# 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

萩原 正人　　関根 聡
楽天株式会社 楽天技術研究所
{masato.hagiwara, satoshi.b.sekine}@mail.rakuten.com

## 1　はじめに

形態素解析をはじめとする日本語の各種解析技術が，様々なテキストに適用されるようになるにつれ，適用先の分野において高い解析精度を低いコストで実現する技術が求められている．日本語の単語分割および形態素解析には，CRF などのアルゴリズムに基づき，コーパスから単語もしく品詞の接続コストを学習し，入力に対して最もコストの低い形態素列を求める**系列予測**の手法が一般的である [7] ．しかし，ある分野で学習したモデルを他分野に適用する際には，全ての単語に単語境界および品詞がアノテーションされた適用先分野のコーパスを新たに用意する必要があり，これには専門知識や，単語・品詞等の日本語の深い理解，もしくはその両方が必要である．

この問題に対して，解析器をいかに低コストで専門分野に適応させるかという分野適応の手法が盛んに研究されている．森ら [4]，Neubig ら [1] は，文字と文字の境目に単語境界があるかどうかを 2 値分類問題として定式化した**点推定**に基づく単語分割を提案している．単語境界の判定は，周囲の文字列から求められた素性に基づき，各境界に対して独立に行われ，周囲の単語境界に依存しない．この手法では，品詞の付与されていない単語リストからなる辞書を言語資源として利用できる．また，単語分割や品詞のアノテーションが部分的にしか付与されていないコーパスから学習できる点を利用し，分類器の確度の低い部分を人間に提示し，それを修正後，教師データに加えて分類器を再度学習するという**能動学習**を通じて，分野適応を実現できる，という利点がある．ただし，能動学習でも，アノテーションには専門知識を持った人手が介入する必要があり，依然としてコストの高い作業である．

ここで，点推定・系列予測等のアルゴリズムに関わらず，形態素解析の不得意とする入力文は，ひらがな・カタカナの連続，長い漢字複合語などを含んだものである．これら字種などの手がかりに乏しい入力を正確に分割するためには，辞書などの語彙的な知識に頼らざるを得ない．そのためには，対象分野における語彙知識を大量に用意する必要があるが，それを人手により構築するのはやはり容易ではない．

この問題に対して本稿では，半教師あり学習により大量のテキストから語彙を自動で抽出し，単語分割モデルにフィードバックすることにより，日本語単語分割の性能を向上させる手法を提案する．本手法は，新たに教師データを追加することなく，機械学習の出力のみに基づきモデルを再学習する点において，自己学習の一手法とみなすことができる．また，分割の難しい複合語や接辞などの分割誤りにロバストに対応するため，**確率的単語分割** [3, 5] の概念を応用する．確率的単語分割とは，各文字の間に単語境界が存在する確率を与える単語分割であるが，その特殊な例として本研究では「分割」「半分割」「非分割」の 3 種類のみの分割のタグが付られた離散確率的単語分割の概念を導入する．この **3 段階単語分割**コーパス中の $n$-gram 出現頻度を確率的に計算することにより，より分割誤りに対して頑健かつ正確な辞書が作成でき，その結果，単語分割の性能が向上することを示す．

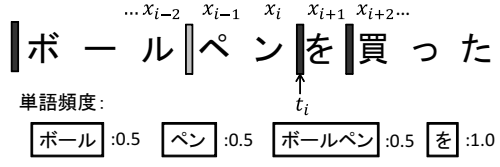


図 1: 点推定による 3 段階単語分割
文字 “ン”“を” の間の分割タグ $t_i$ を求める例

本稿では，まず 2 節にて，森ら [4] の提案する点推定による日本語単語分割について解説する．続いて 3 節にて，確率的単語分割の特殊な場合である 3 段階単語分割を導入する．4 節では，3 段階単語分割を用いてから単語の出現頻度を求め，抽出された語彙を形態素解析辞書としてモデルにフィードバックする手法について述べる．5 節では，同一分野（学習コーパスと適用先大規模コーパスが同分野の場合）および分野適応（異分野の場合）の各場合において単語分割性能を評価した結果を示す．

