知能と情報（日本知能情報ファジィ学会誌）Vol.25, No.6, pp.174－183（2013）

解説

日本語形態素解析とその周辺領域における最近の研究動向†

鍛治 伸裕 *

1．はじめに

形態素解析とは，テキストを単語に分割し，各単語に品詞を割り当てる処理のことである[60, 62]．形態素解析は，日本語や中国語など，分かち書きの習慣がない言語で記述されたテキストを計算処理するためには必要不可欠な技術であり，これまで盛んに研究が行われてきた．

日本語形態素解析の研究は，コスト最小法などの規則ベースの手法に端を発し，主にそれを確率モデル化するという方向で発展を遂げてきた[3, 26, 27, 33, 35, 46]．現在では，ラベル付きコーパスを用いる教師有り学習に基づくアプローチが主流となっており，単語単位の適合率と再現率が共に95%を越えるという，高い精度での解析が実現されている[26]．一方，中国語などの他言語においても，同様に教師有り学習に基づくアプローチが広く用いられている[19, 25, 42, 43, 44, 51, 52]．こうした研究成果の一部は，ソフトウェアとして公開されており，自然言語処理を始めとする多くの研究活動を下支えしている[62]．

このように，形態素解析は成熟した技術であると言えるが，依然として課題も残されている．なかでも，従来の形態素解析モデルが**未知語**（ラベル付きコーパスにも辞書にも出現しない単語）をうまく扱えないという問題[9]は，以前から研究者によって指摘されてきたことであり，未知語に対して頑健な解析モデルを構築することは，現在，形態素解析の研究における主要な目標となっている．

それと同時に，近年，ウェブの拡大と普及により，未知語に頑健な形態素解析の実現に対する要求が急速に高まりつつある．ウェブテキスト上では，多種多様な話題に関する言及が行われるため，辞書に登録されていない固有名詞や新語など，未知語が頻繁に使われる．そのため，自然言語処理をウェブマイニングなどに応用する場合には，未知語を高い精度で解析できる形態素解析技術が重要となる．

このような背景を受けて，この数年の間，形態素解析に関する技術は大きな進歩を遂げつつある．このことを踏まえて，本稿では，日本語形態素解析とその周辺領域における研究成果から，未知語の扱いに関する最近の取り組みを紹介する．ただし，最新の研究動向を伝えることに主眼を置いて話を進め，入門的な解説は省略をする．形態素解析技術に関する基本知識に関しては，教科書[60, 62]やウェブに公開されている資料[55, 56]などを参照されたい．

本稿の構成は以下のようになっている．まず，2節では，導入として，日本語形態素解析とその周辺課題を整理する．3節では，ラベル無しコーパスから未知語の解析に有効な情報を学習する，半教師有り学習に基づく形態素解析手法を紹介する．次に，4節と5節では，未知語の生成過程を考慮することによって，未知語に頑健な解析処理を実現するアプローチを紹介する．最後に6節では，まとめを行うとともに，本稿では詳しく取り上げることができなかった話題を概観する．

2．日本語形態素解析とその周辺

まず始めに本稿が扱う対象を明確にするため，日本語形態素解析というタスクと，その周辺にある研究課題について整理を行う．

形態素解析というタスクはそもそも厳密に定義することが難しいが，本稿では，テキストを単語に分割する処理（単語分割）と，各単語に適切な品詞タグを割り当てる処理（品詞タグ付与）の2つをまとめたものを形態素解析と呼ぶ．分割された単位のことを単語と呼ぶのか，それとも形態素と呼ぶのかなど，上記の定義には議論の余地が残されているが[60, 62]，本稿の主旨から外れるため，これ以上の深い議論は行わない．

日本語形態素解析の研究においては，単語分割と品詞タグ付与を同時に解くモデルが広く用いられている

† Recent Research Trends in Japanese Morphological Analysis and Its Related Areas

Nobuhiro KAJI
* 東京大学 生産技術研究所
Institute of Industrial Science, The University of Tokyo

[26]．しかし，2つのタスクは必ずしも同時に解く必要はなく，実際，それらを順番に解くような方法も提案されている[37]．そのため，本稿では，単語分割と品詞タグ付与という一連の処理のことを，同時に解く解かないに関わらず形態素解析と呼ぶ．

以下では，日本語単語分割や中国語形態素解析など，日本語形態素解析と関わりが深い周辺領域を言語ごとに分類して概観する．

2.1 日本語

日本語における形態素解析の周辺研究としては，その部分問題である単語分割がある[13, 30]．単語分割に関する研究成果は，単語分割と品詞タグ付与を順次行うような形態素解析モデルを前提とすれば，そのまま形態素解析に応用することができる．

さらに，単語分割の中でも，特に複合名詞に焦点を当てた研究が存在する[2, 22, 36, 59]．このような特殊なタスクが設定されている背景には，複合名詞は従来の単語分割モデルによる解析が困難であるため，そこに焦点をあてて問題解決が試みられてきたという経緯がある．複合名詞の単語分割が困難な理由としては，ドメイン依存の用語(domain terms)など，未知語が多く出現するということが指摘できる[2, 22]．これに加えて，複合名詞の単語分割には品詞情報が効きにくいため，単語分割と品詞タグ付与を同時に行って分割精度を向上させるというアプローチの効果が薄いことも，複合名詞の単語分割が困難な理由と言える[2, 22, 36, 59]．

従来，複合名詞の単語分割に関する研究は，一般的な単語分割や形態素解析とは独立に進められてきた．しかし，最近では，複合名詞の単語分割に関する研究成果を取り込んだ形態素解析モデルが提案されるなど[13]，両者には融合の兆しが見られる．

