情報処理学会論文誌 Vol. 52 No. 10 2944–2952 (Oct. 2011)

点予測による自動単語分割

森 信介[†1] ニュービッグ グラム[†2]
坪井 祐太[†3]

本論文では，大量の学習コーパスがある分野で既存手法と同程度かそれ以上の解析精度を保持しつつ，部分的単語分割コーパスなどを利用して安価に分野適応を実現する自動単語分割の設計を提案する．具体的には，推定時の素性として，周囲の単語境界の推定値を参照せずに，周辺の文字列のみを参照する点予測による自動単語分割である．この設計により，単語境界が文の一部にのみ付与された部分的単語分割コーパスを利用することが可能となる．この結果，従来手法に比して格段に高い分野適応性を実現できる．実験では，提案手法と単語 n-gram モデルや条件付き確率場による方法による単語分割の精度を比較し，提案手法が計算時間と精度の両方において優位であることが示された．

A Pointwise Approach to Automatic Word Segmentation

Shinsuke Mori,[†1] Graham Neubig[†2]
and Yuta Tsuboi[†3]

In this paper we propose a design of a word segmenter which allows us a quick domain adaptation keeping a high accuracy in the general domain where a large annotated corpus is available. Our method is based on a pointwise classification which refers only to the neighbouring characters. This design enables us to train our word segmenter by using a partially annotated corpus in which only some parts are annotated. As a result, a high domain adaptability is realized. In the experiments we compared our method and existing methods on word n-gram models or conditional random fields and showed our method is superior to the others in calculation time and accuracy.

†1 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University
†2 京都大学情報学研究科
Graduate School of Infomatics, Kyoto University
†3 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research - Tokyo, IBM Japan, Ltd.

1. はじめに

自動単語分割[1)–3)]は，日本語などの単語境界を明示しない言語の文を単語列に分解する処理である．品詞推定も同時に行う形態素解析[4)–7)]も，学習コーパスの品詞を削除あるいはすべて同一にすることで自動単語分割と見なすことができる[⋆1]．なお，日本語の言語処理においてつねに品詞を推定すべきか否かは議論の余地がある．音声認識や仮名漢字変換の言語モデル作成には品詞は必ずしも必要ではなく，かつ様々な分野において高い単語分割精度が要求されるので，学習コーパス作成のコストも含めた総合設計を考えると単語分割のみで済ませるのがこれらの応用には有望である．

前述の自動単語分割の研究は，すべて系列予測に基づく方法を採用している．すなわち，ある文字間に単語境界があるか否かの決定が，周辺の文字間に単語境界があるか否かの予測に依存する．文は単語の列であり，自動単語分割を系列予測ととらえることは非常に自然である．

しかしながら，この設計により，実装が複雑になるばかりでなく，パラメータ推定や自動単語分割の速度が低下する．また，パラメータ推定には，すべての文字間に単語境界情報を付与したフルアノテーションコーパスが必要となる[⋆2]．この結果，既存手法では，ある分野で高い精度の自動分割器を作成するために必要となるコーパスの作成コストが相対的に大きくなり，ある特定の分野のテキストに対する解析精度を向上させたいという要求に応える速度が低下するとともに，コストが増加する．

このような背景の下，本論文では，大量の学習コーパスがある分野で既存手法と同程度かそれ以上の解析精度を保持しつつ，安価に分野適応を実現する自動単語分割の設計を提案する．具体的には，各文字間に単語境界があるか否かが周辺の文字間の推定値に依存しないと仮定し，パラメータ推定や自動単語分割の速度を低下させることなく，部分的単語分割コーパスなどの言語資源を利用可能とする．すなわち，推定時の素性として，周囲の単語境界の推定値を参照せずに，周辺の文字列のみを参照する方法である．これを点予測による自動単語分割と呼ぶ．点予測と類似の方法による自動単語分割はすでに提案されている[9)]．本論文では，解析精度やコーパスの準備などを考慮した総合設計として点予測による自動単語分割を提案し，系列予測との定量的・定性的比較を行う．

