情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-NL-198 No.2
2010/9/16

HMMを用いて分野適応する仮名漢字変換

黒崎弘光　　山口和紀

近年大規模なコーパスを用いた統計的仮名漢字変換が注目されている．しかし，一般的な分野の辞書を用いると対象分野特有の単語において仮名漢字変換の変換精度は低下してしまう．変換対象の分野に応じた辞書を使うと、仮名漢字変換の精度が向上するが，そのためには変換対象の分野を推定する必要がある。HMM を用いて単語ごとの分野の推定を行うと単語に関連性がない場合推定した分野が大きく変動してしまう．先行研究では 10 単語単位で状態を変化させていたものもあるが，若干の精度の向上にとどまった．そこで本研究では HMM の構造で単語間の関連性を表現して各単語の分野を推定する方法を提案する．HMM で文章の分野を推定し，分野に適した辞書を用いることによる仮名漢字変換の変換精度を調べたところ，適応分野における変換精度が向上した。

Japanese Imput Method Adaptation Using Hidden Markov Model

Hiromitsu Kurosaki and Kazunori Yamaguchi

Statistical approach to Japanese input method is popular these days. But it is difficult to convert in a specific domain. We consider a state as a topic of sentences, and estimate the states with Hidden Markov Model. In this paper, we improve a structure of HMM, because it is difficult to estimate the topics with the basic structure of HMM. We made experimental evaluation on a task of Japanese input method and observed an improvement in the accuracy.

1. はじめに

近年，大量のテキストデータが手に入るようになり，コーパスから自動的に変換規則や辞書を学習する統計的仮名漢字変換が提案されている．[1)2)3)] 統計的仮名漢字変換のシステムで代表的なものとしては Social IME[2)] や ChaIME[3)] などがある．どちらも単語 2gram モデルをベースとしており，Google 日本語 N グラム[⋆1] といった大規模コーパスから推定した統計量を用いることで高精度な変換を可能にしている．また，音声認識や単語分割といったタスクでは対象分野の辞書を用いることで精度を上げることに成功しており[4)]，仮名漢字変換でも同様に対象分野の辞書を用いることが出来れば精度が向上すると考えられる．しかし，これらの仮名漢字変換システムでは言語モデルを一般的な分野のコーパスから構築しており，対象分野特有の単語や表現に対しては精度が低下する．また，話題が化学から歴史に変化する文章，といった話題となっている分野の変化に合わせて変換候補を変化する，といったことは難しい．

文章中のトピックの変化を捉えた先行研究に柴田ら[5)] の HMM を用いてトピックの遷移を推定したものがある．しかし彼らの手法では，ある程度人手によるルールを必要としており，コストがかかる．他の先行研究としては，貞光ら[6)] が評価文章分類という問題に対して文を単位とする HMM を用いることで文書構造を捉えようとしたものなどがある．「引用」を表している文，「異なる対象への評価」を表している文といった単語単位では表現できない情報を，彼らの研究では単語単位の HMM ではなく文単位 HMM の状態で表現することにより文章の構造を捉えることを提案しており，評価文章分類の精度を改善している．

本論文では，文章の分野の変化を捉えるのに隠れマルコフモデルの構造を工夫した HMM を用いることを提案する．提案手法によりより正確に分野の変化を捉えることができ，その結果を仮名漢字変換に活用することで，話題となっている分野の変化を捉えて用いる辞書を変化させることができ，仮名漢字変換の精度が向上すると考えられる．なお，仮名漢字変換にはひらがな文字列を単語ごとに分割するというステップと単語ごとに分割されたひらがな文字列から適切な変換候補を推定するというステップの二つがある．そのうち本論文では後者のステップのみを対象とする．

2. 先行研究

本論文では分野が変化する文章を対象にするため，分野ごとに異なる単語の出現しやすさを学習する必要がある．しかし，人手で単語コスト（あるひとつの単語の出現しやすさ）や単語同士の連接コスト（2 つの単語のつながりやすさ）を記述した辞書を作成するのは大変