2.2 中国語

日本語と同様，中国語もテキストを分かち書きする習慣がないため，形態素解析や単語分割に関する研究が盛んに行われている[19, 25, 42, 43, 44, 51, 52]．提案されている解析モデルは，大半が日本語にも適用可能なものであり，日本語と中国語で方法論に大きな差はないと言えるだろう．実際，日本語と中国語の両言語で実験を行っているような研究も見られる[13, 30, 34]．

中国語の場合，形態素解析と係り受け解析は，完全に独立したタスクではなく，一方の解析結果がもう一方の解析結果に大きな影響を与える．そのため，形態素解析と係り受け解析を同時に行うモデルが提案されており，精度の向上が報告されている[15, 39]．日本語における類似の試みとしては，複合名詞の単語分割と係り受け解析を同時に行うモデルが提案されている[59]．

形態素解析と係り受け解析の同時解析が有効であるのは，孤立語という中国語の特徴に依存する部分が大きいと考えられる．膠着語である日本語の場合は，線形連鎖モデルやセミマルコフモデルでも，助詞や助動詞などを手がかりとして十分に活用できる．そのため，長い複合名詞などの特別な場合を除いて，係り受け構造を考慮する利点は小さいと考えられる．

2.3 英語

英語の場合，テキストは分かち書きされるため，単語分割を行う必要はなく，品詞タグ付与のみが研究対象となる．英語において単語分割と言った場合には，テキストではなく，音素列を単語に分割するという別のタスクを指すことが多い[4, 12]．

このように，英語における単語分割と品詞タグ付与は，タスク設定や位置付けが日本語とは大きく異なっている．しかし，それらは本質的には類似したタスクであるため，独立して研究が行われているわけではなく，提案されている解析モデルには関連性が見られる．例えば，Mochihashiら[30]の提案する単語分割モデルは，英語の単語分割のために提案されたモデル[12]を拡張したものであるが，英語だけでなく日本語と中国語にも適用されている．

2.4 独語

独語は，英語と同様，テキストを分かち書きする言語であるが，複合名詞だけは例外的に分かち書きが行われない．そのため，独語においても複合名詞の単語分割に関する研究が行われており[5, 24]，日本語における研究状況との間に類似性が見られる．

3．半教師有り学習

1節でも触れたように，現在の形態素解析は，ラベル付きコーパスから統計モデルを学習するという，教師有り学習に基づくアプローチが主流となっている．そうした枠組みにおいて未知語の数を削減するためには，より大規模なラベル付きコーパスを用意しなくてはならない．しかし，ラベル付きコーパスは手作業で作成する必要があるため，大規模化することは現実問題として容易ではない．

こうした問題意識から，ラベル付きコーパスだけでなくラベル無しコーパスも学習に利用するという，半教師有り学習に基づく形態素解析の研究が進められて

いる．ラベル無しコーパスは，ラベル付きコーパスと比べてはるかに大量に入手可能である．そうした大規模なラベル無しコーパスから，未知語の解析に有効な情報をうまく取り出すことによって，未知語に頑健なモデルを学習することが，半教師有り学習に基づく形態素解析の狙いである．

3.1 素性駆動型自己学習

近年，自然言語処理においては，自己学習（self-training）[1]の亜種とでも言うべき半教師有り学習アルゴリズムが，単語分割や形態素解析などのタスクにおいて成功を収めている[7, 38, 47, 48, 61]．これは，ラベル付きコーパスの代わりに，素性集合を拡張する方法であるため，ここでは**素性駆動型自己学習**と呼ぶ．

まず始めに，普通の自己学習について簡単に説明を行う．一般的な自己学習の手続きは以下の通りである（アルゴリズム 1）[1]．まず，入力として，ラベル付きコーパス$\mathcal{L}$，ラベル無しコーパス$\mathcal{U}$，特徴量抽出に用いる素性関数の集合fが与えられる．そして，ラベル付きコーパス$\mathcal{L}$と素性集合fを用いてモデルmを学習する（1 行目）．次に，そのモデルを用いてラベル無しコーパス$\mathcal{U}$を解析し，その結果から信頼度の高い部分だけを選択することによって，新たなラベル付きコーパス$\mathcal{U}'$を作成する（2 行目）．そして，最後に，$\mathcal{L}$と$\mathcal{U}'$の両方を用いてモデルを学習する（3 行目）．

自己学習は有名なアルゴリズムであるが，少なくとも自然言語処理の分野においては，いくつかの例外的な場合[16, 29]を除いて，有効性が低いことが経験的に知られている[8]．実際，文献[48]では，自己学習を中国語の形態素解析に適用した結果，効果がなかったことが報告されている．

次に，素性駆動型自己学習の手続きをアルゴリズム 2 に示す[38, 48, 61]．自己学習との違いは，ラベル無しコーパスの解析結果から新たな素性集合f'を導出し，最終的には，ラベル付きコーパス$\mathcal{L}$と，拡張された素性集合$f \cup f'$を用いてモデル学習を行う点である．f'の例としては，解析済みのラベル無しコーパスから抽出された，文字列の分割パターン[48]や単語リスト[38, 48, 61]に基づく素性が提案されている．

Algorithm 1 自己学習 [1].

1: $m \leftarrow \textsc{Train}(\mathcal{L}, f)$
2: $\mathcal{U}' \leftarrow \textsc{Select}(\textsc{Test}(m, \mathcal{U}))$
3: **return** $\textsc{Train}(\mathcal{L} \cup \mathcal{U}', f)$

Algorithm 2 素性駆動型自己学習.