⋆1 中国語の形態素解析において，単語分割と品詞推定を多段に行う場合の速度についての比較はある[8)]．しかしながら，用途や言語資源を含めた包括的な議論は筆者の知る限りない．
⋆2 系列予測に基づく方法でこの制限を取り除くことも提案されている[3)]．

2. 既存手法による自動単語分割

本章では，単語分割問題を定義し，いくつかの既存手法を概説する．主要な既存手法は，単語 n-gram モデル[10] とその拡張であるクラス n-gram モデルを用いて単語列としてモデル化する方法[6] と単語境界の有無をラベルとする条件付き確率場[11] を用いる方法[3),12] であろう．なお，後述の実験では，形態素解析を目的として形態素列をモデル化する条件付き確率場による方法を品詞の数が唯一であるとして自動単語分割器とした場合も比較対象としている．この方法の詳細については文献 7) を参照されたい．

2.1 単語分割問題

単語分割問題は，単語境界を明示しない言語の文を単語列に分解する処理である．すなわち，言語の文字集合 $\mathcal{X}$ の要素の列 $\boldsymbol{x} \in \mathcal{X}^*$ を入力とし，単語 $w \in \mathcal{X}^*$ の列 $\boldsymbol{w}$ を出力とする．ここで，$\boldsymbol{x} = \boldsymbol{w}$ が条件である．

2.2 単語 n-gram モデルによる単語分割

単語 n-gram モデル $M_{w,n}(\boldsymbol{w})$ は，文を単語列 $\boldsymbol{w}_1^h = w_1 w_2 \cdots w_h$ と見なし，その出現確率を以下の式で与える．

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

この式中の w_i（$i \leq 0$）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．完全な語彙を定義することは不可能であるから，未知語を表す特別な記号 UW を用意する．未知語の予測の際は，まず，単語 n-gram モデルにより UW を予測し，さらにその表記（文字列）$\boldsymbol{x}_1^{h'}$ を以下の文字 n-gram モデルにより予測する．

$$M_{x,n}(\boldsymbol{x}_1^{h'}) = \prod_{i=1}^{h'+1} P(x_i|\boldsymbol{x}_{i-n+1}^{i-1}) \tag{2}$$

この式中の x_i（$i \leq 0$）は，語頭に対応する特別な記号であり，$x_{h'+1}$ は，語末に対応する特別な記号である．したがって，未知語は以下のように予測される．

$$P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) = M_{x,n}(w_i) P(\mathtt{UW}|\boldsymbol{w}_{i-n+1}^{i-1})$$

文献 6) では，学習コーパスを 9 つに分割し，その 2 つ以上に出現する単語を語彙とし，それ以外を未知語としている．

式 (1) の確率値 $P(w_i|\boldsymbol{w}_{i-n+1}^{i-1})$ は，単語分割済みコーパスにおける単語 n-gram 頻度 $f(\boldsymbol{w}_{i-n+1}^{i})$ から以下のように最尤推定する．

$$P(w_i|\boldsymbol{w}_{i-n+1}^{i-1}) = \frac{f(\boldsymbol{w}_{i-n+1}^{i})}{f(\boldsymbol{w}_{i-n+1}^{i-1})}$$

この際，すべての未知語を UW に置き換えておく．式 (2) の確率値は，UW に置き換えた未知語を学習コーパスと見なしてその文字 n-gram 頻度から同様に最尤推定する．

単語 n-gram モデルによる自動単語分割[10] は，以下の式で表されるように，文字列 $\boldsymbol{x}$ として与えられる文の生成確率が最大となる単語列を自動分割結果とする．

$$\hat{\boldsymbol{w}} = \mathop{\mathbf{argmax}}_{\boldsymbol{w}=\boldsymbol{x}} M_{w,n}(\boldsymbol{w})$$

2.3 クラス n-gram モデルによる単語分割

クラス n-gram モデル[13] では，あらかじめ単語をクラスと呼ばれるグループに分類しておく．このモデルでは，以下の式のように次のクラス c_i を予測したうえで次の単語を予測する．

$$M_{c,n}(\boldsymbol{w}) = \prod_{i=1}^{h+1} P(c_i|\boldsymbol{c}_{i-n+1}^{i-1}) P(w_i|c_i) \tag{3}$$

