読み推定のための教師なし単語分割

内海慶 [1]　森信介 [2]
[1]SB Intuitions 株式会社　[2] 京都大学
kei.uchiumi@sbintiotions.co.jp　forest@i.kyoto-u.ac.jp

概要

従来，読みの推定を行う形態素解析では教師あり学習が用いられてきた．しかし，人手による読み情報の付与はコストが大きく，読み付きコーパスは一部の言語資源に限られている．一方で，読みの付与された単語辞書や生コーパスは入手可能なものが多数存在している．そこで本研究では，読みを含む単語辞書と生コーパスを用いて読み推定と単語分割の教師なし学習の提案を行う．提案手法を用いることで，教師なし学習手法で F 値 90 程度の読み推定精度を達成した．

1　はじめに

入力文字列を単語やトークン系列へ分割する単語分割やトークナイズは自然言語処理における基盤技術の 1 つである．従来は MeCab[1)] や KyTea[1] をはじめとする形態素解析が用いられてきたが，近年のニューラルネットをベースとした自然言語処理においてはより語彙数を抑えるために単語よりも短い単位であるトークン単位に分割する Sentencepiece[2] などのトークナイザが用いられる．特に入出力が共にテキストであるようなシステムでは，内部的に扱うトークンが単語である必要はなく，コーパスから情報理論的に適した単位に分割する方が都合が良い．しかしながら，対話システムやテキスト読み上げでは，単純にトークンのみではなく，その発音や読みの情報が必要となる．こうしたタスクにおいては従来の形態素解析を用いた単語の読み情報が必要となる．例えば OpenJTalk[2)] では内部的に辞書にアクセント情報を追加した MeCab が用いられており，解析結果の読みを利用している．

読みを推定する形態素解析手法は教師あり学習を用いており，学習には単語境界に加えて正しい読みの情報を付与したコーパスが必要となる．しかしながら，人手による読み情報の付与はコストが大きく，日本語においては京大コーパスや BCCWJ コーパスのコアデータなどの書き言葉を対象とした一部の言語資源に限られる．正しい読みの付与されたコーパスが入手しにくい一方で，読みの付与された単語辞書は多数存在している．一例として，岩波国語辞典大五版コーパスには約 5 万 6 千の辞書項目が含まれており，見出しとして各語のかな表記が与えられている．また，オープンな辞書としては ChaSen や MeCab で利用されている IPA 辞書や，それに多数の新語を加えた mecab-ipa-NEologd[3] がある．同様に，各種アノテーションの行われていない生コーパスであれば，有志によってウェブ上の文書を収集している Common Crawl[3)] のスナップショットや，それをベースにクリーニングを行った各種コーパスが公開されており，大規模なコーパスが利用可能である．

そこで，本研究では上記の入手可能な言語資源を用いて，教師なしで読みと単語分割の学習を行う手法を提案する．具体的には，生コーパスと読みの付与された単語辞書を用いて，読みを潜在変数とした隠れセミマルコフモデルの教師なし学習を行い，読みと単語分割の同時確率を推定する．

2　先行研究

本節では読みの推定を含めた解析を行う手法として教師あり学習に基づく形態素解析を，また生コーパスから単語分割を行う手法として教師なし形態素解析を概観する．

2.1　教師あり形態素解析

教師あり学習に基づく形態素解析は，点予測に基づく手法と構造学習に基づく手法に大別できる．MeCab は京大コーパスを用いて条件付き確率場（CRF）に基づく大域的最適化を行っており，単語分割精度と品詞推定精度を達成しており，日本語

1） https://taku910.github.io/mecab/
2） https://open-jtalk.sp.nitech.ac.jp
3） https://commoncrawl.org

の自然言語処理で広く用いられている．一方で，読みについてはあくまで辞書に含まれている情報を提示するように作られており，読みの推定を目的として作られてはいない．Sudachi[4] は内部のコスト計算に MeCab を利用しており，同様に CRF に基づく．ただし，ビジネス用途を目的としており，形態素解析の際のトークンの粒度の切り替えや辞書の拡充などを行っている．しかし，こちらも MeCab 同様に読みの推定を目的としていない．KyTea は LIBLINEAR を用いた点予測に基づく手法で，部分的にアノテーションされたコーパスからの学習を可能としている．品詞推定や読み推定を単語分割と分けた多段階処理として行っており，BCCWJ コーパス [5] を用いた単語分割，品詞推定精度では CRF とほぼ同等の精度を達成している [1]．また，読み推定については，単語がコーパスに出現する場合はコーパス中に存在する読みの候補の分類問題として解き，コーパスに出現しない場合は辞書の最初の読みを出力，単語が辞書に含まれない場合は文字毎に読みを推定するという処理で行っている．分類の特徴量として前後の文字 n-gram および文字種 n-gram を利用しており，BCCWJ コーパスを用いた実験では 98 ポイントを超える F 値を達成している [6]．