## 2　点推定による単語分割

本節では，点推定による単語分割の概要を，文献 [4] にならい説明する．ここでのタスクは，入力文字列 $\boldsymbol{x} = x_1x_2...x_n$ に対し，単語境界タグ $\boldsymbol{b} = b_1b_2...b_{n-1}$ を割り当てることであり，2 値分類問題として定式化できる．ここで，$b_i$ は，文字 $x_i$ と $x_{i+1}$ との間に単語分割が存在するかどうかを表すタグであり，$b_i = 1$ の場合は分割，$b_i = 0$ の場合は非分割に対応する．図 1 に，“ボールペンを買った” という入力分に対して，文字 “ン” と “を” の間の分割タグを分類する例を示した．タグは，その周辺に存在する文字から求められる素性を参照し，分類器により分類される．分類に用いた素性は，以下の通りである：

- **文字素性**： 境界位置 $b_i$ に接する，もしくは境界位置を内包する長さ $n$ 以下の全ての文字 $n$-gram と，$b_i$ に対する相対位置である．例えば，図 1 において $n = 3$ の場合，-1/ン, 1/を, -2/ペン, -1/ンを, 1/を買, -3/ルペン, -2/ペンを, -1/ンを買, 1/を買っ, の 9 個の素性が作られる[1]．
- **文字種素性**：文字の代わりに文字種を扱うこと以外は，上記の文字素性と同様である．文字種としては，ひらがな，カタカナ，漢字，アルファベット大文字，アルファベット小文字，アラビア数字，漢数字，中黒（・）の 8 種類を考慮した[2]．
- **辞書素性**：境界位置周辺にある長さ $j(1 \leq j \leq k)$ の単語が辞書に存在するかどうかを表す素性であり，境界位置 $b_i$ がその単語の終点 (L)・始点 (R)・内包する (M) かのフラグと，その単語の長さ $j$ の組み合わせである．例えば，図 1 において，辞書に “ペン” “を” が含まれている場合， 各単語に対応して L2, R1 という素性が作られる．なお，本研究では，語彙知識として複数の辞書を与えられるように変更を加えた．その場合，DIC1-L2, DIC2-R1, など，各辞書に対して異なる素性が作られる．

ここで，文字・文字種素性の $n$-gram の最大長 $n$ および辞書素性の単語の最大長 $k$ のパラメータがある．本研究では予備実験の結果， $n = 3, k = 8$ に固定した．

## 3 3 段階単語分割コーパス

日本語の単語分割においては，分割を一意に決定するのが難しい単語というのが存在し，さらに応用によって適切な分割は異なるという問題がある．例えば，“ボールペン” という単語を含んだ文書集合を，キーワード検索する場合を想定する．分割しなければ，“ペン” という単語にマッチしない（再現率の低下）が，一方で，“ボール／ペン” と 2 単語に分割すれば，例えばユーザーが野球の “ボール” を探しているような場合にマッチしてしまう（精度の低下）．このような英語起源の複合名詞に対して,『現代日本語書き言葉均衡コーパス』の分割認定規定 [2] では,「原語で 1 語となるものの結合体が『リーダーズ英和辞典』第 2 版で 1 語として扱われている」という短単位の認定基準を採用している．しかし “ボールペン” のような和製英語および英語起源ではない外来語の場合，客観的な定義を定めるのは難しく，このような基準が適切かどうかについては疑問が残る．

この問題に対して，文字間の分割を，2 値ではなく確率的に与える確率的単語分割が提案されている [3, 5]．しかしながら，全ての文字間に確率値を付与するため計算コストが高く，また，全ての部分文字列が単語候補となるため，そこから抽出される語彙サイズが非常に大きくなる．この問題に対応するために文献 [3] では，各単語境界を，確率に応じてランダムに「分割」「非分割」のどちらかに寄せた「疑似確率的単語分割コーパス」を生成し，それを用いて単語 $n$-gram の頻度を決定的に計算している．この手法は一種のサンプリングであり，乱数を使い単語の頻度を近似的に求めるため，この操作を複数回繰り返す必要がある．

