変換ログを用いた仮名漢字変換精度の向上

山口 洋平　　森 信介　　河原 達也
京都大学 情報学研究科
{yohei, mori, kawahara}@ar.media.kyoto-u.ac.jp

1　はじめに

統計的仮名漢字変換 [1] の拡張として，既知語に加えてテキストの部分文字列も変換候補に挙げることができる方法が提案されている [5].

この方法により，Web テキストやメールなどの文書に含まれている任意の文字列が変換候補に挙がるので，ユーザによる単語登録がほぼ不要になる．ユーザがこれを変換結果として選択することで，単語候補がその入力記号列とともに変換ログに記録される．このように，変換ログは，ユーザの入力の傾向のみならず，単語候補とその入力記号列の組，およびその文脈情報を含む.

本論文では，このような変換ログを言語資源とみなし，これを用いて，単語候補とその文脈を自動的に学習する仮名漢字変換システムを提案する．

このような仮名漢字変換システムを実装し，上限としてシステムの出力を完全に修正するユーザを，下限として全く修正しないユーザを想定し，提案手法による精度上昇曲線の上限と下限を計算した．さらに，その中間としての現実的なユーザを想定し，提案手法として計算した．これらの結果から提案手法による変換ログを利用することで自動的に変換精度を向上する仮名漢字変換システムが実現出来ることが示された.

2　統計的仮名漢字変換

本節では，単語と入力記号列の組を言語モデルの単位とする仮名漢字変換システムについて説明する [7].

2.1　定式化

統計的手法による仮名漢字変換 [6] 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，日本語の文字 $\mathcal{X}$ の正閉包を変換結果として出力する．この際，以下の式が示すように，単語 w と入力記号列 $\boldsymbol{y}$ の組 $u = \langle w, \boldsymbol{y} \rangle$ を単位とする言語モデルによる生成確率を評価基準とする.

$$\underset{\boldsymbol{w}}{\mathbf{argmax}}\, P(\boldsymbol{w}|\boldsymbol{y}) \;=\; \underset{\boldsymbol{w}}{\mathbf{argmax}}\, \frac{P(\boldsymbol{w},\boldsymbol{y})}{P(\boldsymbol{y})} \;=\; \underset{\boldsymbol{w}}{\mathbf{argmax}}\, P(\boldsymbol{u}) \qquad (1)$$

ここで単語列 $\boldsymbol{w}$ は文字であることに注意されたい.

式 (1) の $P(\boldsymbol{u})$ は，u を単位とする n-gram モデルを用いて，以下のようにモデル化される.

$$P(\boldsymbol{u}) \;=\; \prod_{i=1}^{h+1} P(u_i|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1})$$

$$P(u_i|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1}) \;=\; \begin{cases} P(u_i|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1}) & \text{if } \ u_i \in \mathcal{U} \\ P(\mathtt{UU}|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1}) M_{u,n}(u_i) & \text{if } \ u_i \notin \mathcal{U} \end{cases} \qquad (2)$$

ここで，$\mathcal{U}$ は言語モデルの語彙 (単語と入力記号列の組の集合) を表す．この式の中の u_i ($i \quad 0$) と u_{h+1} は，文頭と文末に対応する特殊な記号 BT である．また UU は未知の組を表す記号である.

式 (2) の $M_{u,n}(u_i) = M_{u,n}(\langle w, \boldsymbol{y} \rangle)$ は未知語モデルである．大きな学習コーパスを用いれば，実際の使用における未知語率は極めて低く，また未知語に対する正確な仮名漢字変換は困難であると考えて，アルファベット $\mathcal{U}$ 上の代わりにアルファベット $\mathcal{Y}$ 上の未知語モデル $M_{y,n}(\boldsymbol{y})$ を用いることとする．これは，y を単位とする n-gram モデルを用いてモデル化される.

