# 仮名漢字変換ログの活用による言語処理精度の自動向上

森 信介　　Graham Neubig
京都大学 情報学研究科

## 1　はじめに

学習に基く音声言語処理は一定の成果を上げるているが、多くの研究は、データの集収方法を含めた全体戦略を欠いている。実際、ある問題を設定し、コストを投じてその問題の入力と出力の組の例を作成し、機械学習を適用して一定の精度で入力から出力を推定した結果の報告は枚挙に暇がない。規則による方法からこのような方法への移行により、アルゴリズムとデータが分離された。その結果、言語処理システムの構築において、ある程度の分業や並行作業が可能になり、さらに分野適応も比較的容易になった。しかしながら、データをどのように集収あるいは作成するかという課題が依然として残されたままである。

この言語資源の収集の問題を、有用な音声言語処理システムの一般への提供により解決するというのが本論文の提案である。音声言語処理システムの利用過程で得られる情報は、利用結果から得られる情報よりもはるかに多い。実際、本論文で具体例とする仮名漢字変換システムは、未知語も変換候補として列挙することが可能で、この利用過程の情報には、未知語の単語境界やその読みが含まれる。これらは、完成された文には含まれない情報である。

インターネットが十分に普及した現在、ありとあらゆる分野の大量のテキストが利用可能であるとの主張もあるが、実際には、企業内での活動報告書やカルテなど、インターネットではアクセスできない分野の文も多数ある。このような分野での音声言語処理にも高い需要があり、この需要に短い開発期間で応えるためにも、活動報告書やカルテの執筆過程において副次的に産生される情報を蓄積しておくことは、公開・非公開にかかわらず非常に重要である。

このような考察に基づいて、本論文では、有用な音声言語処理システムをまず一般ユーザーに提供し、その利用過程で得られる音声言語に関する情報を活用し更なる精度向上を図るという新しい音声言語処理パラダイムを提案する。このパラダイムに沿った具体例として、未知語も変換候補として列挙する仮名漢字変換システムを挙げる。これを一般ユーザーの利用に供することで単語境界と読み情報が付与された文の断片が得られる。このデータを用いることで、単語分割や読み推定など解析系の音声言語処理の精度向上がコストなしで実現できることを示す。

## 2　仮名漢字変換システム

この節では、変換ログの収集に用いた仮名漢字変換システムについて説明する。

### 2.1　確率的モデルによる仮名漢字変換

変換ログの収集には、確率的モデルによる仮名漢字変換 [1] を用いた。確率的モデルによる仮名漢字変換は、キーボードから直接入力可能な入力記号 $\mathcal{Y}$ の正閉包 $\boldsymbol{y} \in \mathcal{Y}^+$ を入力として、変換候補文字列 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots)$ を確率 $P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ の降順に提示する。ここで、$P(\boldsymbol{y}|\boldsymbol{x})$ は、確率的仮名漢字モデルであり、日本語文 $\boldsymbol{x}$ を所与とした入力記号列の生成確率を表す。また、$P(\boldsymbol{x})$ は確率的言語モデルである。

#### 2.1.1　確率的言語モデル

確率的言語モデルには、単語 2-gram モデルを用いている。このモデルは、文を単語列 $\boldsymbol{w}_1^h = w_1 w_2 \cdots w_h$ とみなし、これらを文頭から順に以下の式を用いて予測する。

$$M_{w,2}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|w_{i-1}) \tag{1}$$

この式の中の $w_i$ $(i \leq 0)$ は、文頭に対応する特別な記号であり、$w_{h+1}$ は、文末に対応する特別な記号である。完全な語彙を定義することは不可能であるから、未知語を表わす特別な記号 `UW` を用意する。未知語の予測の際は、まず、単語 2-gram モデルにより `UW` を予測し、さらにその表記 (文字列)$\boldsymbol{x}_1^{h'}$ を以下の文字 2-gram モデルにより予測する。

$$M_{x,2}(\boldsymbol{x}_1^{h'}) = \prod_{i=1}^{h'+1} P(x_i|x_{i-1}) \tag{2}$$

この式の中の $x_i$ $(i \leq 0)$ と $x_{h'+1}$ は、それぞれ、語頭と語末に対応する特別な記号である。確率 $P(w_i|w_{i-1})$ の値は、単語に分割された文からなるコーパスから最尤推定する。

#### 2.1.2 確率的仮名漢字モデル

確率的仮名漢字モデルは、日本語文を単語列 $\boldsymbol{w}$ とみなし、単語と入力記号列との対応関係がそれぞれ独立であると仮定することで以下の式で表される。

$$M_{PM}(\boldsymbol{y}|\boldsymbol{w}) = \prod_{i=1}^{h} P(\boldsymbol{y}_i|w_i) \quad (3)$$