†1 東京大学総合文化研究科
Department of General Systems Studies,University of Tokyo

⋆1 http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html

である．そこで，森ら[1)]が提案した統計的仮名漢字変換を利用する．統計的仮名漢字変換ではコーパスから自動的に辞書に相当する確率モデルを学習するため，上記の問題に対して有効である．

また、文章中の分野の変化をモデル化する必要があり，本論文では HMM を用いることで文章の分野の変化をモデル化することを考える．ただし単語単位で HMM を用いると過適応を起こしやすいため，貞光ら[6)]が提案した，文を単位とする HMM や隠れマルコフモデルの構造を工夫した HMM を用いることでスムージングを試みる．

2.1 統計的仮名漢字変換

統計的仮名漢字変換では，仮名文字列の入力 y が与えられたとき変換候補 $(x_1, x_2, ..x_k)$ に対して以下の（1）式を満たすように x の添字をつける．

$$i \geq j \Leftrightarrow P(x_i|y) \geq P(x_j|y) \tag{1}$$

したがって，仮名漢字変換の主な役割は各変換候補の確率値 $P(x|y)$ の順序関係を求めることである．したがって，この順序関係が保持されていればよいので，ベイズの定理を用いて次のように変形する．

$$P(x_i|y) \geq P(x_j|y) \tag{2}$$

$$\Leftrightarrow \frac{P(y|x_i)P(x_i)}{P(y)} \geq \frac{P(y|x_j)P(x_j)}{P(y)} \tag{3}$$

$$\Leftrightarrow P(y|x_i)P(x_i) \geq P(y|x_j)P(x_j) \tag{4}$$

$P(x)$ は確率的言語モデルと呼ばれ，仮名漢字混じりの文 x がある言語である確率を表す．$P(y|x)$ は確率的仮名漢字モデルと呼ばれ，仮名漢字混じりの文 x が与えられたときのキーボードからの入力が仮名文字列 y である確率を表す．また，以下ではそれぞれのモデルについて説明する．

まずは確率的言語モデルについて述べる．確率的言語モデルとは与えられた文字列がある言語の文である尤度を数値化したものであり，よく使われる確率的言語モデルとしては単語 n-gram モデルが挙げられる．このモデルは文を単語列 $w_1^h = w_1 w_2 \cdots w_h$ とみなし，これらを先頭から順に予測する．

$$M_{w,n} = \prod_{i=1}^{h+1} P(w_i|w_{i-n+1}^{i-1}) \tag{5}$$

この式の中の $w_i (i < 1)$ 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．仮名漢字変換は単語列 w_1^h の末端の単語を推測するものなので，$x_i = w^{h_1}$ に対応する．次に確率的仮名漢字変換モデルについて述べる．確率的仮名漢字モデルとは，仮名漢字混じりの文 x が与えられた時のキーボードからの入力文字列 y の確率である．単語列 w が与えられたときの確率的仮名漢字モデルによる確率は以下の式で表す．