1: $m \leftarrow \textsc{Train}(\mathcal{L}, f)$
2: $f' \leftarrow \textsc{Induce}(\textsc{Test}(m, \mathcal{U}))$
3: **return** $\textsc{Train}(\mathcal{L}, f \cup f')$

素性駆動型自己学習において重要なのは，大規模なラベル無しコーパスを利用して，新しい素性を導出している点である．これによって，ラベル付きコーパスに出現する単語と出現しない単語の間で，より多くの素性が共有されるようになり，モデルの汎化能力が高められていると考えられる．ラベル無しコーパスを素性導出に利用するというタイプの半教師有り学習手法は，ここで紹介した素性駆動型自己学習の他にも多く提案されており，単語分割や形態素解析における適用事例としては，風間ら[63]，Sunら[44]，持橋ら[58]，Zengら[49]などの研究がある．

3.2 自然注釈

一方，アルゴリズム上の工夫を行うのではなく，タスクに固有のヒューリスティクスを駆使することによって，ラベル無しコーパスを学習に利用するアプローチも提案されている．そうした方法は，一般的には半教師有り学習と呼ばないのかもしれないが，ラベル付きコーパスとラベル無しコーパスの両方を使ってモデル学習を行っていることには変わりないため，本稿では半教師有り学習の 1 つとして扱う．

日本語や中国語のように分かち書きを行う習慣のない言語においても，句読点やマークアップを擬似的な単語区切りとみなせば，ラベル無しコーパスから，部分的に単語境界の情報が付与されたコーパスを作成することができる[21, 50]．例えば，テキストが下記のようにマークアップされていれば「形」と「解」の直前に単語境界が存在すると考えることができる．

頑健な＜a＞形態素＜/a＞解析を行う

このようなマークアップなどの擬似的な注釈情報のことを**自然注釈**（natural annoation）と呼ぶ[21]．

自然注釈付きコーパスからは，通常のラベル付きコーパスと違って，完全な単語分割結果を得ることはできない．例えば，上記の文の場合であれば「態」の直前に単語境界が存在しないことや「を」の直前に単語境界が存在することは分からない．そのため，このような不完全な情報をどうモデル学習に利用するかが問題となる．

Jiangら[21]は，ウィキペディアのマークアップを基にして390万文の自然注釈付きコーパスを作成し，それを用いて自己学習を行うことによって，単語分割モデルの精度を向上させることに成功している．自然注釈付きコーパスが与えられたとき，自然注釈に違反しないように制約を加えながら解析を行えば，普通に解析を行うよりも精度の高い分割結果が得られると考えられる．Jiangらの提案は，このような直感に基づ

Algorithm 3 自然注釈に基づく自己学習 [21].

```
1: m ← Train(L)
2: L' ← ∅
3: for x ∈ U do
4:     y ← Test(m, x)
5:     ỹ ← ConstrainedTest(m, x)
6:     if y ≠ ỹ then
7:         L' ← L' ∪ (x, ỹ)
8:     end if
9: end for
10: return Train(L ∪ L')
```

いて，アルゴリズム 1 における Select 関数を設計するというものである．

Jiangらの提案する学習方法をアルゴリズム 3 に示す（アルゴリズム 1 とは若干表記方法が異なる）．まず，ラベル付きコーパス $\mathcal{L}$ を用いてモデル m を学習する（1 行目）．そして，自然注釈付きコーパス $\mathcal{U}$ の各文 x に対して，普通に解析した結果 y と，自然注釈に違反しないように制約を加えて解析した結果 $\tilde{y}$ を取得する（3, 4 行目）．このとき，$y \neq \tilde{y}$ であれば，$\tilde{y}$ を正解だとみなして新たなラベル付きコーパス $\mathcal{L}'$ を作成する（5 から 7 行目）．最後に，$\mathcal{L} \cup \mathcal{L}'$ から最終的なモデルを学習する（8 行目）．

3.3 議論

本節では，半教師有り学習に基づく形態素解析の研究として，素性駆動型自己学習と自然注釈を紹介した．紹介した手法は，2 つとも自己学習に基づくものであるが，自己学習という枠組み自体が本質的に重要なわけではない．それぞれ，ラベル無しコーパスからの素性導出，単語分割というタスクに固有のヒューリスティクスの利用，という点が精度向上に寄与している本質的な要因と考えられる．

従来，形態素解析におけるラベル無しコーパスの利用と言えば，未知語の抽出が主流であった [31, 32, 65]．抽出した未知語のリストは，教師有り学習に基づくアプローチを前提とした場合，素性導出に使うのが一般的であると考えられる．そのため，現在では，未知語抽出もラベル無しコーパスから素性導出を行う方法の 1 つと位置付けることができるであろう．

自然注釈は，タスクに固有のヒューリスティクスに基づく手法であるため，ナイーブな印象を受けるかもしれない．しかし，実用的には大きな効果が期待できることから，手段にこだわることなく，こうした方向性も今後大いに研究されるべきであろう．例えば，Tsuboiら [45] の提案するようなアルゴリズムを使って，自然注釈付きコーパスからモデル学習を行うなど，様々な展開が考えられる．

4．異表記のモデル化

従来，学習時に観測されない単語は，全て未知語として一括りに扱われてきた．しかし，未知語は，固有名詞のように完全に新規な単語と，既知語の異表記として捉えるのが自然なものに分けて考えることができる [57]．近年では，後者のような未知語を扱うために，異表記の生成過程を考慮した手法が提案されている．