この式中の c_i（$i \leq 0$）は，文頭に対応する特別な記号であり，c_{h+1} は，文末に対応する特別な記号である．単語 n-gram モデルと同様に，未知語を表す特別な記号 UW を用意し，未知語の予測の際は，まず，単語 n-gram モデルにより UW を予測し，さらにその表記（文字列）$\boldsymbol{x}_1^{h'}$ を文字 n-gram モデル $M_{x,n}$ により予測する．

式 (3) の確率値 $P(c_i|\boldsymbol{c}_{i-n+1}^{i-1})$ と $P(w_i|c_i)$ は，単語分割済みコーパスとその各単語をクラスに置き換えたコーパスから最尤推定される．

各単語がどのクラスに属するかは，クロスエントロピーを目的関数とする単語クラスタリング[14] により決定される．

クラス n-gram モデルによる自動単語分割[6] は，単語 n-gram モデルの場合と同様に，文の生成確率が最大となる単語列を自動分割結果とする．

2.4 条件付き確率場による単語分割

単語境界の有無をラベルとする条件付き確率場（CRF）[11] による単語分割[3),12] では，ある出力系列 $\boldsymbol{t}$ の入力文 $\boldsymbol{x}$ に対する条件付き確率を以下のように定式化する．

$$P(\boldsymbol{t}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x}}} \exp\left(\sum_{i=1}^{l} \sum_{k} \lambda_k f_k(x_{i-m+1} x_{i-m+2} \cdots x_{i+m}, t_{i-1} t_i) \right)$$

ここで $Z_{\boldsymbol{x}}$ は全系列に対する確率の和が 1 となるようにする正規化項である．f_k は k 番目の素性であり，λ_k はその重みである．素性は，点予測の場合と同様に，判定点の前後 m 文字の文字列の部分文字列である文字 n-gram と文字種 n-gram，および辞書素性である．これらに加えて，窓幅 m 内のラベルも素性として参照される．

単語境界の有無をラベルとする条件付き確率場による自動単語分割は，条件付きの生成確率が最大となるラベル列 $\mathbf{argmax}_{\boldsymbol{t}}\, P(\boldsymbol{t}|\boldsymbol{x})$ を自動分割結果とする．

3. 点予測を用いた自動単語分割

前章で説明したいずれの既存手法も，ある文字がある単語に属するかやある文字境界に単語境界があるかを周辺の推定値を含めてモデル化している．本章では，周辺の推定値を参照しない点予測を用いる手法による単語分割を提案する．

提案手法では，単語分割の問題を入力文字列の部分文字列への分解と考え，各文字間において，それが部分文字列の境界であるか否かを判別する 2 値分類として定式化する．すなわち，文字 x_i と x_{i+1} の間の単語境界の有無を示すタグ t_i を出力する．単語境界タグ t_i の値域は，文字 x_i と x_{i+1} の間に単語境界が「存在する」か「存在しない」の 2 値とする．入力文字列（長さを l とする）の両端に単語境界があるのは自明なので，$t_1 t_2 \cdots t_{l-1}$ を予測する問題となる．

点予測を用いた単語分割では，分類器は，ある文字間の単語境界の有無の予測時に，周辺の文字列情報のみから得られる以下の 3 種類の素性を参照する（**図 1** 参照）．

1. 文字 n-gram：判別するタグ位置 i の前後の部分文字列であり，窓幅 m と長さ n のパラメータがある．素性は，長さ $2m$ の文字列 $x_{i-m+1} \cdots x_{i-1} x_i x_{i+1} \cdots x_{i+m}$ の長さ n のすべての部分文字（文字 n-gram）である．
2. 文字種 n-gram：文字を文字種に変換した列を対象とする点以外は文字 n-gram と同じである．文字種は，漢字，片仮名，平仮名，ローマ字，数字，その他の 6 つである．
3. 単語辞書素性：各長さ k に対する以下のフラグである．
   - 判別点の左の部分文字列 $x_{i-k+1} x_{i-k+2} \cdots x_i$ が単語として辞書にあるか
   - 判別点の右の部分文字列 $x_{i+1} x_{i+2} \cdots x_{i+k}$ が単語として辞書にあるか
   - 判別点をまたぐ部分文字列 $x_{i-j+1} x_{i-j+2} \cdots x_{i-j+k}$（$1 \leq \forall j < k$）が単語として辞書にあるか