$$M_{kk}(y|w) = \prod_{i=1}^{h} P(y_i|w_i) \tag{6}$$

ここで入力文字列 y_i は単語 w_i に対応する入力文字列であり，

$$y = y_1 y_2 \cdots y_h \tag{7}$$

を満たす．

$f(e)$ を事象 e が起こる頻度とすると，確率 $P(y|w)$ の値はコーパスから最尤推定することによって決定する．

$$\prod_{i=1}^{h} P(y_i|w_i) = \frac{f(y_i, w_i)}{f(w_i)} \tag{8}$$

以上を統合すると，統計的な仮名漢字変換が列挙する変換候補は以下の式を最大にするような変換候補 x である．

$$P(y|x)P(x) = \prod_{i=1}^{h} P(y_i|w_i)P(w_i) \tag{9}$$

$$P(y_i|w_i)P(w_i) = P(w_i|w_{i-n+1}^{i-1})P(y_i|w_i) \tag{10}$$

例えば単語 2gram モデルならば，次のようになる．

$$P(y|x)P(x) = \prod_{i=1}^{2} P(y_i|w_i)P(w_i) \tag{11}$$

$$= P(w_1|w_0^0)P(y_1|w_1)P(w_2|w_1^1)P(y_2|w_2) \tag{12}$$

2.2 HMM による文章の構造のモデル化

この節では貞光らが提案した文を単位とする HMM について述べる．大局的には肯定的な評価だが局所的に見れば否定的な表現が多い文章に対して単語単位でこの文章を評価した場合否定的評価文章に分類される可能性が高い．このような文章に対して，貞光らはある対象に対する評価を含む文章（評価文章）には「評価対象への評価」を表している箇所以外に「評価対象以外の対象に対する評価」や「他者の意見の引用」を表している箇所があるこ

とに着目し，単語よりも大きな文を単位とした HMM を用いて文章構造をモデル化することで評価文章分類の精度を上げることを提案した．本論文ではこれを文単位 HMM と呼ぶ．そこでは文章の各々の文はある隠れた状態（「引用」状態や「異なる対象への評価」状態）を持ち，その状態が遷移することで文章構造が形成される，というように仮定している．この仮定においては文自体を出力シンボルとしているため，文章 d_k に対し文単位 HMM を用いて付与される確率 P_H を以下のように定義している．

$$P_H(d_k|\mathbf{a},\mathbf{b}) = \sum_{T_k}\prod_{t=1}^{T_k} a_{q_{t-1}q_t} b_{q_t}(\mathbf{s}_{kt}) \tag{13}$$

$$= \sum_{T_k}\prod_{t=1}^{T_k} a_{q_{t-1}q_t} \prod_{n=1}^{|\mathbf{s}|} b_{q_t}(w_{ktn}) \tag{14}$$

$$\tag{15}$$

ここで $\mathbf{s_{kt}}$ は t 番目の単語シーケンスを表し，t は文章 d_k の文番号，T_k は文章 d_k に含まれる文数，w_{ktn} は文 s_{kt} の n 番目に出現した単語，q_t は t 番目の文が滞在する HMM の状態数を表す．また，$\mathbf{a}$，$\mathbf{b}$ はモデルパラメータで，$a_{q_{t-1}q_t}$ は状態 q_{t-1} から状態 q_t へ遷移する確率を表し，$b_{q_t}(w_{ktn})$ は状態 q_t にいるときに単語 w_{ktn} を出力する確率を表す．

文を単位とする上記の文を単位とする HMM では HMM の表す状態に対して特別な制約はない．したがって，貞光らの提案を参考に各状態の表しているものを置き換えることで本論文で対象としている「話題が化学から歴史に変化する文章」といった話題が変化する文章の構造をモデル化することを考える．たとえば文章中の各々の文の状態を「化学」状態や「歴史」状態とし，その状態の遷移が話題の変化だと考えると，文単位 HMM を用いることで「話題が化学から歴史に変化する文章」の文章構造をモデル化することができる．

3. 段階的 HMM を用いた統計的仮名漢字変換の提案

この節では，まず，どのように HMM を用いて文章の分野を推定するか，また推定された各単語の分野をどのように仮名漢字変換に活用するかについて述べる．その後，提案手法である段階的 HMM について説明する．