異表記とは，例えば「コンピューター」と「コンピュータ」のように，同一の単語に対する異なる文字列表記のことを指す．よく使われる異表記は，ほとんどが辞書に登録されているため，これまで異表記が未知語として問題になることは稀であった．しかし，近年，くだけたウェブテキストの普及によって，小書き（例：おいしぃ），長音化（例：すごーい），過剰なひらがな化（例：らーめん）など，極めて多様な異表記を扱うことが必要となっている．しかし，そうした異表記の多くは未知語であり，解析に失敗してしまうことが問題となっている [41, 57]．

異表記の大半は，単純な編集操作によって，辞書に登録されている正規形から自動生成することができる [41, 64]．このことに着目し，単語の正規形から出現形（正規形と異表記が混在したもの）が生成される過程を確率モデルの枠組みで捉えようとする試みや，前処理によって辞書を自動拡張する試みが行われている．

4.1 拡張品詞 n-gram

筆者の知る限り，形態素解析において最初に異表記のモデル化を行ったのは風間ら [64] である．彼らは，品詞2-gramモデル [3] を拡張することによって，単語の出現形と正規形を同時に生成する確率モデルを提案している（**拡張品詞2-gram**と呼ぶ）．拡張品詞2-gramにおいては，単語の出現形 $\mathbf{w} = (w_1, w_2, \ldots w_n)$，正規形 $\mathbf{v} = (v_1, v_2, \ldots v_n)$，品詞タグ $\mathbf{t} = (t_1, t_2, \ldots t_n)$ の同時生成確率が以下のように定義される．

$$p(\mathbf{w}, \mathbf{v}, \mathbf{t}) = \prod_{i=1}^{n} p(t_i|t_{i-1}) p(v_i|t_i) p(w_i|v_i) \quad (1)$$

これと同一の生成モデルは，工藤ら [57] によっても独立に提案されている．また，英語の単語分割においても，出現形と正規形を生成する確率モデルが提案されている [4]．

拡張品詞2-gramを用いて形態素解析を行うには，正規形 $\mathbf{v}$ を消去した確率 $p(\mathbf{w}, \mathbf{t}) = \sum_{\mathbf{v}} p(\mathbf{w}, \mathbf{v}, \mathbf{t})$ を考えて，入力文 x に対して確率最大となる $(\mathbf{w}, \mathbf{t})$ を求めれば良い．

$$(\mathbf{w},\mathbf{t}) = \underset{(\mathbf{w},\mathbf{t})\in\mathcal{G}(x)}{\operatorname{argmax}}\ p(\mathbf{w},\mathbf{t}) \quad (2)$$

もしくは，(**w**, **v**, **t**)の確率を最大化すれば，形態素解析と同時にテキスト正規化(text normalization)を行うこともできる．

$$(\mathbf{w},\mathbf{v},\mathbf{t}) = \underset{(\mathbf{w},\mathbf{v},\mathbf{t})\in\mathcal{G}'(x)}{\operatorname{argmax}}\ p(\mathbf{w},\mathbf{v},\mathbf{t}) \quad (3)$$

上記の式において，$\mathcal{G}(x)$と$\mathcal{G}'(x)$は，入力文xに対して考えられる全ての(**w**, **t**)および(**w**, **v**, **t**)の集合を生成する関数である．$\mathcal{G}(x)$と$\mathcal{G}'(x)$の与え方に関する詳細な議論は見られないが，異表記を考慮しながら辞書引きを行うことによって，候補を生成しているものと推測できる(文献[64]の3節などを参照)．

拡張品詞2-gramに基づく形態素解析を実現するうえで技術的に問題となるのは，モデルの学習方法である．既存のラベル付きコーパスを用いる場合，$p(t_i|t_{i-1})$と$p(v_i|t_i)$は容易に推定可能であるが，$p(w_i|v_i)$の推定は難しい．少なくとも現時点において，ウェブのようなくだけたテキストに対して，単語の出現形と正規形をアノテートしているような大規模コーパスは存在していない．

風間らは，文字単位の生成モデルを使って確率$p(w_i|v_i)$を定義し，最終的には人手で確率値の調整を行っている．一方，工藤らはEMアルゴリズムを用いて，ウェブコーパスから確率値$p(t_i|t_{i-1})$，$p(v_i|t_i)$，$p(w_i|v_i)$を直接推定することを提案している．

4.2 辞書拡張

拡張品詞2-gramのような生成モデルに代わる簡便なアプローチとして，前処理によって辞書を拡張する方法が提案されている[41, 53]．この方法では，辞書登録語の異表記を機械的に生成することによって辞書を拡張し，拡張された辞書と既存の形態素解析モデルを用いることによって解析を行う．これを**辞書拡張**と呼ぶ．

このような方法で長音化を扱おうとした場合，任意の数の長音記号が挿入される可能性があることから，あらゆる異表記を事前に全て列挙しておくことは不可能となる．そのため，辞書を拡張するのではなく，テキストを正規化しながら辞書引きを行うという実装が提案されている[41]が，本質的には辞書の拡張を行っているのと同じことである．

辞書拡張においては，拡張品詞2-gramと異なり，ほぼ自明なモデル学習の方法が存在する．すなわち，既存のラベル付きコーパスを用いて，従来の形態素解析モデルと全く同じ方法で学習を行うことができる[53]．

