自動未知語獲得による仮名漢字変換システムの精度向上

森 信介 †　　小田 裕樹

1　はじめに

生成的な自然言語処理システムは、何らかの記号列を入力として受け取り、自然言語の文を出力する。例えば、音声認識の入力は音響特徴量の列であり、仮名漢字変換ではキーボードからの入力記号列 (読み) である。これらの入力は、自然言語に関する何らかの情報を有していると考えられる。したがって、この情報を用いることで、利用するにつれて精度が向上する自然言語処理システムを構築することが可能であろう。

本論文では、仮名漢字変換システムを例にとり、入力記号列を自動変換するたびにモデルを更新し、これによりシステムの性能を向上させることができることを示す。具体的には、以下の通りである。まず、登録語に加えて、分割情報のないテキストコーパスに出現する全ての部分文字列も変換候補とする仮名漢字変換システムを構築する。次に、その利用の際に得られる登録語以外の表記と入力記号列の組 (単語候補) を用いて辞書を更新する。さらに、単語候補の表記を参照して言語モデルを再推定し、単語候補の文脈情報を獲得する。この結果、より性能が高い仮名漢字変換システムが構築される。

上述のシステム更新の手続きは、人手を一切必要としない。したがって、提案手法は教師無し学習とみることもできる。つまり、出力の必要性にかかわらず入力を処理し、性能向上を図ることが可能である。この性質は、音声認識や機械翻訳など、大量の入力が容易に集められる言語処理において特に有用と考えられる。

2　確率的モデルによる仮名漢字変換

この節では、生成的な自然言語処理システムとして、確率的モデルによる仮名漢字変換について説明する。

2.1　確率的言語モデル

最も一般的な言語モデルは単語 n-gram モデルである。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予測する。

$$M_{w,n}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) \qquad (1)$$

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 n-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 n-gram モデルにより予測する。

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

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。

2.2　仮名漢字変換

確率的モデルによる仮名漢字変換 [1] は、キーボードから直接入力可能な記号 $\mathcal{Y}$ 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。この式において、後半の $P(\boldsymbol{x})$ は確率的言語モデル (LM) であり、上述の単語 n-gram モデルを用いることができる。前半の $P(\boldsymbol{y}|\boldsymbol{x})$ は、確率的仮名漢字モデル (PM) と呼ばれ、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。これは、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。

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

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以下の式を用いて最尤推定する。

$$P(\boldsymbol{y}_i|w_i) = \frac{f(\boldsymbol{y}_i,\ w_i)}{f(w_i)}$$

この式中の $f(e)$ は、事象 e のコーパスにおける頻度を表す。

2.3　無限語彙の仮名漢字変換

前項の仮名漢字変換システムは、学習コーパスに出現する表記と入力記号列の組のみが変換候補になる。

† 日本アイ・ビー・エム東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623-14

この制限を取り払うために、サブワードモデルによる候補の列挙と確率的単語分割コーパスを用いた文脈の記述により語彙をテキストコーパスの部分文字列全てに拡張する方法が提案されている [2] 。

2.3.1 サブワード モデル

文字を単位とするサブワードモデルは、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、単漢字辞書を検索することで得られる。例えば、$\boldsymbol{\mathcal{Y}}_{日} = \{\,カ, ジツ, ニチ, ニッ, ヒ, ビ\,\}$, $\boldsymbol{\mathcal{Y}}_{テ} = \{\,テ\,\}$, $\boldsymbol{\mathcal{Y}}_{レ} = \{\,レ\,\}$ であり、$P(ニッテレ \mid 日テレ) = \frac{1}{|\boldsymbol{\mathcal{Y}}_{日}|}\frac{1}{|\boldsymbol{\mathcal{Y}}_{テ}|}\frac{1}{|\boldsymbol{\mathcal{Y}}_{レ}|} = \frac{1}{6}\times\frac{1}{1}\times\frac{1}{1} = \frac{1}{6}$ となる。

2.3.2 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 n-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 n-gram 確率を期待頻度から計算する。

単語境界確率の決定方法は次の通りである。まず、単語に分割されたコーパスに対して自動単語分割システムの境界推定精度 α を計算しておく。次に、テキストコーパスを自動単語分割し、その出力において単語境界であると判定された点では単語境界確率を α とし、単語境界でないと判定された点では単語境界確率を $1-\alpha$ とする。

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。

$$P(w_i|H_i) = \lambda_g P_g(w_i|H_i) + \lambda_r P_r(w_i|H_i) \quad (5)$$

この式中の H_i は、単語 w_i を予測する際の履歴である。λ_g と λ_r は補間係数であり、削除補間 [4] によって求める。

2.3.3 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 モデルの更新

前節で説明した仮名漢字変換システムは、未登録語であっても、対象分野のテキストコーパスに出現する文字列であれば出力する可能性がある。この節では、出力に含まれる登録語でない表記と入力記号列の組 (未登録語) を利用し、精度向上を図る方法を提案する。