3.1 HMM による分野の推定

まず，仮名漢字変換ではすでに変換を終えた語と次に変換する単語の読みという異なる種類の情報を利用するため，前節の HMM の方法をそのまま利用することができない．そこでまず，仮名漢字変換の場面で想定している HMM の状態推定の仕方について述べる．文章の分野を HMM の状態で表し，HMM の出力を各分野における単語の出現分布とする．また，単語を 1 単語ずつ順次変換をしていく，という状況とする．このように，単語を 1 単語ずつ変化していく処理を本論文ではオンライン処理と呼ぶ．文章の冒頭から k 番目の単語を変換するとき，HMM が使える情報としては，すでに変換を終えた単語列 $w_1^{k-1} = w_1 \cdot w^{k-1}$ と仮名文字列 y_k が利用出来る．しかし，HMM の出力を各分野における単語の出現分布とするため，仮名文字列 y_k では不都合である．したがって，各状態における変換候補のうち最尤の変換候補（x_{kj}）を状態毎に一旦求める．その後，w_1^{k-1} に変換候補 x_{kj} を加えた単語列について最尤の状態列とその時の確率を求める．これを変換候補を変えて求め直し，最も高かった状態列を HMM の推定した状態列とする．例えば,「酸素、還元、」まで変換を終えていて,「ヨウカイ」という読みを変換変換したい、という状況を考える.「ヨウカイ」の読みを持つ変換候補は，溶解，妖怪，と複数あるが，化学状態では「溶解」が最も出現確率が高かったとすると，**図 1** のように「酸素、還元、溶解」で HMM の状態を推定し，この場合は変換候補の分野は化学となる．一方，歴史状態では「妖怪」が最も出現確率が高かったとすると，**図 2** のように「酸素、還元、妖怪」で HMM の状態を推定をし，変換候補の分野が歴史になることがありうる．変換候補の分野はそれまでに変換を終えた単語に影響をうけるためある程度前の分野と同じ分野になる確率が高いが，文章の分野が変化すると，不適切な分野では単語の出現確率が低いため，HMM の推定する分野もやがては他の分野に変化する．

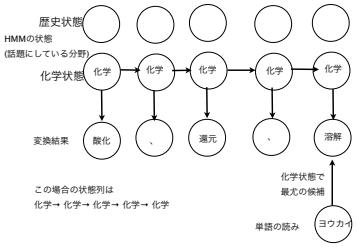

図 1 HMM による分野の推定 1（出典：Wikipedia）
Fig. 1 The estimated sequence of domain

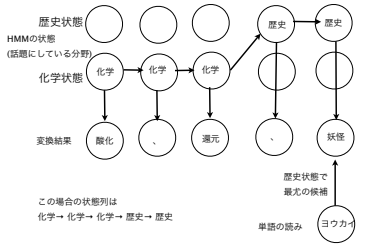

図 2 HMM による分野の推定 2（出典：Wikipedia）
Fig. 2 The estimated sequence of domain

3.2 分野が分かっているときの仮名漢字変換

このようにすると，HMM を用いることで，文章のどのあたりがどの分野を話題にしているのかを推定できる．次に，HMM が推定した分野に応じて仮名漢字変換で用いる辞書を

Vol.2010-NL-198 No.2
2010/9/16

変化させる．そうすることで，変換対象の語の分野に適応した辞書を用いることができる．例えば先程の例で言えば，化学が尤もらしい分野だった場合，化学の辞書を用いて仮名漢字変換を行う．例えば，HMM の出力を各分野における単語の出現確率とし，仮名漢字変換で単語 2gram モデルを用いると，HMM で単語の大雑把な分布をみて文章の分野の変化を考慮し，仮名漢字変換の単語 2gram により文法的な要素も反映した変換となり，変換精度が向上すると考えられる．

3.3 段階的 HMM

一般に文章の分野は急激に変化することは少なく，なだらかに変化すると考えられる．しかし，HMM を直接最尤推定した場合，過適応を起こしやすく，単語単位で状態が大きく変動することが起こりうる．そこで，隠れマルコフモデルの構造を工夫することで状態の変化がゆるやかになるようなモデルを考えた．例えば HMM のマルコフモデルの構造は分野の数が 3 のときは**図 4** のようになるような，隠れマルコフモデルの構造である．図の化学，歴史，一般とは，それぞれ文章の分野を表しており，一般（化），一般（歴），一般（一）はそれぞれ化学，歴史，一般状態に遷移しやすい一般状態である．こうすることで，なるべく同じ状態に長くいるようになると考えられる．また，化学状態から 1 度の遷移で歴史，一般状態には遷移できず，同様に歴史状態から化学，一般状態にも 1 度の遷移では遷移できないようにしてある．そうすることで，たまたま出てきただけの単語の影響を減らし，状態の変化が緩やかになると考えられる．