4.3 議論

本節では，形態素解析において異表記をモデル化するためのアプローチとして，拡張品詞n-gramと辞書拡張を紹介した．

拡張品詞n-gramは，多様な異表記をエレガントに扱うことのできる魅力的な枠組みであり，今後の発展が大いに期待できる．しかし，最先端の統計モデル[26, 37]との比較が行われていないなど，有効性が不明確な部分も残されており，これからの研究の進展が待たれる．

一方，辞書拡張は，考え方や実装がシンプルであることが大きな利点となる．拡張品詞2-gramのような，異表記を取り込んだモデルがまだ未成熟であることを考慮すると，形態素解析において異表記を扱うための方法として，現時点ではベストプラクティスと言える．しかし，このような単純な方法で，ひらがな化のような副作用の多い異表記をうまく扱うことができるのか[57]，拡張品詞2-gramと精度にどの位の差が生じるのかなど，解消されていない疑問も多く，引き続いての研究調査が望まれる．

異表記をモデル化するというアプローチは，テキスト正規化[14]やスペル誤り訂正[17]とも深く関連する．これらの研究との関連性についても，今後の研究の中で議論が深まることを期待したい．

5．言語投影

英語からの借用は，日本語において未知語が形成される代表的な要因の1つとなっている[22, 36]．借用語は片仮名を使って表記されることが多いため，片仮名語とも呼ばれる．本節では，この片仮名語という未知語に着目した研究を紹介する．

片仮名語は複合名詞(例：パセリソース，ジャンクフード，ブラキッシュレッド)を形成しやすいことが知られているが，これは従来の単語分割モデルによる解析が困難となっている[22, 36]．その理由として，上記のように片仮名語には未知語が多いことや，2節で説明したように，品詞情報が利用できないことなどが挙げられる．

一方，英語は，日本語と異なり，単語を分かち書きして表記する．このことに着目し，何らかの方法で片仮名複合名詞を英語に変換して，英語と片仮名語の対応関係を利用することによって，片仮名複合名詞の単語分割を行うという手法が提案されている[13, 22, 36]．そうしたアプローチのことを**言語投影**(language projection)と呼ぶ．

表 1 「パセリソース」に対する分割候補，対訳辞書を用いた各候補の英訳，英語コーパスにおける英訳の頻度［36］.

分割候補	対訳辞書による英訳	頻度
パセリ/ソース	parsely sauce	20,600
パセ/リソース	pase resource	3

5.1 対訳資源に基づく手法

Nakazawaら［36］は，対訳辞書を利用して単語分割候補を英語に翻訳し，得られた英語表現の自然さに基づいて適切な単語分割候補を選択する方法を提案している．

以下では具体例を用いてNakazawaら［36］の方法を説明する．説明を簡単にするため，複合名詞「パセリソース」の分割候補として「パセリ／ソース」と「パセ／リソース」の 2 つが与えられたと仮定し，そのどちらか一方を選択するという問題設定を考える．ただし，／は単語境界を表す．

対訳辞書を使って 2 つの分割候補を英語に変換すると「parsely sauce」と「pase resource」という英訳が得られる．そして，大規模な英語コーパスを使ってそれらの出現頻度を調べると，どちらが英語として自然な表現であるかが分かり，その結果として，どちらの分割候補が適切であるのかを判断することができる．この例の場合であれば「parsely sauce」の方が頻度が大きくなるため「パセリ／ソース」が正しい分割結果であると判断できる（表 1 ）．

同様に対訳資源を用いるアプローチとしては，対訳コーパスから単語対応（word alignment）を発見することにより，独語複合名詞の分割規則を学習する試みが報告されている［6, 24］.

5.2 翻字モデルに基づく手法

対訳辞書を使うというアプローチは，高い精度で英訳を得ることができることが利点となる．しかし，対訳辞書は高価な言語資源であるため，対応できる片仮名語が限られてしまうことが問題となる．これと同様のことは，対訳コーパスを用いたアプローチにも当てはまる．このような問題意識から，最近では，翻字モデルに基づく手法が提案されている［13, 22］.

翻字とは文字の置き換えに基づく翻訳のことであり（例：computerに対するコンピュータ），片仮名語は基本的には英語の翻字となっている．また，これとは逆に，片仮名語を元の英語に戻す操作のことを逆翻字と呼ぶ.

英語を片仮名語に翻字したり，片仮名語を英語に逆翻字したりする処理は，機械翻訳などにおいて必要となる．そのため，単語分割とは独立に，翻字処理に関する研究が進められている．例えばJiampojamarnら［18］は，英語 e とその翻字 f の組（例：e =computer，f =コンピュータ）を同時に生成する確率モデルを提案している：

$$\log\ p(e,f) = \sum_{(\bar{e},\bar{f})} \log\ p(\bar{e},\bar{f}) \tag{4}$$

ただし（$\bar{e}$, $\bar{f}$）は対応関係にある文字列（例：$\bar{e}$ =com，$\bar{f}$ =コン）であり，どの文字列が対応関係にあるのかという情報は潜在変数として扱われる．以下ではこのような確率モデルのことを翻字モデルと呼ぶ.

翻字モデルを利用すれば，例えば f に対して同時確率を最大化する e を求めることによって，任意の片仮名語を英語に逆翻字することが可能となる．そこで，Kajiら［22］は，対訳辞書のような高価な言語資源の代わりに，翻字モデルを使うことを提案している．Kajiらの提案する方法では，まずウェブ上の括弧表現［28］から対訳関係にある片仮名語と英語の対を抽出し，翻字モデル［18］に基づいて片仮名語と英語の単語対応を発見することによって，片仮名語内部の単語境界を認識する（表 2 ）．このようにして，分かち書きされた片仮名語のリストを大量に獲得し，これを分割処理に利用する．具体的には，分割候補に含まれる片仮名語 n -gram（n=1, 2）に対して，それが獲得されたリストに出現するか否かという 2 値素性を用いる.