一方，一般的に人間の認識できる単語分割の強さは多くて数段階であると考えられる．さらに，本研究で用いた EC 分野のテキストには句読点・記号等が数多く含まれているため，多くの単語境界が「分割」$b_i = 1$ であり，全ての文字境界において連続確率値を考えることは無駄である．よって本研究では「分割」「非分割」の他に，$b_i = 0.5$「半分割」を加えた 3 段階の離散確率的単語分割を考え，これを **3 段階単語分割**と呼ぶ．“ボール-ペン” のような複合名詞，“折り-たたむ” のような合成的な複合動詞，“お-すすめ” のような，接辞も含めて語彙化しているような単語の中の分割は，半分割として定義するのが自然である．他にも，“充-電-池” のような「AB + BC → ABC」型の複合語，“妊-産-婦” のような「AC + BC → ABC」型の複合語についても半分割として定義した．

確率的単語分割コーパスから，単語ユニグラム $w$ の頻度 $f_r$ は，単語 $w$ の表記の出現 $O_1 = \{(i,k)|\boldsymbol{x}_{i+1}^{k} = w\}$ を用いて，

$$f_r(w) = \sum_{(i,k)\in O_1} b_i \left[\prod_{j=i+1}^{k-1}(1-b_j)\right] b_k \qquad (1)$$

として計算される．図 1 の例において，例えば “ボールペン” の出現頻度は 1.0 * 1.0 * 1.0 * 0.5 * 1.0 * 1.0 = 0.5 となる．この 3 段階単語分割は，そのまま 3 値分類問題として定式化ができるため，2 種類の分類器を使い容易に求められる[3]．本手法は，既存の CRF 等から分割確率を求める場合 [6] と比較し，学習・テストともに高速であるという特徴がある．また，3 段階単語分割コーパスから，文中に含まれる単語およびその頻度（重み）を求めるのも高速かつ単純であるため，既存の検索システム・プラットフォーム等のインデクシング時に容易に実装できるという特徴がある．また，半分割と判断された単語については，単語全体および単語の構成要素（例 : “ボール” “ペン”）の両者が抽出されるため，人間にとって分割・非分割の判断が難しい単語をとりあえず半分割としておくということが可能であり，アノテーションも容易になる．

## 4 語彙抽出および再学習

本節では，上記の 3 段階単語分割モデルを用いてアノテーション無しの大規模テキストを解析し，そこから抽出された語彙を用いてモデルを再学習する手法について述べる．

[1] 森ら [4] は，境界位置から離れた文字 $n$-gram（例えば -2/ペ 等）も用いているが，これらは分類性能にほとんど寄与しないことが予備実験の結果分かったため，本研究では用いていない．

[2] 森ら [4] はアルファベットの大文字小文字，漢数字，中黒の区別をしていないが，本研究において，これらの分割は単語分割の性能向上に対して有用であることが予備実験の結果分かった．

[3] なお，今回扱った分野の文書では，単語の平均長が約 3 文字であり，したがって過半数の単語境界が $b_i = 0$（非分割）である．そのため，まず非分割かそれ以外かという 2 値分類器により判定し，それで非分割以外と判定された場合のみ，分割か半分割かを判断する 2 番目の分類器を用いて判断すれば効率的である．