4. 実　　験

HMM を用いた統計的仮名漢字変換では，HMM による単語の分野の推定と統計的仮名漢字変換による仮名漢字変換の独立した二つのステップに分かれているため，単語の分野の推定方法を工夫することにより仮名漢字変換の部分を変えなくても仮名漢字変換の変換精度を向上させることが可能である．そこで，この節では HMM の構造に工夫をすることでどの程度仮名漢字変換の変換精度の向上に役立つかを調べるために，HMM を用いた仮名漢字変換システムを実装し，その変換精度を評価した．

この章では実験条件，比較対象のモデル，評価基準について述べた後に実験結果とそれに対する考察を述べる．

4.1 実 験 概 要

まず，文章中の分野の変化を捉えられるのかを調べたいので，**図 5** のような文章の途中で分野が変化する文章を仮名漢字変換の対象とする。(図 5 ならば，分野は化学から歴史に変化している)

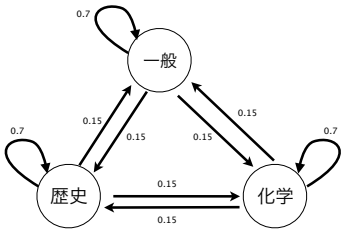

図 3 HMM のマルコフモデルの構造
Fig. 3 the structue of HMM

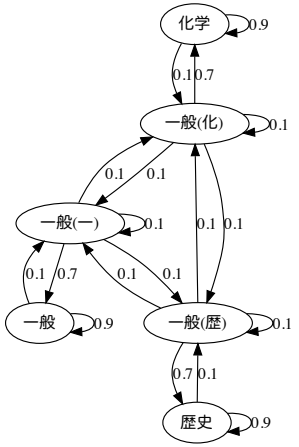

図 4 段階的 HMM のマルコフモデルの構造
Fig. 4 the structure of step HMM

実装した仮名漢字変換システムの概要は以下のようなものである．まず、仮名漢字変換システムは HMM を用いて単語の分野の推定をするステップと単語の読みから仮名漢字変換を行う 2 つのステップに分けられる．HMM を用いて単語の分野の推定をするステップでは，単語分割された仮名漢字混じりの文を与えたときに，推定した変換対象の語の分野を出力する．仮名漢字変換を行うステップでは，仮名漢字変換システムに単語分割された単語の読みを与えると，仮名漢字変換システムは HMM によって推定された分野の単語 1gram,2gram の辞書を用いて統計的仮名漢字変換を行う．

4.2 コ ー パ ス

このシステムを実現するためには，一般的なコーパスから作成した単語 1gram，単語 2gram の辞書，適応分野のコーパスから作成した単語 1gram，単語 2gram の辞書，単語分割されて読みが付与されたテストコーパスが必要になる．以下ではそれらの作成方法について述べる．

実験に用いるコーパスには Wikipedia の文章を用いた．というのも，Wikipedia は百科

事典であるため様々な分野の文章があり，用語に関する説明だけでなく歴史や関連分野についても言及しているため，本論文が対象にしているような文章が入手しやすいからである．Wikipedia の文章は Wikipedia データダンプ XML ファイル[7] という形で入手可能である．ただし，このファイルはテキストデータには不要な Wiki タグや他のページへのリンクなどが含まれており，そのままでは日本語のコーパスとしては扱いにくい．そこで WP2TXT[8] を用いてテキスト抽出を行ない，記事タイトルや項目の見出しを除いて，1 個以上の句点（。）を持った行のみを有効な言語データとして利用した．次に，このままでは単語分割されておらず，また単語の読みも不明であるので，実装した仮名漢字変換システムで利用するためには文を単語単位で分け，単語の読みを付与する必要がある．そこで，形態素解析エンジン MeCab 0.96[9] を用いて形態素解析を行った．形態素解析辞書には ipadic-2.7.0[9] を用いた．