表 2 括弧表現から抽出された英語と片仮名語の例．片仮名語と英語の単語対応に基づいて認識された単語境界は / で表現されている.

片仮名語	英語
ウィキペディア	wikipedia
ランキング	ranking
フライド/コーク	fried coke
タップ/ウォーター/プロジェクト	tap water project

Hagiwaraら［13］も同様のアプローチを提案しているが，いくつかの拡張が行われている．まず，従来の研究［22, 36］のように片仮名複合名詞の単語分割だけを個別に扱うのではなく，言語投影の仕組みを形態素解析モデルの中に組み込んでいる．また，解析時にオンラインで逆翻字を行うことによって，リストを事前に作成するような方法［22］よりも，多くの片仮名語に対応できるよう工夫をしている.

Hagiwaraらの手法は具体的には次のようになっている．まず，通常の形態素解析と同様に単語ラティスを構築する．そして，ラティスの経路探索を行うが，その時，経路上に片仮名語の n -gram（n=1, 2）が出現

した場合，翻字モデルを用いて逆翻字する．そして，英語コーパスを用いて，得られた英語 n -gramの出現確率を計算し，その対数値を素性として用いる．

5.3 議論

本節は，英語からの借用によって生成される未知語に対応するためのアプローチである言語投影を紹介した．言語投影においては，片仮名語の英語訳を取得することが技術課題となるが，対訳辞書を利用する方法と，翻字モデルを用いる方法の２つがこれまでに提案されている．

言語投影の利点は，例えば「パセリ／ソース」という単語の並びは意味が通るが「パセ／リソース」は意味が通らないというような違いを，容易に認識できる点である．同様の認識処理は，既存の教師有り学習でも原理的には実現可能であるが，疎データ問題が発生するため，現実的には実現困難であると考えられる．

もう１つの利点は，単語分割と多義性解消を結合処理として扱うことが可能な点である．5.1節の議論では省略したが「ソース」の語義には曖昧性があるため，その英訳には「sauce」以外に「source」も考えられる．しかし，この曖昧性も英語コーパスでの頻度を求めることによって解消できる．つまり「parsely source」の頻度は「parsely sauce」より小さくなるので，この場合の「ソース」の意味は「sauce」であることが分かる．

言語投影は，対応先として英語以外の言語を考えることもできる．また，特殊な形として，日本語同士の対応関係（言い換え）を利用することもできる[22]．このように，英語以外の言語表現との対応関係を利用することは，今後の研究の方向性として興味深い．

6．おわりに

本稿では，半教師有り学習，異表記のモデル化，言語投影，という３つの話題を取り上げて，形態素解析における最近の研究動向の紹介を行った．いずれも歴史が浅い方法論であるため，評価が定まっていないものも多いが，形態素解析という技術が今も活発に進化を続けている様子を感じ取って頂ければ幸いである．

本稿では，紙面の都合があり，最近の研究を網羅的に紹介することはできなかった．そこで，最後に，今回取り上げることができなかった話題を概観することによって，本稿の結びとしたい．

高速化　一般的に形態素解析は高速な処理として知られているが，これは未知語を考慮しない場合の話である．未知語の可能性を考慮して形態素解析を行う場合は，入力文の長さ（文字数）に対して２乗オーダの計算量が必要となるため，解析速度が大きく低下する．この問題に対しては，探索アルゴリズムの改良や再順位付け（reranking）など，高速化に関する取り組みが行われている[20, 23, 51, 52, 54]．

大域モデル　多くの形態素解析モデルは局所的な素性のみを使用して解析を行っているが，大域的な情報をモデル化することによって，解析精度の向上を実現しようとする研究も存在する[34, 40]．典型的には，同じ表層形の単語には同じ品詞タグが割り当てられやすいなど，一貫性がモデル化されており，未知語の解析に有効であると考えられる．

教師無し形態素解析　未知語に頑健な形態素解析を実現するための枠組みとしては，教師無し学習に基づく形態素解析モデルが提案されている[30]．しかし，教師無し形態素解析は，人間の直感に合う結果を出力することが保証されていないため，今すぐ（半）教師有り学習に基づく既存手法に取って代わる可能性は低いだろう．今後は，半教師有り学習におけるコンポーネントとして利用するなどなど，その応用方法に関する研究が重要になると考えられる[58]．

言語資源の整備　ここまで紹介をしてきた研究は，基本的には解析モデル（または，モデルが用いる素性）の改良に関するものであった．しかし，形態素解析の性能を向上させるという目的を実現させるためには，辞書やラベル付きコーパスといった言語資源を整備することも重要である[10, 11, 37]．どの問題を解析モデルの改良によって解消し，どの問題を言語資源の拡充によって解消するのかという，システム設計に関しても今後の研究において議論されるべきであろう．

謝辞

本稿を執筆する際には，東京大学の吉永直樹氏からは有益なコメントを多数頂きました．また，東京工業大学の笹野遼平氏からは，本稿執筆時には未発表であった論文情報を提供して頂きました．記して感謝致します．

参 考 文 献

[1] Steven Abney. *Semisupervised Learning for Computational Linguistics*. Chapman and Hall/CRC, 2007.

[2] Rie Kubota Ando and Lillian Lee. Mostly-unsupervised statistical segmentation of Japanese Kanji sequences. *Natural Language Engeering*, Vol.9, No.2, pp.127-149, 2003.