教師あり学習に基づく形態素解析手法はいずれも高い精度で単語分割，品詞推定を行うことができる．構造学習を用いた手法では読みの推定を目的としていないものの，原理的には特徴量として単語の読みや連接などを加えることができるが，いずれにしても正しい読みが付与された教師データが必要となる．

2.2 教師なし形態素解析

形態素解析の教師なし学習には大きくはヒューリスティックに基づく手法とベイズに基づく手法，およびニューラルネットに基づく手法がある．ヒューリスティックに基づく手法では，最小記述超原理に基づき記述超をを小さくするように文字列を連接することで単語を獲得する [7][8]．これに対し，ベイズに基づく手法では単語 n-gram に基づく単語分割確率を定義し，確率的に単語分割をサンプルすることで単語の獲得と n-gram 言語モデルの更新を行う [9]．ベイズに基づく手法では適切な事前分布を与えることで，MDL に基づく手法と比較しても高い単語分割精度を達成できる．また，単語分割だけではなく潜在変数として品詞を導入し，単語と品詞の同時学習を行うことでより高い単語分割精度を達成できることも報告されている [10]．また，近年ではニューラル言語処理での利用を前提として，語彙数に制約を加えた手法も提案されている．基本的なアイデアは MDL やエントロピーに基づく手法と同様であるが，これらの手法ではニューラル言語モデルで扱う語彙数を減らし，入出力の系列超を文字ベースほど長くすることなく各時刻での単語予測に掛かる計算量を軽減することに貢献している [11][2]．また，ニューラルネットを用いた教師なし単語分割手法も提案されている [12]．ニューラルネットを用いた手法では確率的に単語分割をサンプリングするのではなく，隠れセミマルコフモデルと同様に入力系列に対する単語分割の周辺確率を計算し，それを直接最大化するよう学習している．

教師なし学習に基づく形態素解析では，辞書を用いないことから単語分割を行うことや潜在変数として品詞クラスを推定することはできるが，読みの推定を行うことは難しい．

3 提案手法

本節では読みの推定を目的として，辞書を用いた教師なし形態素解析手法を提案する．提案手法の生成モデルを (1) に示す．

$$
\begin{aligned}
p(c) &= \sum_{W,Y} P(W,Y) \\
&= \sum_{W} \sum_{Y} \prod_{i}^{n} P(w_i|y_i) P(y_i|y_{i-1}) \qquad (1)
\end{aligned}
$$

c は観測文字列 $c = \{c_0, c_1, \ldots, c_T\}$ を，W，Y はそれぞれ単語列 $W = \{w_0, w_1, \ldots, w_n\}$ とその読み $Y = \{y_0, y_1, \ldots, y_n\}$ を表す．観測文字列から生起し得る単語列とその読みについて周辺化したものを観測文字列の確率としており，これは読みを潜在変数とした隠れマルコフモデルと見做せる．

遷移確率　2 つの単語 w_i，w_j とその読み $y_i = \{p_0^i, p_1^i, \ldots, p_n^i\}$，$y_j = \{p_0^j, p_1^j, \ldots, p_n^j\}$ が与えられた時，遷移確率は y_i，y_j を連結した系列 $y = \{p_0, p_1, \ldots, p_{n^i+n^j}\}$ に対する n-gram 確率の積 (2) で計算できる．

$$
\begin{aligned}
P(y_j|y_i) &\propto P(y) \\
&= \prod_{t} p(p_t|p_{t-l+1:t-1}) \qquad (2)
\end{aligned}
$$

l は読みの n-gram の次数を表す．

ただし，単語には複数の読みの候補が存在するため，このままでは長い読みの確率が小さくなる．そ

Algorithm 1 学習アルゴリズム

Require: $c \in D$

Init $W, Y \sim Uniform$
Add $\forall w, y \in S, N$ to Θ
for $j = 1 \cdots J$ **do**
  for w, y in randperm W, Y **do**
    Remove customers of w, y from Θ
    Draw w, y according to (4)
    Add customers of w, y to Θ
  end for
  Sample hyperparameters of Θ
end for

こで，実際には長さで幾何平均を取ることで読みの長さの影響が出ないようにしている．

遷移確率の n-gram 確率 $p(p_t|p_{t-l+1:t-1})$ には，階層 Pitman-Yor 言語モデル（HPYLM）[13] を用いた．

出力確率　出力確率は読みが与えられた時の単語の条件付き確率 $P(w|y)$ は直接計算することは難しい．そのため，提案手法では単語と読みの確率変数を独立と見做し，近似することで計算する．

$$P(w|y) \propto P(w, y) = P(w)P(y) \tag{3}$$

$P(y)$ の計算には遷移確率同様 HPYLM を，$P(w)$ の計算には単語 unigram 確率の事前分布に文字 n-gram の HPYLM を導入した Nested Pitman-Yor 言語モデル（NPYLM）[9] を用いた．

