした。初期規則は、アルファベットの子音と母音の組み合わせに対する読みと、全てのアルファベット1文字に対する読みの候補を標準的な日本人の発音に基づいて系統的に列挙したもので、約400個であった。100個の英日単語対からなるサブセットを10個用意し、対応付け作業を繰り返した結果、初期規則は約1000個まで増えた。この対応規則を使って外来語辞書の英日単語対（約6万個）を対応付け、英日文字列対応データを作成した。

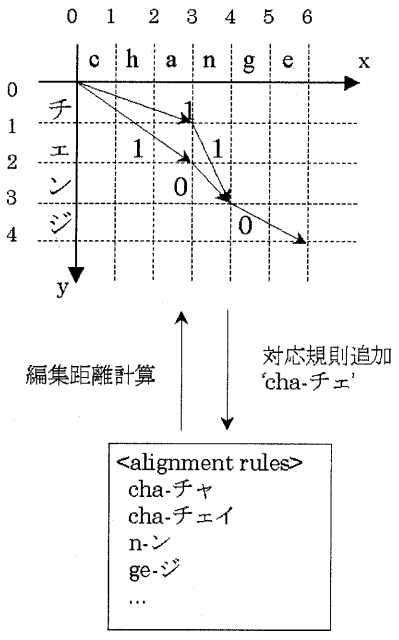

図 1. 対応付け例

5. 実験

英日文字列対応データの9割（約5.4万個）から(2)式の文字列bigramモデルを学習し、残りの1割をオープンテストセットとした。このテストセットを音訳し、外来語辞書のエントリと比較したところ、1位の出力では80%、上位3位までででは92%の正解率が得られた。

6. VCWebへの適応

この英日音訳アルゴリズムをVCWEbに追加し、英語サイトではリンク上の英単語を日本語へ上位3候補音訳させた。この拡張版VCWebで幾つかの英語サイト（新聞、ポータルサイトなど）の閲覧を試みたところ、日本語サイトと遜色無い使用感を得ることができた。この結果には2つの要素があると考えら得る。1つは、通常リンク上には複数の単語があり、システムが全ての単語を誤認識することは稀であるということである。もう1つは、音訳結果は音声認識の内部で使用されるので、必ずしも音訳が正しくなくてもシステムがうまく動くことが多いということである。それは、もし音訳結果が誤っていても、音声認識部でユーザーの発話に最も近い単語をリンクの単語から選択するからである。音声認識精度は英語でも日本語でも同程度の精度（95%以上）である。これは我々の音訳アルゴリズムは十分実用に耐えることを示している。

7. まとめ

本研究は英日音訳の新しい手法を提案し、1位の出力では80%、上位3位までででは92%の正解率を得た。この音訳手法を日本語音声制御ブラウザVCWebに組込んで日本人英語の音声認識を実現した。

参考文献

[1] 渕・加藤: WWWブラウザの音声による制御, 情報処理学会研究報告 97-SLP-16-7, pp37-42, 1997.

[2] T. Fuchi and S. Takagi: Japanese Morphological Analyzer using Word Co-occurrence -JTAG-, COLING-ACL 98, pp409-413, 1998.

[3] 野田・山口・大附・小川・中川・今村: 音声認識エンジンVoiceRexの開発, 音響学会講演論文集, 2-1-19, pp91-92, 1999-09.

[4] K. Knight and J. Graehl: Machine Transliteration. Computational Linguistics, 24(4) pp599-612, 1998

[5] 増田・梅村: 人名辞書から名前読み規則を抽出するアルゴリズム, 情報処理学会論文誌, 40(7) pp2927-2936.