2.2　パラメータ推定

式 (2) の $P(u_i|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1})$ と $P(\mathtt{UU}|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1})$ の値は，単語に分割され，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される.

$$P(u_i|\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1}) = \frac{f(\boldsymbol{u}_i^{\,i\ \ }_{\,n+1})}{f(\boldsymbol{u}_i^{\,i\ \ 1}_{\,n+1})} \qquad (3)$$

ここで，$f(e)$ は事象 e のコーパス内頻度である.

2.3　単語候補の列挙

文献 [7] では，式 (3) のコーパスとして，人手による単語分割と入力記号列が付与された文書に加えて，Web テキストやメールなどの文書に対して，擬似確

システムの提示:	エンゲ/えんげ/UNK 困難/こんなん 、/、 頻脈/ひんみゃく/SUB 、/、
ユーザの選択:	嚥下/えんげ 困難/こんなん 、/、 頻脈/ひんみゃく/SUB 、/、
システムの提示:	旧約/きゅうやく 等/とう 適切/てきせつ な/な 処置/しょち
ユーザの選択:	休薬/きゅうやく/SUB 等/とう 適切/てきせつ な/な 処置/しょち

表記と入力記号列の組の後の「UNK」はそれが未知語であることを、「SUB」はテキストの部分文字列であることを表す。

図 1: 変換ログの例

率的単語分割および擬似確率的入力記号列付与を行った結果を用いることで，既知語に加えてこれらの文書の部分文字列を変換候補として挙げる仮名漢字変換システムを実現している．本論文では，これを実際に用いることで得られる変換ログを利用する．

3 変換ログとユーザのモデル

本節では，まず変換ログについて説明し，次に仮名漢字変換システムを利用するユーザのモデルについて述べる．

3.1 変換ログ

提案手法で用いる仮名漢字変換システムの変換ログは，ユーザが入力した入力記号列とシステムの最尤解とユーザの確定結果からなる (図 1 参照)．

提案手法で用いる仮名漢字変換システムの変換ログは，ユーザがよく変換するドメインや表現といったユーザ個人の言語活動の特性を反映しているだけでなく，語彙には含まれないがテキストの部分文字列として出現する単語候補も含まれる．図 1 の「頻脈/ひんみゃく」は，システムがテキストの部分文字列を変換候補として提示し，ユーザはそれを確定している.「休薬/きゅうやく」は，システムが誤って提示している第 1 候補をユーザが棄却し，2 番目以降の候補から適切な部分文字列を選択した結果である．これらの例からわかるように，変換ログからは，単語と入力記号列の組がその文脈とともに得られる．

3.2 ユーザモデル

現実的なユーザをモデル化するために，まず 2 つの極端なユーザモデルについて考える．

1 つ目は，システムの提示に誤分割や誤変換があっても修正せずにそのまま確定するユーザである．ここでは，このユーザをシステム (system) と呼ぶ．この場合，変換結果に誤変換が含まれていても，そのまま変換ログに書き込まれる．これは，変換ログを言語資源として見たとき，ノイズとなると考えられる．

2 つ目は，システムの提示に誤分割や誤変換があれば必ず修正し，確定するユーザである．ここでは，このユーザをオラクル (oracle) と呼ぶ．この場合，変換ログには必ず正しい変換結果が書き込まれる．このとき，修正された変換結果の中に人手で修正したコーパスや辞書に含まれない変換候補があれば，単語候補として獲得される．

3.3 現実のユーザ

現実のユーザ (user) は上述のモデルのように単純ではなく，両者のある一定の割合での組み合わせであると考えられる．例えば，現実のユーザは一定の割合で誤変換をそのまま確定し，次に誤確定箇所を消去してもう一度入力記号列を入力し正しい変換結果を選択するといった行動をとると考えられる[1]．

このようなユーザの変換ログから得られる確定結果 L_u は，誤確定確率を　として，

$$L_u = L_o + \quad L_s$$

となる．ここで，L_o と L_s はそれぞれ，オラクルとシステムの確定結果である．

このようにして得られる変換ログを言語資源として仮名漢字変換システムの学習コーパスに加えることでユーザの単語登録なしに未知語を文脈とともに自動獲得している仮名漢字変換システムを提案する．