このようにして作成した単語分割された文章から一般的な単語 1gram，2gram の辞書，適応分野の単語 1gram，2gram の辞書，分野に変化があるテスコトーパスを作成した．一般的な単語 1gram，2gram の辞書は様々な分野を含んだ大量の文章から作成した．したがって，この単語 1gram，2gram の辞書は未知語は少ないが，特定の分野以外では余り出てこない単語については扱いにくいと考えられる．適応分野の単語 1gram，2gram の辞書は著者が文章を分野にあわせて分類し，各分野ごとに作成した辞書である．一般的な単語 1gram，2gram の辞書よりは未知語も多く学習コーパスのサイズが小さいために偶然性の影響を受けやすいが，適応分野においては一般的な単語 1gram，2gram の辞書よりもより精度の高い仮名漢字変換を行うのに役立つと考えられる．テストコーパスも著者が話題の変化がある文章を選び出すことで作成した．テストコーパスは，例えば図 5 のように局所的には異なる話題の単語が混ざっていつつも単語より大きな単位で見た場合は 1 つの分野であり，また例で言えば化学から歴史へと話題が移っているような話題にしている分野が変化している文章である．

コーパスの大きさは表 1 の通りである．テストコーパスは 3 つの文章について変換を行ない，実験ではその時の平均正答率，標準偏差について記載した．また，表 1 にはテストコーパスの合計の大きさを記載した．

4.3 比較対象のモデル

提案手法を評価するために，単語 2gram モデル，HMM-単語 2gram モデル，文単位 HMM-単語 2gram モデル，段階的 HMM-単語 2gram モデルの 4 つのモデルにおける変換精度を調べた．各モデルの詳細は次のようになる．

炎は有機物の酸化反応によって放出される熱エネルギーの現れであるから、化学の歴史は人類が火を扱いはじめたときから始まっているとも考えられる。金あるいは銀以外の金属は、自然界には酸化物ないしは硫化物として産出するので、古代における青銅器・鉄器などの金属精錬も化学反応である還元反応を知らずと利用しているのである。 — 主に化学に関する文章

古代ギリシアにおける学問の発展はアリストテレスにより大成されたが、 — 主に歴史に関する文章

図 5 テストコーパスの例：出典（Wikipedia）
Fig. 5 example

・単語 2gram モデル

統計的仮名漢字変換システムで代表的なものとしては Social IME[2] や ChaIME[3] などがある．どちらも確率的言語モデルに単語 2gram モデルを用いており，単語 2gram モデルは統計的仮名漢字変換でよく使われるモデルだと考えられる．したがって，他の HMM-単語 2gram モデルなどを比較するときのベースラインとして，単語 2gram モデルによる仮名漢字変換の変換精度を用いることにする．

単語 2gram モデルでは 0 頻度問題の影響を受けやすいため，次の式のようにして影響を

表 1 コーパス

分野	用途	文数	単語数	文字数
一般	学習	118189	約 400 万	6932052
化学	学習	280	8044	13488
歴史	学習	230	5559	18414
化学+歴史	テスト	272	7266	16423

緩和している．

$$P(y|x)P(x) = \lambda_1 P(x_i) + \lambda_2 P(x_i|x_{i-1}) \quad (16)$$

重み λ_1，λ_2 は本論文では最大エントロピー法を用いて学習データにおける仮名漢字変換の精度が最尤になるように決定した．

・HMM-単語 2gram モデル

他の工夫をした HMM-単語 2gram モデルを評価するために，単純に HMM と単語 2gram モデルを組み合わせたモデルにおける仮名漢字変換の変換精度を調べることにする．HMM-単語 2gram モデルは具体的には，HMM を用いて各単語ごとに分野を推定し，推定された分野に対応した辞書を用いて単語 2gram モデルと同様にして変換を行う．HMM のマルコフモデルの構造は分野の数が 3 のときは**図 3** のようになる．