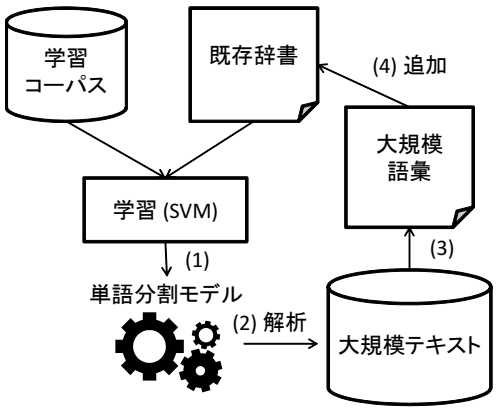

図 2: 語彙抽出および再学習のプロセス

再学習の全体のプロセスを図 2 に示した．まず，3 段階単語分割のアノテーションされた学習コーパスおよび既存の辞書を用いて，ベースラインのモデルを学習する（1）．続いて，ベースラインのモデルを用いて，大規模テキストを解析し，3 段階単語分割を自動でアノテーションしたコーパスを作る（2）．こうしてできた 3 段階単語分割コーパスから，式 (1) に従って単語ユニグラムの頻度を計算し，大規模語彙リストを抽出する（3）．最後に，抽出された語彙リストを，単語分割モデルの素性に用いる辞書として追加し（4），単語分割モデルを再学習する（1）．語彙を辞書として追加する手法として，以下の 4 種類を比較検討した．なお，獲得された語彙の集合を $V$ とする．

1. **APPEND** $V$ のうち頻度の高いものだけを既存の辞書（UniDic）に追加する．
2. **TOP** $V$ のうち頻度の高いものだけを別の辞書として追加する．
3. **ALL** $V$ の全体を別の辞書として追加する．
4. **MULTI** $V$ を頻度ごとに別の辞書として追加する．$V$ の頻度上位 $x$ 位までの部分集合を $V_x$ とすると，例えば，$V_{1000}, V_{2000}, V_{3000}$ を作成し，それぞれを別の辞書として分割モデルを再度学習する．ここで，$V_x \subset V_y$ iff $x < y$ が成り立つ．これにより，例えば頻度が上位 1000 位以内の単語であれば全ての辞書に対応する素性が作られるが，例えば頻度が上位 2500 位の語に対しては $V_{3000}$ の素性しか対応しないため，分類器が単語の頻度に応じて異なる重みを学習することが期待できる．

# 5 評価実験

単語分割の性能評価指標には，精度 (Prec)・再現率 (Rec)・F 値を用いた．正解コーパスに含まれる延べ単語数を $N_{REF}$，解析結果に含まれる延べ単語数を $N_{SYS}$，解析結果と正解コーパスの両者に含まれる延べ単語数を $N_{COR}$ とすると，Prec $= N_{COR}/N_{SYS}$, Rec $= N_{COR}/N_{REF}$, F $= 2Prec \cdot Rec/(Prec + Rec)$ として計算される．なお，性能評価の際には，純粋に単語分割性能を比較するため，半分割を分割に統一することにより，正解コーパスおよび解析結果の両方を 2 段階単語分割コーパスに変換した．

表 1: 同一分野における単語分割性能比較 (%)

| コーパス | 手法 | 精度 | 再現率 | F 値 |
|---|---|---|---|---|
| ベースライン | | 97.68 | 97.93 | 97.80 |
| 2 段階 | APPEND | 97.79 | 98.01 | 97.90 |
| | TOP | 97.96 | 98.19 | 98.08 |
| | ALL | 98.06 | 98.19 | 98.12 |
| | MULTI | 98.13 | 98.29 | 98.21 |
| 3 段階 | APPEND | 97.98 | 98.13 | 98.06 |
| | TOP | 98.04 | 98.25 | 98.14 |
| | ALL | 98.20 | 98.23 | 98.21 |
| | MULTI | **98.27** | **98.34** | **98.30** |

ベースラインの辞書としては，UniDic [4]の見出し語リスト（異なり 304,267 語）を用いた．分類に用いたサポートベクトルマシンの実装には，LIBLINEAR[5]（線形カーネル）をデフォルトパラメータで使用した．なお，学習コーパスおよび対象テキスト中のいわゆる半角文字（英数字・記号・カタカナ）は全て全角に統一したが，それ以上の正規化は行わなかった．