[3] Masayuki Asahara and Yuji Matsumoto. Extended

models and tools for high-performance part-of-speech tagger. In *Proceedings of COLING*, pp.21-27, 2000.
[4] Benjamin Börschinger, Mark Johnson, and Katherine Demuth. A joint model of word segmentation and phonological variation for English word-final /t/-deletion. In *Proceedings of ACL*, pp.1508-1516, 2013.
[5] Martin Braschler and Bärbel Ripplinger. How effective is stemming and decompounding for German text retrieval? *Information Retrieval*, Vol.7, pp.291-316, 2004.
[6] Ralf D. Brown. Corpus-driven splitting of compound words. In *Proceedings of TMI*, 2002.
[7] Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto, Yujie Zhang, and Hitoshi Isahara. Dependency parsing with short dependency relations in unlabeled data. In Proceedings of IJCNLP, pp.88-94, 2008.
[8] Stephen Clark, James Curran, and Miles Osborne. Bootstrapping POS-taggers using unlabelled data. In *Proceedings of CoNLL*, pp.49-55, 2003.
[9] Thomas Emerson. The second international Chinese word segmentation bakeoff. In *Proceedings of SIGHAN*, pp.123-133, 2005.
[10] Dan Garrette and Jason Baldridge. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL*, pp.138-147, 2013.
[11] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffery Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of ACL (Short papers)*, pp.42-47, 2011.
[12] Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. Contextual dependencies in unsupervised word segmentation. In *Proceedings of COLING-ACL*, pp.673-680, 2006.
[13] Makoto Hagiwara and Satoshi Sekine. Accurate word segmentation using transliteration and language model projection. In *Proceedings of ACL (Short Papers)*, pp.183-189, 2013.
[14] Bo Han and Timothy Baldwin. Lexical normalization of short text messages: Makin sens a #twitter. In *Proceedings of ACL*, pp.368-378, 2011.
[15] Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese. In *Proceedings of ACL*, pp.1045-1053, 2012.
[16] Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. Improving a simple bigram HMM part-of-speech tagger by latent annotation and self-training. In *Proceedings of NAACL (Short Paper)*, pp.213-216, 2009.
[17] Zhongye Jia, Peilu Wang, and Hai Zhao. Graph model for Chinese spell checking. In *Proceedings of SIGHAN*, pp.88-92, 2013.
[18] Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. Applying many-to-many alignment and hidden Markov models to letter-to-phoneme conversion. In *HLT-NAACL*, pp.372-379, 2007.
[19] Wenbin Jiang, Liang Huang, and Qun Liu. Automatic adaptation of annotation standards: Chinese word segmentation and POS tagging - a case study. In *Proceedings of ACL-IJCNLP*, pp.522-530, 2009.
[20] Wenbin Jiang, Haitao Mi, and Qun Liu. Word lattice reranking for Chinese word segmentation and part-of-speech tagging. In *Proceedings of Coling*, pp.385-392, 2008.
[21] Wenbin Jiang, Meng Sun, Yajuan Lii, Yating Yang, and Qun Liu. Discriminative learning with natural annotations: Word segmentation as a case study. In *Proceedings of ACL*, pp.761-769, 2013.
[22] Nobuhiro Kaji and Masaru Kitsuregawa. Splitting noun compounds via monolingual and bilingual paraphrasing: A study on Japanese katakana words. In *Proceedings of EMNLP*, pp.959-969, 2011.
[23] Nobuhiro Kaji and Masaru Kitsuregawa. Efficient word lattice generation for joint word segmentation and POS tagging in Japanese. In *Proceedings of IJCNLP*, pp.153-161, 2013.
[24] Philip Koehn and Kevin Knight. Empirical methods for compound splitting. In *Proceedings of EACL*, pp.187-193, 2003.
[25] Canasai Kruegkrak, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Ketaro Torisawa, and Hitoshi Iahara. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In *Proceedings of ACL*, pp.513-521, 2009.
[26] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP*, pp.230-237, 2004.
[27] Sadao Kurohashi and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pp.22-38, 1994.
[28] Dekang Lin, Shaojun Zhao, Benjamin Van Durme, and Marius Paşca. Mining parenthetical translation from the Web by word alignment. In *Proceedings of ACL*, pp.994-1002, 2008.
[29] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of NAACL*, pp.152-159, 2006.
[30] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of ACL*, pp.100-108, 2009.
[31] Shinsuke Mori and Makoto Nagao. Word extraction from corpora and its part-of-speech estimation using distributional analysis. In *Proceedings of Coling*, pp.1119-1122, 1996.
[32] Yugo Murawaki and Sadao Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proceedings of EMNLP*, pp.429-437, 2008.
[33] Masaki Nagata. A part of speech estimation method for Japanese unknown words using a statistical model