・文単位 HMM-単語 2gram モデル

貞光らは文という単語よりも大きな単位で文章の構造を捉えるという，文を単位とする HMM モデルを提案している．文単位 HMM-単語 2gram モデルは，貞光らの提案した文単位 HMM を用いることで，文という単語よりも大きな単位で分野をモデル化することを意図したモデルである．文単位 HMM では文を単位として状態を変化させるが，貞光らの実験では 10 単語単位で状態が変化するようにしていたので本論文でも 10 単語を文とみなし 10 単語単位で状態が変化するようなモデルにした．文単位 HMM で状態を状態を推定した後は，他のモデルと同様に，推定された分野に対応した辞書を用いて単語 2gram モデルと同様にして変換を行う．文単位 HMM は状態の変化のタイミングが HMM と異なるものなので，マルコフモデルの構造は HMM と同じである．

・段階的 HMM-単語 2gram モデル

3 章で説明した HMM のマルコフモデルの構造は分野の数が 3 のときは**図??**のようになるような隠れマルコフモデルの構造を用いたモデルである．HMM を用いて各単語ごとに分野を推定た後は，他と同様に，推定された分野に対応した辞書を用いて単語 2gram モデルと同様にして変換を行う．

表 2 辞書

分野	種類	単語数
一般	単語 1gram	101338
一般	単語 2gram	878793
化学	単語 1gram	1293
化学	単語 2gram	4355
歴史	単語 1gram	2128
歴史	単語 2gram	6472

正解の文 (8 単語)：固体 、 液体 、 気体 など が ある
変換結果　　：固体 、 液体 、 機体 など が ある

図 6 正答率
Fig. 6 percentage of correct answers

4.4 評価基準

仮名漢字変換の変換精度の評価には単語単位で変換した場合の第一候補の変換精度（正答率，再現率，適応率）を用いた．正答とは，本論文では仮名漢字変換の出力がテストコーパスと一致した場合のことを指す．

正答率，再現率，適合率については以下の式で求められる値であり，再現率，適合率に関しては森[1] と同一の方法で計算した．

正答率について**図 6** の例を用いて説明する．もとの文章の単語数は 8 単語であり，単語が正しく変換された数は 7 単語である．この場合，正答率は 7/8 である．

$$\text{正答率} = \frac{\text{正答した単語数}}{\text{テストコーパスの単語数}} \quad (17)$$

$$\text{再現率} = \frac{\text{正答した文字数}}{\text{テストコーパスの文字数}} \quad (18)$$

$$\text{適合率} = \frac{\text{正答した文字数}}{\text{仮名漢字変換の出力の文字数}} \quad (19)$$

4.5 パラメータ

HMM のパラメータを以下に挙げる．

・隠れ状態 s_i：文章の分野にあたる．本論文では，歴史，化学，一般の 3 種類を考える．歴史，化学はそれぞれ歴史，化学に関する語を出力しやすい状態，一般は特定の分野を想定しておらず，一般的にな傾向に従って語を出力する状態

・状態遷移確率 a_{ij}：状態 s_i から状態 s_j へ遷移する確率．学習データに対して何度か実験を行い，最も仮名漢字変換の変換精度が高くなるような値に決定した．

・単語の出力確率

状態 s_i のとき単語 w_k を出力する確率．**b** は学習コーパス中の単語の出現確率を用いた．

4.6 実験結果

提案手法の評価するために，オンライン処理をした場合のテストコーパスに対する変換精度を調べた結果，表 3 のようになった．オンライン処理とは，3 章で説明したような HMM による単語の分野の推定にすでに仮名漢字変換を終えた単語と変換対象の語の読みから推

定される尤もらしい単語の情報を用いて分野の推定を行った場合の仮名漢字変換である．読みから推定される尤もらしい単語は，各分野の辞書を用いた場合に最尤となる変換候補である．表から段階的 HMM-単語 2gram モデルが他の HMM を用いた仮名漢字変換よりも高い変換精度を示しており，提案手法の有効性が確認できた．HMM, 文単位 HMM を用いた場合はベースラインである単語 2gram よりも変換精度がわずかながら低下しており，これは上手く分野を捉えられず変換対象の単語と関係ない分野の辞書を用いたためだと考えられる．また，オンライン処理では変換対象の語よりも後ろの語の情報が使えないため，文章の後ろの方の情報を使える場合と比べて HMM の推測は難しくなる．