単語と読みのサンプリング　単語と読みのサンプリングには，blocked Gibbs sampling を用いる．即ち各時刻における単語と読みの同時確率を前向きアルゴリズムによって求め，前向き確率に従って後ろ向きに単語と読みをサンプリングする．時刻 t で長さ k の単語 $c_{t-k+1:t}$ と読み y が現れる前向き確率を (4) に示す．

$$P(c_{t-k+1:t}, y) = \sum_j \sum_z P(c_{t-k+1:t}|y)P(y|z)P(c_{t-k-j+1:t-k}, z) \tag{4}$$

4　学習アルゴリズム

提案手法の学習アルゴリズムを 1 に示す．式 (4) で単語と読みの同時確率を求めていることから分かるように，提案法では単語と読みを同時にサンプリングする．サンプリングされた単語とその読みの系列を用いて，遷移確率計算で用いる読み n-gram HPYLM と出力確率で用いる NPYLM の更新を行う．遷移確率の HPYLM の次数は 5 とした．単語は unigram としているが，事前分布として用いる文字 n-gram HPYLM の次数は 10 とした．

表 1　使用した BCCWJ のデータ

コーパス	訓練	評価
OC	5,904	500
OW	5,563	504
OY	7,058	509
PB	9,582	511
PM	12,543	495
PN	57,281	505

表 2　読み推定の性能評価

	HSMM			KyTea		
	Pre.	Rec.	F1	Pre.	Rec.	F1
OC	87.19	91.46	89.27	99.03	99.16	99.09
OW	89.43	91.54	90.47	99.57	99.56	99.57
OY	82.40	85.89	84.11	99.09	98.78	98.93
PB	90.71	92.69	91.68	99.13	99.08	99.10
PM	89.08	91.73	90.39	98.76	98.71	98.73
PN	86.49	87.93	87.21	99.43	99.43	99.43

5　評価実験

テキストの読み推定を教師なし学習で行えるかを検証するため，BCCWJ のコアデータを用いた評価を行う．実験では，各ドメインのコーパスから各種アノテーションを除去し生テキストとしたものを用意し，単語分割と読みの教師なし学習を行なった．使用したデータのサイズを表 1 に示す．実験には Unidic と KyTea で利用されている数字辞書，及び未知語処理用に単漢字辞書を利用した。評価尺度には，Neubig ら [6] と同様に，最長共通部分列に基づく精度，再現率，F 値を用いた．評価結果を表 2 に示す．また，既存手法では教師なし学習による読み推定手法がないため，リファレンスとして KyTea による教師あり学習を行った場合の数値を載せる．また，提案手法は教師なし学習手法であるが，教師データがある場合はそれをサンプリング結果として扱うことで教師あり学習も可能である．そこで，表 1 に示す BCCWJ のコアデータを教師データとして利用し，BCCWJ のサブデータから無作為に抽出した 10 万文を生コーパスとして用いる半教師あり学習も行った．半教師あり学習の評価結果を表 3 に示す．表 2 より，教師なし学習手法でもドメインによっては F 値 90 程度と比較的高い精度で読みの推定が行えていることが分かる．一方で，教師あり学

表 3 半教師あり学習の評価

	教師なし学習			半教師あり学習		
	Pre.	Rec.	F1	Pre.	Rec.	F1
OC	87.19	91.46	89.27	87.64	91.89	89.72
OW	89.43	91.54	90.47	88.61	91.05	89.82
OY	82.40	85.89	84.11	82.38	85.98	84.14
PB	90.71	92.69	91.68	90.51	92.50	91.50
PM	89.08	91.73	90.39	88.79	91.57	90.16
PN	86.49	87.93	87.21	86.66	88.19	87.41

習手法である KyTea と比較すると精度，再現率ともに大きく下回っている．KyTea と同様の教師データを用いて，さらに BCCWJ サブデータから生コーパスを追加した半教師あり学習の実験結果（表 3）でも，ほとんど教師なし学習と変わらない結果となった．

6 エラー分析

高頻度の誤りの例を表 4 に示す．誤りの大半は空白や記号，英数字，接尾辞が占めている．記号や空白については，提案手法では記号辞書に含まれる場合は読みを出さないように実装している．また，アラビア数字やアルファベットについては正解が表記と同じであるが，それらの読みは未知語処理で扱う単漢字辞書にのみ含まれていた．提案手法ではラティスの構築時に既知語辞書を用いて辞書引きが行えた単語については未知語辞書を引いていないため，英数字については表記をそのまま読みとするノードがラティスに含まれておらず，正解を出力することができていない．これらの問題は既知語辞書に全角スペースを含めたり，記号辞書に読みを与えることで解決可能と考えられる．