3.1 未登録語の獲得

出力に含まれる未登録語の表記はテキストコーパスの部分文字列であり、入力記号列は実際にタイプされている。さらに、前後の文脈を勘案して出力されているので、ある程度の確からしさで、適切な表記と入力記号列の組合せとみなせる。本論文では、このような未登録語をある一定量の文を自動変換した後に集計し、ある頻度以上の表記とそれに対応する入力記号列を準登録語として記憶し、サブワードモデルの更新と言語モデルの再推定を行なうことを提案する。

3.2 サブワード モデルの更新

前節で説明したサブワードモデルは、可能な入力記号列の出現頻度を 1 とした場合の最尤推定とみなすことができる (式 (4) 参照)。準登録語を含むサブワードモデルによる生成確率は、獲得された単語候補の変換結果における頻度 $f_s(\boldsymbol{y}, w)$ の k 倍をこれに加えた相対頻度として、以下のように算出することとする。

$$P(\boldsymbol{y}|w) = \frac{1 + kf_s(\boldsymbol{y}, w)}{F + kf_s(w)}$$

ここで $F = \prod_{i=1}^{m} |\boldsymbol{\mathcal{Y}}_{x_i}|$, $f_s(w) = \sum_{\boldsymbol{y}} f_s(\boldsymbol{y}, w)$ としている。この式は、頻度の関数となる補間係数

$$\lambda_1 = \frac{F}{F + kf_s(w)}, \quad \lambda_2 = \frac{kf_s(w)}{F + kf_s(w)}$$

を用いて

$$P(\boldsymbol{y}|w) = \lambda_1 \prod_{i=1}^{m} \frac{1}{|\boldsymbol{\mathcal{Y}}_{x_i}|} + \lambda_2 \frac{f_s(\boldsymbol{y}, w)}{f_s(w)}$$

と表せるので、従来の仮名漢字モデルと準登録語辞書との補間と見ることができる。

表 1: コーパス

	分野	表記	入力記号	単語境界	文字数	入力記号数	単語数	文 数
C_g	一般	○	○	○	598,264	811,084	406,021	20,808
C_r	医療	○	×	×	2,270,705	—	—	53,915
C_s	医療	×	○	×	—	50,866	—	900
C_t	医療	○	○	○	4,079	5,403	2,831	100

3.3 言語モデルの更新

確率的単語分割コーパスの単語境界確率の推定に用いる自動単語分割システムは、生成確率 $\mathrm{P}(\boldsymbol{w})$ が最大となる単語列を自動分割結果とする。この生成確率の計算には、式 (1)(2) で与えられる単語 n-gram モデルを用いる [5] 。自動単語分割システムが準登録語を勘案するようにするために、式 (2) において登録語の生成確率を以下の式のように準登録語に均等に分配する。

$$M'_{x,n}(w) = \begin{cases} 0 & \text{if } w \in \mathcal{W}_k \\ M_{x,n}(w) + \dfrac{\sum_{w \in \mathcal{W}_k} M_{x,n}(w)}{|\mathcal{W}_d|} & \text{if } w \in \mathcal{W}_d \\ M_{x,n}(w) & \text{otherwise} \end{cases} \tag{6}$$

ここで $\mathcal{W}_k$ と $\mathcal{W}_d$ はそれぞれ登録語と準登録語の集合を表す。

準登録語辞書が更新された時点で、式 (1) と式 (6) に基づく自動単語分割器を用いてテキストコーパスの単語境界確率を再計算する。その結果得られる確率的単語分割コーパスから言語モデルを、式 (5) の補間係数も含めて再推定する。これにより、提案する仮名漢字変換システムの変換精度が利用するにつれて向上することが期待される。

4 評価

前節で説明した利用するにつれて精度が向上する仮名漢字変換システムの評価として、初期のシステムとある一定の量の入力を処理した後のシステムの性能の比較実験を行なった。この節では、実験の結果を提示し、提案手法を評価する。

4.1 コーパス

実験に用いたコーパスは、主に新聞記事や辞書の例文からなる一般コーパスと医療文書からなる医療コーパスである (表 1 参照)。一般コーパス C_g の各文は正しく単語に分割され、各単語に入力記号列 (読み) が付与されている。医療コーパス C_r はテキストコーパスであり、単語境界や入力記号の情報を持たない。この医療コーパスの 1,000 文に入力記号列を付与し、この内の 900 文 C_s に対してシステムを適用した状態で、残りの 100 文 C_t に対してモデルの性能を測定した (C_r は C_s と C_t を含まない)。

4.2 実験の手順

まず、初期モデルの構築の手順を詳述する。

1. 一般分野のコーパス C_g から、一般分野の単語 2-gram モデル LM_g と、仮名漢字モデル PM_g を作成する。
2. 上記の単語 2-gram モデル LM_g を用いて医療分野のテキストコーパス C_r の各文字間の単語境界確率を推定し、これを確率的単語分割コーパス SSC_r とする。
3. 確率的単語分割コーパス SSC_r から医療分野の単語 2-gram モデル LM_r を作成する。
4. LM_g と LM_r とを補間し、初期モデルの言語モデルとする。補間係数は、医療分野のテキストの生成確率が最大となるように削除補間法により決定する。
5. 単漢辞書から作成した仮名漢字モデル PM_d と一般分野のコーパス C_g から作成した仮名漢字モデル PM_g を初期モデルの仮名漢字モデルとする。