では，どの程度難しくなっているのかを調べるために，HMM が変換対象よりも後ろの語の情報も使って分野の推定ができた場合について調べた．具体的には HMM に文章全体を与えて仮名漢字変換を行う前に HMM に各単語の状態を推定してもらい，その推定結果に従って適応分野の辞書を用いて仮名漢字変換を行った結果，表 4 のようになった．この場合は HMM-単語 2gram モデルが最も高い正答率を示し，HMM-単語 2gram モデルでは変換対象よりもあとにある単語の情報がわかることが大きく正答率に影響することがわかった．

表 3 オンライン処理の場合の精度

モデル	用いた辞書の分野	正答率 [%]	適合率 [%]	再現率 [%]
HMM-単語 2gram モデル	一般, 化学, 歴史	94.29	94.34	93.81
標準偏差		0.303	1.30	1.24
文単位 HMM-単語 2gram モデル	一般, 化学, 歴史	94.24	94.12	93.87
標準偏差		0.933	1.73	1.32
段階的 HMM-単語 2gram モデル	一般, 化学, 歴史	94.99	94.99	94.62
標準偏差		1.35	1.95	1.29
単語 2gram モデル（ベースライン）	一般	94.31	94.47	94.40
標準偏差		3.35	1.94	1.85

表 4 分野の推定が最高の場合の精度

モデル	用いた辞書の分野	正答率 [%]	適合率 [%]	再現率 [%]
HMM-単語 2gram モデル	一般, 化学, 歴史	96.2	95.27	94.87
標準偏差		1.04	0.616	0.661
文単位 HMM-単語 2gram モデル	一般, 化学, 歴史	94.37	94.51	93.83
標準偏差		2.21	2.86	1.98
段階的 HMM-単語 2gram モデル	一般, 化学, 歴史	95.57	95.5	95.34
標準偏差		1.96	2.87	2.80
単語 2gram モデル（ベースライン）	一般	94.31	94.47	94.40
標準偏差		3.35	1.94	1.85

5. おわりに

本論文では文章の分野の変化を精度よく捉えるための方法として，マルコフモデルの構造を工夫した HMM を提案した．HMM の推定した分野にしたがって用いる辞書を変化する統計的仮名漢字変換を行ったところ，適応分野における変換精度が向上した．

今後の課題としては，さらに辞書を増やすことでさらなる変換精度の向上を目指すことや，単語分割されていない状態でも適切に状態の推定や仮名漢字変換を出来るのかを調べることが挙げられる．

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，情報処理学会論文誌， Vol.40, No.7, pp.2496–2953 (1999).
2) 奥野　陽，荻原将文：インターネットを用いた日本語入力システム，情報処理学会研究報告， Vol.2009, No.36, pp.1–6 (2009).
3) 小町　守，森　信介，徳永拓之：あいまいな日本語のかな漢字変換，情報処理学会夏のプログラミング・シンポジウム，pp.51–55 (2008).
4) 倉田岳人，森　信介，雅史西村：講義関連コーパスを利用した音声認識システムの自動適応，電子情報通信学会論文誌， No.9, pp.2530–2540 (2007).
5) 柴田知秀，黒橋禎夫：言語情報と映像情報を統合した隠れマルコフモデルに基づくトピック推定，情報処理学会論文誌， Vol.48, No.6, pp.2129–2139 (2007).
6) 貞光九月，山本幹雄：文を単位とする文章構造を用いた評価文章分類，言語処理学会第 13 回年次大会発表論文集， Vol.13, pp.230–233 (2007).
7) ： Wikipedia データダンプ XML(jawiki-lastest-pages-articles.xml.bz2)，http://download.wikipedia.org/jawiki/latest/ (2010 年 7 月アクセス).
8) 長谷部陽一郎：Wikipedia 日本語版をコーパスとして用いた言語研究の手法，言語文化， Vol.9, No.2, pp.373–403 (2006).
9) 工藤　拓：形態素解析エンジン MeCab（和布蕪），http://mecab.sourceforge.net/ (2010 年 7 月アクセス).