of morphology and context. In *Proceedings of ACL*, pp.277－284, 1999.
［34］ Tetsuji Nakagawa and Yuji Matsumoto. Guessing parts－of－speech of unknown words using global information. In *Proceedings of COLINGACL*, pp.705－712, 2006.
［35］ Tetsuji Nakagawa and Kiyotaka Uchimoto. A hybrid approach to word segmentation and postagging. In *Proceedings of ACL Demo and Poster Sessions*, pp.217－220, 2007.
［36］ Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. Automatic acquisition of basic Katakana lexicon from a given corpus. In *Proceedings of IJCNLP*, pp.682－693, 2005.
［37］ Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of ACL (Short Papers)*, pp.529－533, 2011.
［38］ Fuchun Peng, Fangfang Feng, and Andrew McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of COLING*, pp.562－568, 2004.
［39］ Xian Qian and Yang Liu. Joint Chinese word segmentation, pos tagging and parsing. In *Proceedings of the EMNLP－CoNLL*, pp.501－511, 2012.
［40］ Alexander Rush, Roi Reichart, Michael Collins, and Amir Globerson. Improved parsing and POS tagging using inter－sentence consistency constraints. In *Proceedings of EMNLP*, pp.1434－1444, 2012.
［41］ Ryohei Sasano, Sadao Kurohashi, and Manabu Okumura. A simple approach to unknown word processing in Japanese morphological analysis. In *Proceedings of IJCNLP*, pp.162－170, 2013.
［42］ Weiwei Sun. A stacked sub－word model for joint chinese word segmentation and part－of－speech tagging. In *Proceedings of ACL*, pp.1385－1394, 2011.
［43］ Weiwei Sun and Xiaojun Wan. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations. In *Proceedings of ACL*, pp.232－241, 2012.
［44］ Weiwei Sun and Jia Xu. Enhancing chinese word segmentation using unlabeled data. In *Proceedings of EMNLP*, pp.970－979, 2011.
［45］ Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. Training conditional random fields using incomplete annotations. In *Proceedings of Coling*, pp.897－904, 2008.
［46］ Kiyotaka Uchimoto, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, and Hitoshi Isahara. Morphological analysis of a large spontaneous corpus in Japanese. In *Proceedings of ACL*, pp.479－488, 2003.
［47］ Gertjan van Noord. Using self-trained bilexical preferences to improve disambiguation accuracy. In *Proceedings of IWPT*, pp.1－10, 2007.
［48］ Yiou Wang, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Ken-taro Torisawa. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of IJCNLP*, pp.309－317, 2011.
［49］ Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. Graph－based semi－supervised model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of ACL*, pp.770－779, 2013.
［50］ Longkai Zhang, Li Li, Zhengyan He, Houfeng Wang, and Ni Sun. Improving Chinese word segmentation on micro－blog using rich punctuations. In *Proceedings of ACL (Short Papers)*, pp.177－182, 2013.
［51］ Yue Zhang and Stephen Clark. Joint word segmentation and POS tagging using a single perceptron. In *Proceedings of ACL*, pp.888－896, 2008.
［52］ Yue Zhang and Stephen Clark. A fast decoder for joint word segmentation and POS tagging using a single discriminative model. In *Proceedings of EMNLP*, pp.843－8526, 2010.
［53］ 岡部晃，小町守，小木曽智信，松本裕治．表記のバリエーションを考慮した近代日本語の形態素解析．人工知能学会全国大会，2013.
［54］ 岡野原大輔，辻井潤一．Shift－Reduce 操作に基づく未知語を考慮した形態素解析．言語処理学会第 14 回年次大会発表論文集，2008.
［55］ 海野裕也．形態素解析の過去・現在・未来．http://www.slideshare.net/pfi/ss－9805912.
［56］ 工藤拓．Mecab 汎用日本語形態素解析エンジン．http://www.jtpa.org/files/MeCab.pdf.
［57］ 工藤拓，市川宙，David Talbot，賀沢秀人．Web 上のひらがな交じり文に頑健な形態素解析．言語処理学会第 18 回年次大会論文集，pp.1272－1275，2012.
［58］ 持橋大地，鈴木潤，藤野昭典．条件付確率場とベイズ階層言語モデルの統合による半教師あり形態素解析．言語処理学会第 17 回年次大会発表論文集，pp.1071－1074，2011.
［59］ 小林義行，徳永健伸，田中穂積．名詞間の意味的共起情報を用いた複合名詞の解析．自然言語処理，Vol.3，No.1，pp.29－43，1996.
［60］ 長尾真（編）．自然言語処理．岩波書店，1996.
［61］ 萩原正人，関根聡．半教師あり学習に基づく大規模語彙に対応した日本語単語分割．言語処理学会第 18 回年次大会論文集，pp.1280－1283，2012.
［62］ 萩原正人，中山敬広，水野貴明（訳）．入門自然言語処理．オライリー・ジャパン，2010.
［63］ 風間淳一，宮尾祐介，辻井潤一．教師なし隠れマルコフモデルを利用した最大エントロピータグ付けモデル．自然言語処理，Vol.11，No.4，pp.3－23，2003.
［64］ 風間淳一，光石豊，牧野貴樹，鳥澤健太郎，松田晃一，辻井潤一．チャットのための日本語形態素解析．言語処理学会年次大会論文集，pp.590－512，1999.
［65］ 鍛治伸裕，福島健一，喜連川優．大規模ウェブテキストからの片仮名用言の自動獲得．電気情報通信学会論文誌 D（データ工学特集号），Vol.J92－D，No.3，pp.293－300，2009.

（2013年10月14日　受付）

[問い合わせ先]
〒153-8505　東京都目黒区駒場4-6-1
東京大学 生産技術研究所
鍛治　伸裕
TEL：03-5452-6098
FAX：03-5452-6457
E-mail：kaji@tkl.iis.u-tokyo.ac.jp

著 者 紹 介

鍛治（かじ）　伸裕（のぶひろ） [非会員]

2005年東京大学情報理工学系研究科博士後期課程修了．博士(情報理工学)．同年東京大学生産技術研究所産学官連携研究員．特任助手，特任助教を経て，現在，同大学生産技術研究所特任准教授．自然言語処理に関する研究に興味を持つ．