## 5.1 同一分野における自己学習

本節では，学習コーパスおよび適用先の大規模テキストが同じ分野である場合についての実験結果を示す．まず，学習コーパスを作成するために，楽天市場[6]の全商品からジャンルの偏り無くランダムに抽出した 590 商品のタイトルおよび説明文，および，楽天プロダクト[7]からランダムに抽出した商品説明 50 商品分を 2 人の評価者がアノテーションすることより 3 段階単語分割コーパスを作成した．単語分割基準は，半分割を除き文献 [2] に従った．本コーパスの単語数は約 11 万，文字数にして約 34 万文字である．このコーパスを用い 10 分割の交差検定により性能を評価した．

適用先大規模テキストとしては，楽天市場全商品データのタイトルおよび説明文を用いた．なお，重複・類似商品が多数あるため，タイトル・説明文の文字バイグラムから作成したベクトル間の余弦類似度が 0.8 以上である商品対を重複商品として判断し，片方を除外した．重複排除後の商品数は約 2700 万であり，文字数にして約 160 億文字である．

ベースラインモデルにより大規模テキストを解析した後，2 段階単語分割を用いた場合，異なり 576,954 語，3 段階単語分割を用いた場合，異なり 603,187 語が抽出された．いずれの場合も，頻度 20 以上の単語のみを用いている．手法 MULTI を用いて辞書を追加する場合，得られた単語リストの，頻度上位 10 万, 20 万, 30 万, 40 万, および全体を別々に辞書として追加した．手法 TOP の場合，上位 10 万のみを用いた．図 1 に，ベースライン，2 段階単語分割・3 段階単語分割を使って再学習した結果を示す．

2 段階単語分割を使いモデルを再学習した場合，どの手法 (APPEND/TOP/ALL/MULTI) を用いて大規

[4]http://www.tokuteicorpus.jp/dist/
[5]http://www.csie.ntu.edu.tw/~ cjlin/liblinear/
[6]http://www.rakuten.co.jp/
[7]http://product.rakuten.co.jp/

表 2: 分野適応における単語分割性能比較 (%)

| コーパス | 手法 | 精度 | 再現率 | F 値 |
|---|---|---|---|---|
| ベースライン | | 96.13 | 96.62 | 96.38 |
| 分野適応 | | 96.60 | **96.99** | 96.79 |
| 2 段階 | APPEND | 96.50 | 96.55 | 96.52 |
| | TOP | 96.44 | 96.19 | 96.32 |
| | ALL | 96.69 | 96.40 | 96.55 |
| | MULTI | 96.67 | 96.45 | 96.56 |
| 3 段階 | APPEND | 96.38 | 96.53 | 96.46 |
| | TOP | **96.80** | 96.97 | **96.89** |
| | ALL | 96.64 | 96.89 | 96.76 |
| | MULTI | 96.69 | 96.78 | 96.73 |

模語彙に追加した場合にも，F 値が向上していることが分かり，提案する大規模テキストを用いた半教師あり学習が有効であることを示している．また，F 値の増加幅は APPEND<TOP<ALL<MULTI の順で大きく，語彙を追加する場合，既存の辞書に追加するよりも別の辞書として追加した方が，また，全体を単一の辞書として追加するよりも頻度に応じて別の辞書として追加した方が効果的である．この結果より，分類器によって，単語の頻度に応じて異なる重みが自動的に学習できていると考えられる．2 段階単語分割による単語フィードバック後，ベースラインと比較して単語分割が改善した例としては “テレ／キャスター” → “テレキャスター”（エレキギターの一種），“被-包-材” → “被包-材” などがあり，それぞれ “テレキャスター”，“被包” という単語を正しく獲得できていることが分かる．

さらに，3 段階単語分割を使いモデルを再学習した場合，全ての場合においてベースラインおよび 2 段階単語分割よりも性能が向上している．単語分割の改善した例としては “テクビヨウ” → “テクビ／ヨウ”，“表面積” → “表／面積” などがあり，半分割を考慮することによりこれら接辞を伴う単語がより正確に獲得できていることが分かる．