ここで、部分入力記号列 $\boldsymbol{y}_i$ は単語 $w_i$ に対応する入力記号列であり、$\boldsymbol{y} = \boldsymbol{y}_1\boldsymbol{y}_2\cdots\boldsymbol{y}_h$ を満たす。確率 $P(\boldsymbol{y}_i|w_i)$ の値は、単語ごとに入力記号列が付与されたコーパスから最尤推定する。

### 2.2 無限語彙の仮名漢字変換

変換ログの収集に用いた仮名漢字変換システムは、テキストコーパスの部分文字列を変換候補とするように拡張されている [2]。以下では、この拡張について説明する。

#### 2.2.1 サブワード モデル

ある文字列と入力記号列との関係を記述するために、文字と入力記号列の組を単位とするサブワードモデルを用いる。このモデルでは、まず、ある表記 $w = x_1x_2\cdots x_m$ に対応する入力記号列を各文字 $x_i$ の入力記号列 $y_i$ の連接とし、次に、その出現確率 $P(\boldsymbol{y}|w)$ を各文字に対応する入力記号列が一様に出現すると仮定して、以下のように計算する。

$$P(\boldsymbol{y}|w) = P(\boldsymbol{y}|x_1x_2\cdots x_m) = \prod_{i=1}^{m} \frac{1}{|\boldsymbol{\mathcal{Y}}_{x_i}|} \quad (4)$$

ここで、$\boldsymbol{\mathcal{Y}}_x$ は文字 $x$ に対応する可能な入力記号列の集合であり、単漢字辞書を参照することで得られる。

#### 2.2.2 文脈の記述

サブワード モデルが列挙する単語候補を適切に選択するために、その文脈を適切に記述する必要がある。このためには、仮名漢字変換を適用する分野のコーパスから言語モデルを推定することが望ましい。これを実現するために、単語分割情報がないテキストコーパスから文献 [3] の方法を用いて推定した単語 2-gram モデルを用いる。この方法では、テキストコーパスの各文字間に単語境界確率を付与し、確率的単語分割コーパスとし、単語 2-gram 確率を期待頻度から計算する。

単語境界確率は、単語分割済みコーパスから推定した最大エントロピー法に基づくモデルを用いた [4]。

テキストコーパスの部分文字列も候補にする仮名漢字変換においては、単語分割済みコーパスから推定した言語モデル $P_g$(式 (1)(2)) とテキストコーパスから推定した言語モデル $P_r$ を以下のように補間して用いる。

$$P(w_i|w_{i-1}) = \lambda_g P_g(w_i|w_{i-1}) + \lambda_r P_r(w_i|w_{i-1}) \quad (5)$$

$\lambda_g$ と $\lambda_r$ は補間係数であり、削除補間によって求める。

#### 2.2.3 無限語彙の仮名漢字変換

テキストコーパスの部分文字列も候補にする仮名漢字変換は、式 (3) と式 (4) で表記の候補をその生成確率とともに列挙し、式 (5) で与えられる言語モデルの確率を掛けることで得られる文全体での生成確率の降順に変換候補を提示する。

## 3 仮名漢字変換ログの活用

本論文では、仮名漢字変換ログの自然言語処理への活用の例として、自動単語分割と読み推定の精度向上について述べる。自動単語分割や読み推定は、学習コーパスがある分野では実用上十分な精度が実現されている。したがって、現在の課題は分野適応であり、これを容易に実現するために、部分的にアノテーションされたコーパスからの学習が可能である必要がある。この機能があるテキスト解析器として KyTea[1] [5] がある。この節では、まず KyTea の概略を説明し、次に仮名漢字変換ログについて述べる。

### 3.1 自動単語分割

単語分割は、各文字間を判定点とし、点推定の 2 値分類器で行われる[2] 。参照される特徴量は、判定点の前後 $w$ 文字の文字 $n$-gram と文字種[3] $n$-gram、そして単語辞書情報である。単語辞書情報は、判定点で終わる文字列が辞書に含まれているか、判定点から始まる文字列が辞書に含まれているか、判定点を含む文字列が辞書に含まれているかを表す特徴量である。それぞれの特徴量は単語長によって区別される。

### 3.2 読み推定

読み推定は、単語に分割された文を入力とし、単語毎に独立に以下の分類に基づいて行われる。

$\mathbf{Q}_1$ 学習コーパスに出現しているか
- はい
  - $\mathbf{Q}_2$ 読みが唯一か複数か
    - 複数 ⇒ 分類器 [6] を用いて読みを選択
    - 唯一 ⇒ その読みを選択
- いいえ