このようにして構築された初期モデルを元に、医療分野の入力記号列 C_s を 100 文づつ自動一括変換し、頻度 2 以上の未登録語を取り出し準登録語辞書に追加し、3 節で説明したモデルの更新を行なうことを繰り返した。この際に、サブワードモデルでの頻度の倍率 k を 10,000 とした。

上記の処理の結果得られる言語モデルの予測力と辞書 (登録語と準登録語) のカバレージをテストコーパス C_t に対して測定した。さらに、これらの言語モデルと辞書をもつ仮名漢字変換システムの変換精度をテストコーパス C_t に対して計算した。

表 2: 仮名漢字変換の精度の変化

累積変換文数	カバレージ	パープレキシティ	変換精度 適合率	再現率
0	88.1%	44.0	96.8%	97.2%
100	89.0%	43.2	96.8%	97.2%
300	91.0%	41.7	97.0%	97.4%
900	93.4%	40.8	97.1%	97.5%

4.3 評価基準

提案手法では、準登録語辞書に変換候補が追加されていく。この評価基準として、テストコーパスに出現する延べ単語数に対する準登録語を含む登録語の割合 (カバレージ) を採用した。

確率的言語モデルの予測力の評価に用いた基準は、テストコーパスにおける単語あたりのパープレキシティである。まず、テストコーパス C_t に対して未知語の予測も含む文字単位のエントロピー H を以下の式で計算する。

$$H = -\frac{1}{|C_t|}\log_2 \prod_{\boldsymbol{w}\in C_t} M_{w,n}(\boldsymbol{w})$$

ここで、$M_{w,n}(\boldsymbol{w})$ は単語 n-gram モデルによる単語列 $\boldsymbol{w}$ の生成確率を、$|C_t|$ はテストコーパス C_t の文字数を表す。次に、単語単位のパープレキシティを以下の式で計算する。

$$PP = 2^{H\times\overline{|\boldsymbol{w}|}}$$

ここで $\overline{|\boldsymbol{w}|}$ は平均単語長 (文字数) である。

仮名漢字変換の評価基準は、各入力文の一括変換結果と正解との最長共通部分列 (LCS; longest common subsequence)[6] の文字数に基づく再現率と適合率である。正解コーパスに含まれる文字数を N_{COR} とし、一括変換の結果に含まれる文字数を N_{SYS} とし、これらの最長共通部分列の文字数を N_{LCS} とすると、再現率は N_{LCS}/N_{COR} と定義され、適合率は N_{LCS}/N_{SYS} と定義される。

4.4 評価

表 2 は、初期の仮名漢字変換システムとある一定の文数の入力記号列を処理した後の仮名漢字変換システムのテストコーパスに対する性能を示す。

まず、処理した文数の増加に伴うカバレージの上昇から、利用するにしたがって獲得された単語を辞書に追加することで未登録語が減少していることが分かる。つまり、テストコーパスに出現する有用な単語が獲得されているといえる。

次に、言語モデルの予測力についてであるが、パープレキシティが単調に減少していることから、予測力が向上していることがわかる。これは、獲得された単語を考慮に入れて言語モデルを再推定した効果であり、獲得された単語の文脈情報を言語モデルが適切に記述していることが分かる。

最後に、変換精度であるが、適合率と再現率の双方において、処理した文数の増加に伴う精度向上が見られる。従来手法 (初期モデル) の再現率は 97.2%と高いので、提案手法により実現された誤りの削減率 (約 10.5%) は十分有意義であろう。

以上の結果から、「自然言語処理システムが実際に用いられる際の入力にも言語に関する何らかの情報が含まれる」との仮定が正しく、その情報を用いる提案手法により、利用するにつれて変換精度が向上していく仮名漢字変換が実現された。

5 結論

本論文では、「自然言語処理システムが実際に用いられる際の入力にも言語に関する何らかの情報が含まれる」との着想から、利用するにつれて精度が向上する自然言語処理システムを構築することが可能であることを示した。実験では、仮名漢字変換を自然言語処理システムの例にとり、一定の文数の入力記号列 (読み) を変換するたびにモデルを更新し、システムの性能を向上させることが可能であることを示した。この結果、利用するにつれて精度が向上する自然言語処理システムを構築すことができることが確認された。

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, 情報処理学会研究報告, Vol. NL172 (2006).

[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. 47 (2007).

[4] Jelinek, F., Mercer, R. L. and Roukos, S.: Principles of Lexical Language Modeling for Speech Recognition, *Advances in Speech Signal Processing*, Dekker, chapter 21, pp. 651–699 (1991).

[5] Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. of the COLING94*, pp. 201–207 (1994).

[6] Aho, A. V.: 文字列中のパターン照合のためのアルゴリズム, コンピュータ基礎理論ハンドブック, Vol. I: 形式的モデルと意味論, Elseveir Science Publishers, pp. 263–304 (1990).