### 5.2 異分野における自己学習

本節では続いて，学習コーパスと適用先の大規模テキストの分野が異なる場合の本手法の有効性について検証する．学習コーパスとしては上述の商品ドメインの参照コーパスを用いたが，適用先の大規模テキストとして，楽天トラベル[8]のユーザーレビューを用いた．本コーパスは，ユーザーのレビュー，宿泊施設名，宿泊プラン名，宿泊施設からの返答，で構成され，合計 348,564 件，文字数にして約 1 億 2600 万文字である．そのうち 150 件と 50 件のレビューをランダムに抽出し，人手による単語分割後，それぞれテストコーパスおよび能動学習用コーパス（学習コーパスに対する追加分）として用いた．

まず，上記商品分野の学習コーパスから学習したベースラインを用い，対象分野の大規模テキストを解析した．その時点での性能が図 2 の「ベースライン」である．ここで抽出された語彙をモデルにフィードバックし，再学習したところ，単語分割性能の向上は見られなかった．異なる分野のモデルで解析して得られた語彙の信頼性は低く，かつ，学習コーパスにも現れるものは少ないことから，分類器が適切な重みを学習できず，辞書として追加しても効果的ではないことが分かる．

次に，上記商品分野の学習コーパスに分野適応用のコーパスを加えて単語分割モデルを学習した後，それを用いて対象分野の大規模テキストを解析した．その時点での性能が図 2 の「分野適応」である．大規模テキストを解析した後，2 段階単語分割を用いた場合，異なり 41,671 語，3 段階単語分割を用いた場合，異なり 44,247 語が抽出された．いずれの場合も，頻度 5 以上の単語のみを用いている．これら得られた単語を辞書として追加し，学習コーパスおよび分野適応用コーパスを用いてモデルを再学習した結果が表 2 である．2 段階単語分割を使用した場合は分野適応に比べて総じて性能が低下している一方，3 段階単語分割の場合はわずかながら性能の向上が見られる．これは，獲得された語彙の少なくとも一部が，学習コーパス（追加分）に含まれるため，分類器が重みを学習できているためである．しかし，性能向上は前述の同一分野の場合に比べるとわずかであり，適用先分野が異なる場合，特に分野適応用コーパスが小さい場合には，本手法による語彙フィードバックの効果は限定的であることが分かる．分野適用後も “素泊まり” などの分野依存語や “美ら海” などの固有名詞で分割を誤っており，これらを正しく解析するためにはさらに多くのアノテーションが必要であると考えられる．

[8]http://travel.rakuten.co.jp/

## 6 おわりに

本稿では，大量のテキストから語彙を抽出し，単語分割モデルにフィードバックすることにより，日本語単語分割の性能を向上させる半教師あり学習手法を提案した．また，確率的単語分割の特殊な場合としての 3 段階単語分割を提案・実装し，通常の 2 段階単語分割と比較して性能向上に有効であることを示した．品詞も含めた統合的な形態素解析にどのように応用するかは今後の課題である．

## 参考文献


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

[2] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版. 国立国語研究所, 2011.

[3] 森信介, 小田裕樹. 擬似確率的単語分割コーパスによる言語モデルの改良. 自然言語処理, Vol. 16, pp. 77–84, 2009.

[4] 森信介, 中田陽介, Graham Neubig, 河原達也. 点予測による形態素解析. 言語処理学会論文誌, Vol. 18, No. 4, pp. 367–381, 9 2011.

[5] 岡野原大輔, 工藤拓, 森信介. 形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用. NLP 若手の会第 1 回シンポジウム, 2006.

[6] 工藤拓. 形態素周辺確率を用いた分かち書きの一般化とその応用. 言語処理学会第 11 回全国大会, 2005.

[7] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学会研究報告, NL161 巻, 2004.