[1] `http://www.phontron.com/kytea/` (2010 年 1 月 10 日)
[2] 現在の KyTea の分類器は LibLinear[6] である。
[3] 文字種は、漢字、片仮名、平仮名、ローマ字、数字、その他である。

表 1: 変換ログの例

| 入力記号列 | 確定結果 | 備考 |
|---|---|---|
| ごいかし | 語彙/ごい/RC, 化/か/IN, し/し/IN | 適切な単位と入力記号列 |
| けいたいそ | 形態素/けいたいそ/RC | 単位の不一致 (正: 形態, 素) |
| ひんしを | ひんし/ひんし/UW, を/を/IN | 誤った確定 (正: 品詞, を) |

IN: 一般分野の言語モデルの語彙, RC: テキストコーパスの部分文字列, UW: 未知語

表 2: コーパスの諸元

| | 分野 | 単語境界 | 入力記号 | 文字数 | 入力記号数 | 単語数 | 文 数 |
|---|---|---|---|---|---|---|---|
| $C_{g,a}$ | 一般 | ○ | ○ | 1,945,254 | 2,621,067 | 1,352,161 | 56,924 |
| $C_{g,r}$ | 一般 | × | × | 60,065,893 | — | — | 8,335,449 |
| $C_{t,r}$ | 論文 | × | × | 9,990,893 | — | — | 210,085 |
| $C_{t,a}$ | 論文 | ○ | ○ | 12,775 | 17,205 | 8,666 | 265 |

$\mathbf{Q}_2'$ 辞書に入っているか
はい ⇒ 最初の項目の読みを選択
いいえ ⇒ 各文字の一般的な読みの連接

### 3.3 仮名漢字変換ログの利用

前節で説明したテキストコーパスの部分文字列も変換候補とする仮名漢字変換システムによって得られる変換ログの例を表 1 に示す。変換ログの主な情報は、ユーザーが確定した表記と入力記号列の組の列である。入力の単位は、多くの場合完全な文ではなく、文断片である。また、誤まって確定した結果や、2 文字の人名などを他の単語を用いて 1 文字ずつ入力する過程などを含む。したがって、変換ログは、ノイズありの単語分割済みかつ入力記号列付与済みの文断片からなるコーパスと見做すことができる。

本論文では、このような仮名漢字変換ログをコーパスや辞書として、自動単語分割器や読み推定器から参照することを提案する。

## 4 評価

仮名漢字変換ログの利用による効果を評価するために、仮名漢字変換システムを実装し、これを実際に用いて変換ログを収集し、これを利用した自動単語分割と読み推定の実験を行なった。この節では、実験の結果を提示し、本論文で提案する枠組みを評価する。

### 4.1 言語資源

表 2 は実験に用いたコーパスの諸元である。単語境界と入力記号列を付与した一般分野のコーパス $C_{g,a}$ は、現代日本語書き言葉均衡コーパス (33,147 文) [7] と、日常会話のための辞書の例文 (14,754 文) と新聞記事 (9,023 文) からなる。一般分野の生コーパス $C_{g,r}$ は新聞記事 (8,335,449 文) からなる。対象分野の生コーパス $C_{t,r}$ は 2004 年から 2009 年の言語処理学会年次大会の論文である。テストコーパスとして用いるために単語境界と入力記号列を付与した対象分野のコーパスは、文献 [8] と本論文の本文[4] である。

また、タグ付きコーパスと同じ基準に従う単語と入力記号列の組 154,201 個を含む辞書 $D_u$ [9] を用いた。

### 4.2 仮名漢字変換ログ

仮名漢字変換エンジンは、2 節で説明した確率的モデルによる方法で実現されており、式 (1) と式 (2) の言語モデルは $C_{g,r}$ の自動分割結果[5] と $C_{g,a}$ の合計 8,392,373 文からを推定した。語彙は、9 分割した学習コーパスの 2 箇所以上に出現し、かつタグ付きコーパスか辞書に含まれる 98,448 語とした。式 (3) の仮名漢字モデルは、タグ付きコーパス $C_{g,a}$ と辞書から推定した。コーパスにおける頻度を計数した上で、辞書に含まれる組の頻度を 2 加算した。

部分文字列を変換候補とするために、対象分野の生コーパス $C_{t,r}$ を用いた。このコーパスを倍率 2 の疑似確率的単語分割コーパスとし、頻度 2 以上の部分文字列を単語候補とした。

上述の仮名漢字変換エンジンを文献 [8] と本論文の執筆に用い、その変換ログを収集した。なお、変換ログには、論文執筆期間におけるメールなどの文章作成によるログも少量ながら含まれる。本実験で利用した変換ログは、ユーザーが最終的に選択した単語と入力記号列の組の列であり、表 3 はその諸元である。これ