4 実験

本論文で提案する変換ログを利用する仮名漢字変換システムの変換精度が、変換ログの増加に伴ってどのように変化するかをシステム，オラクル，現実のユーザのそれぞれのユーザモデルについて実験を行なった．本節では、その結果を提示し提案手法を評価する．

4.1 コーパス

学習コーパスとして，単語境界と入力記号列が人手で付与されている現代日本語書き言葉均衡コーパス [3] と，これらの情報がない医療分野のコーパスを用いた (表 1 参照)．医療分野のコーパスに対しては，倍率 2

[1]現実のユーザの振る舞いは，誤確定箇所の部分的な消去や，他の箇所からのコピーを含めて様々であると考えられ，本論文での仮定は近似に過ぎない．

表 1: 言語資源の詳細

	分野	単語境界	入力記号	文字数	単語数	文数
学習コーパス	一般	人手	人手	3,919,935	899,030	33,141
学習コーパス	医療	自動	自動	18,132,813	—	275,091
ユーザの入力	医療	自動	自動	256,967	—	30,000
テストコーパス	医療	人手	人手	208,381	45,318	1,000

で疑似確率的単語分割し，さらに倍率 2 で疑似確率的入力記号列付与を行った [7][2]．ユーザが入力する文書として，自動で単語境界と入力記号列を付与した医療分野の 30,000 文を想定した．テストコーパスとして，人手で単語境界と入力記号列を付与した医療分野の 1,000 文を用いた．

4.2 実験手順

想定する仮名漢字変換システムのユーザが現実のユーザ user である場合の実験手順について説明する．

1. 学習コーパスから言語モデルを構築する．
2. ユーザの入力から 300 行を変換する．
3. その確定結果 L_u を学習コーパスに追加し，仮名漢字変換システムを再構築する．L_u に含まれる L_s は，確定結果それぞれに対して，0 から 1 までの乱数を発生させ，その乱数の値が　よりも小さいならば，その確定結果を選択することを表す．
4. テストコーパスを用いた評価を行う．
5. ユーザの入力全てに行き渡るまで 2. から 4. までを繰り返す．

想定するユーザが oracle である場合には L_u を L_o とし，想定するユーザが system である場合には L_u を L_s とする．誤確定確率は，あるユーザ 1 名の変換ログから推定した結果として　$= 13/200$ とした．

4.3 評価基準

評価基準は，カバー率と変換精度である．カバー率は，テストコーパスの単語と入力記号列の組の内の仮名漢字変換システムの語彙に含まれている割合である．仮名漢字変換システムの語彙は，人手で準備された一般分野の学習コーパスと変換ログから構成される．したがって，カバー率は変換ログから得られた単語候補がどの程度有効であったかを示す．

変換精度は，各入力文の一括変換結果と正解との最長共通部分列 [2] の文字数に基づく F 値である．正解コーパスに含まれる文字数を N_{REF} 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文字列の文字数を N_{LCS} とすると，F 値は次のように定義される．

$$\text{F 値} = \frac{2N_{LCS}}{N_{REF} + N_{SYS}}$$

4.4 評価

変換ログを学習コーパスにある回数，追加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図 2, 図 4 に示す．

まず，図 2 によると，カバー率において，変換結果の追加回数の増加に伴うカバー率の上昇から，単語候補が変換ログに出現することで語彙に変化し，未知語が獲得できていると言える．oracle と user のカバー率は最終的にはほぼ同じになることがわかる．一方で，図 3 によると，語彙サイズに関しては user の方が oracle よりもわずかに大きい．これは擬似確率的に生成したコーパスに含まれる分割誤りなどによる単語候補が変換ログに出現したからだと考えられる．

次に，図 4 によると，変換精度において，変換結果の追加回数の増加に伴う精度向上が見られる．user, oracle, system の順で精度上昇の度合いが大きいことがわかる．