接尾辞や単漢字を正しく読むためには，文脈によって判断する必要がある．提案手法では遷移確率と出力確率の 2 つを用いて単語分割と読みの同時確率を計算しており，遷移確率で文脈を考慮した計算を行っている．今回の実験では，遷移確率の HPYLM の unigram の事前分布として一様分布を用いており，読みで用いられる文字数を 256 と仮定した．一方で，出力確率の NPYLM では単語 unigram の事前分布を文字 10-gram の HPYLM から求めている．文字 unigram の事前分布は遷移確率と同様に一様分布とし，その文字数を 5000 とした．そのため，読みの n-gram 確率は出力確率と比較して低頻度の場合に大きくなっており，遷移確率が本来よりも高くなったと考えられる．また，出力確率に NPYLM を用いているため，読みが与えられた時の単語の条件付き確率を直接計算することができない．そこで，条件付確率を単語と読みの同時確率で近似し，さらに単語と読みの間に独立性を仮定している．よって，本来存在する単語ごとの読みの出やすさが扱えていない．これらの近似や平滑化のため，本来であれば確率が小さくなるような文脈における単語の読みの確率が高くなっていると予想できる．この問題を解決するためには，遷移確率のハイパーパラメータの調整や，出力確率に読みが与えられた時の単語の条件付き確率を直接モデル化することが可能なニューラルネットを用いることが考えられる．

表 4 読みを誤った単語の例

単語	正解の読み	予測した読み	タイプ	頻度
全角スペース	_		空白	1136
,	、		記号	355
等	とう	ひとし	接尾辞	180
1	1	ひと	英数字	135
3	3	すりー	英数字	127
2	2	にい	英数字	125
者	しゃ	もの	接尾辞	106
円	えん	まどか	接尾辞	86
-	ー		記号	76
5	5	ふぁいぶ	英数字	74
行	おこな	ぎょう	単漢字	70
4	4	ふぉー	英数字	70
〓	げた		記号	69
■	しかく		記号	68
万	まん	よろず	接尾辞	63
二	に	にい	数字	61
年	ねん	とし	接尾辞	61
五	ご	いつ	数字	60
%	%	ぱーせんと	記号	59

7 おわりに

本稿では，辞書と生コーパスから単語分割と読みの教師なし学習を行う手法の提案を行った．教師なしで読み推定を最適化するよう単語分割を学習する初の試みであり，BCCWJ コーパスを用いた実験によって，F 値 90 程度の読み推定が行えることを示した．単語の出力確率に単語と読みの独立性を仮定した近似を入れていることは課題であり，今後は出力確率をニューラルネットに置き換えることなどが考えられる．

参考文献

[1] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In **ACL 2011**, pp. 529–533, 2011.

[2] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In **ACL 2018**, pp. 66–75, 2018.

[3] 奥村学佐藤敏紀. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017), pp. NLP2017–B6–1. 言語処理学会, 2017.

[4] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In **LREC 2018**, 2018.

[5] Makoto Yamazaki Toshinobu Ogiso Takehiko Maruyama Hideki Ogura Wakako Kashino Hanae Koiso Masaya Yamaguchi Makiro Tanaka Maekawa, Kikuo and Yasuharu Den. **Balanced corpus of contemporary written Japanese**. Language Resources and Evaluation 48, 2014.

[6] Graham Neubig and Shinsuke Mori. Word-based partial annotation for efficient corpus construction. In **LREC 2010**, 2010.

[7] Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In Hang Li and Lluís Màrquez, editors, **EMNLP 2010**, pp. 832–842, 2010.

[8] Pierre Magistry, Benoît Sagot, et al. Can mdl improve unsupervised chinese word segmentation? In **SIGHAN Workshop on Chinese Language Processing**, pp. 2–10, 2013.

[9] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In **ACL-IJCNLP 2009**, pp. 100–108, 2009.

[10] Kei Uchiumi, Hiroshi Tsukahara, and Daichi Mochihashi. Inducing word and part-of-speech with pitman-yor hidden semi-markov models. In **ACL-IJCNLP-2015**, pp. 1774–1782, 2015.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In **ACL 2016**, pp. 1715–1725, 2016.

[12] Kazuya Kawakami, Chris Dyer, and Phil Blunsom. Learning to discover, ground and use words with segmental neural language models. In **ACL 2019**, pp. 6429–6441, 2019.

[13] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In **COLING-ACL-2006**, pp. 985–992, 2006.