[4] 本論文は 4.2 項の最後の文までをテストコーパスとした。
[5] 単語自動分割器は $C_{g,a}$ から学習した KyTea である。

表 3: 変換ログの諸元

| 断片数 | 文字数 | 入力記号数 | 単語数 |
|---|---|---|---|
| 3,873 | 12,926 | 19,062 | 7,934 |

「単語数」はユーザーが確定した単位の数を表す。

をコーパスとして利用する場合は $C_l$ と表記し、頻度や文脈情報を捨象して辞書として利用する場合は $D_l$ と表記する。

### 4.3 評価

以下の 3 つの方法で自動単語分割器と読み推定器を構築し、テストコーパス $C_{t,a}$ における単語分割と読み推定の精度を測定した。

**BL:** 学習コーパス $C_{g,a}$ と辞書 $D_u$ から推定

**DA:** 学習コーパス $C_{g,a}$ と辞書 $D_u + D_l$ から推定

**CA:** 学習コーパス $C_{g,a} + C_l$ と辞書 $D_u$ から推定

自動単語分割の評価基準は、単語境界か否かが正しく推定された文字境界の割合である。読み推定の評価基準は、推定結果の入力記号と正解の入力記号との最長共通部分列の文字数を推定結果の文字数で除することで得られる適合率と正解の文字数で除することで得られる再現率の F 値 [10] である。

表 4 はそれぞれのモデルによる精度を示す。まず、ログを利用しない場合の **BL** とログを頻度や文脈情報を捨象し辞書として利用する場合の **DA** との比較についてである。読み推定の精度は少し向上しているが、単語分割の精度はわずかながら低下している。単語分割の精度の低下の原因は、変換ログの単位が単語分割の基準に必ずしも一致しないことにより本来単語ではない断片が辞書に加えられたことであろう。分割誤りの増加にもかかわらず読み推定の精度が向上している理由は、分野特有の読みが獲得できていることである。この典型的な例は、変換ログにおける「両/りょう 端/たん」であり、単語分割基準では「両端」と 1 単語とされるが、自動分割の結果では 2 単語になり誤りとなる。しかしながら、**BL** で「りょうはし」となる読みが適切な「りょうたん」になる。

コーパスとして利用する **CA** の場合には、単語分割と読み推定の双方において精度が向上している。特に読み推定の精度向上は著しい。これは、文脈を参照することで、「2/2 値/ち」における「値」のように、複数の読みがあり一般分野とは異なる読みをする単語に対して読みが適切に推定されていることによる。

以上の結果から、仮名漢字変換ログを利用することで、タグ付きコーパスを準備するコスト無しに単語分割や読み推定の精度が向上することが確かめられた。

表 4: 単語分割と読み推定の精度

| 手法 | 単語分割 | 読み推定 |
|---|---|---|
| **BL** | 98.11% | 98.55% |
| **DA** | 98.10% | 98.57% |
| **CA** | 98.20% | 99.14% |

## 5 結論

本論文では、未知語も変換候補として列挙する仮名漢字変換システムを構築し、その利用過程で得られる変換ログを用いて単語分割と読み推定の精度向上がコストなしで実現できることを示した。これにより、有用な音声言語処理システムを一般ユーザーに提供することと、その利用過程で得られる情報を活用することの有用性が示された。

## 参考文献


[1] 森信介, 土屋雅稔, 山地治, 長尾真: 確率的モデルによる仮名漢字変換, 情処論, Vol. 40, No. 7, pp. 2946–2953 (1999).

[2] 森信介: 無限語彙の仮名漢字変換, 情処論, Vol. 48, pp. 3532–3540 (2007).

[3] 森信介, 宅間大介, 倉田岳人: 確率的単語分割コーパスからの単語 N-gram 確率の計算, 情処論, Vol. 48, No. 2, pp. 892–899 (2007).

[4] 森信介, 小田裕樹: 擬似確率的単語分割コーパスによる言語モデルの改良, 自然言語処理, Vol. 16, No. 5, pp. 7–21 (2009).

[5] Neubig G., 中田陽介, 森信介: 点推定と能動学習を用いた自動単語分割器の分野適応, 言語処理学会年次大会 (2010).

[6] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classication, *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874 (2008).

[7] Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102 (2008).

[8] 森信介, 小田裕樹: 3 種類の辞書による自動単語分割の精度向上, 情報処理学会研究報告, Vol. NL193 (2009).

[9] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵: コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用, 日本語科学, Vol. 22, pp. 101–122 (2007).

[10] 長野徹, 森信介, 西村雅史: N-gram モデルを用いた音声合成のための読み及びアクセントの同時推定, 情処論, Vol. 46 (2006).