以上の結果から，変換ログを学習コーパスに追加することによって，未知語が獲得でき，さらに分野適応できることがわかる．

5 最後に

仮名漢字変換タスクにおいて変換ログを学習コーパスに追加することで精度の向上を確認し，変換ログを用いた仮名漢字変換タスクにおける提案手法での精度の上限と下限について述べた．

今後の課題として，今回の実験では文単位での変換を採用したが，ユーザが実際に日本語入力システムを用いる状況を考えると，文節単位程度での実験を行う必要があると考えられる．また，変換ログの利用方法

[2] 単語境界確率や入力記号列の確率の推定には KyTea[4] を用いた．

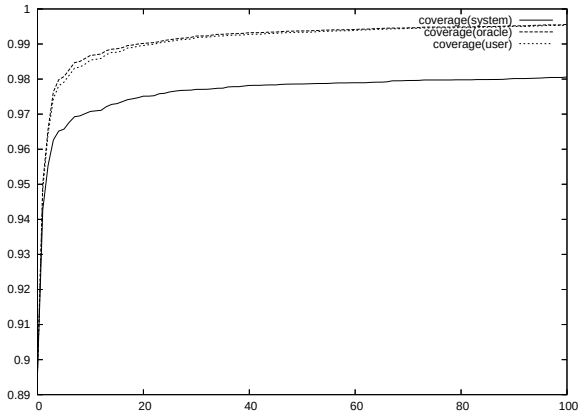

図 2: カバー率

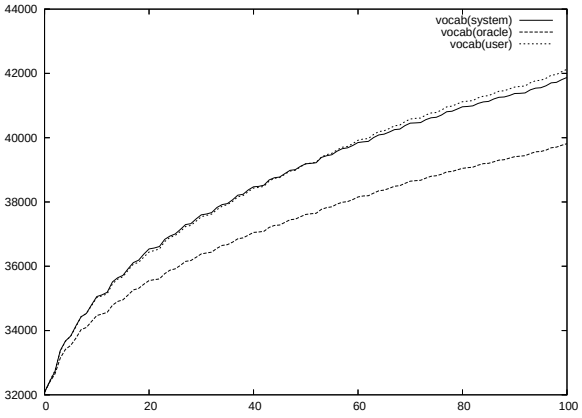

図 3: 語彙サイズ

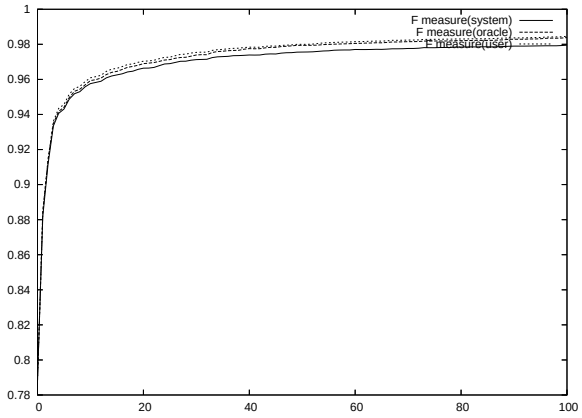

図 4: F 値

を工夫することによる下限の精度上昇曲線の底上げがある．

参考文献

[1] Zheng Chen and Kai-Fu Lee. A new statistical approach to chinese pinyin input. In *Proc. of the ACL00*, pp. 241–247, 2000.

[2] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. The MIT Press, 1990.

[3] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources*, pp. 101–102, 2008.

[4] Graham Neubig and Shinsuke Mori. Word-based partial annotation for efficient corpus construction. In *Proc. of the LREC10*, 2010.

[5] 森信介. 無限語彙の仮名漢字変換. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2006, No. 36, pp. 17–24, 2006-03-27.

[6] 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. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 98, No. 48, pp. 93–99, 1998-05-28.

[7] 森信介, 笹田鉄郎, Neubig Graham. 擬似確率的タグ付与コーパスからの言語モデル構築. 情報処理学会研究報告, 2010.