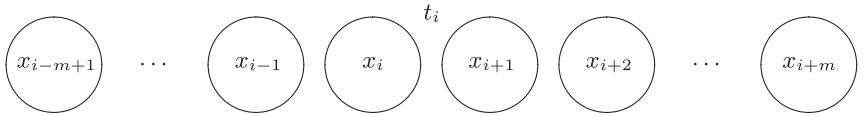

図 1　ある判定点 t_i の分類に参照する文字
Fig. 1　The characters to be referred to for a decision point t_i.

系列予測では，周辺の予測結果も参照するが，点予測では，これらの誤りの可能性がある情報を参照しないことに留意されたい．SVM やロジスティック回帰などのある識別関数を $f(\boldsymbol{x}, t_i)$ とすると，点予測による単語分割は以下の計算を各 i に対して行うことである．

$$\hat{t}_i = \mathop{\mathbf{argmax}}_{t_i} f(\boldsymbol{x}, t_i)$$

ここで，窓幅を m としているので $\boldsymbol{x} = x_{i-m+1} x_{i-m+2} \cdots x_{i+m}$ である．

4. 点予測と系列予測の定性的な比較

この章では，提案する点予測と条件付き確率場を代表とする系列予測の定性的な比較について述べる．

4.1 参照する素性の範囲とその影響

自動単語分割における点予測と系列予測の差異は，ある文字間の単語境界の有無の予測に際して，周辺の単語境界の有無の予測値を参照するか否かである．これを参照する系列予測では，素性として直接参照する範囲の外にある入力文字列も間接的に参照する．たとえば，t_i の予測に際して，t_{i-1} の予測を参照することにより，窓幅 m よりも 1 つ前の文字 x_{i-m} を間接的に参照することになる（図 1 参照）．すなわち，同じ素性（文字 n-gram など）を参照する点予測に比べて，結果的により広い範囲の入力文字列を参照することになる．結果として，その分の精度向上が期待される．一方で，周辺の単語境界の有無は予測値であるので単語分割精度と同程度にのみ信頼できるので，単語分割精度の向上にはつながらないと考えられる．すなわち，ある点の周辺に単語境界があるか否かの予測が正しければ，その点に単語境界があるか否かの予測が正しくなる傾向があるが，逆に周辺での予測が誤りであれば，その点での予測を誤る傾向があると考えられる．つまり，周辺の単語境界の有無の予測値を参照するか否かは，直接精度向上には直結しておらず，より広い範囲の入力文字列を参照することで精度向上が期待されるのみである．もし，より広い範囲の入力文字列を参照することが有効であるなら，点予測においてもより広範囲の入力文字を素性として参照することで精度が向上するはずなので，精度という観点では，点予測と系列予測の本質的な差異はないと考えられる．

4.2　実装および計算時間

実装という観点からは，点予測は系列予測に比べて格段に簡潔である．これは，パラメータ推定のときでも単語分割のときでも同じである．このことは，点予測においては，ある文字間の単語境界の有無が周辺の単語境界の有無と独立であると仮定していることから明らかであろう．同一のフルアノテーションコーパスから学習する場合のパラメータ推定と単語分割に要する時間の定量的比較については，次章で述べる．

点予測では，単語分割のときには，他の判定点に対するラベルの予測値を参照しないので，判定点を単位とする並列化も可能である．ただし，素性の元情報である入力文字列の多くを周辺の判定点と共有しているので，密結合である必要がある．また，ロジステック回帰など，ラベルとともにその確率を推定する分類器を用いる場合には，単語境界確率を推定することが可能になる．これにより確率的単語分割[15]などの枠組みが利用可能になる．これは，系列予測でも各判定点における周辺化により可能であるが，計算時間が大きいという欠点がある．さらに，単語境界が部分的に付与された文を完全に単語分割する場合も点予測の枠組みでは，未判定の箇所を判定するだけで，実装も実行も非常に単純であるが，系列予測では確定している点が確定している値になる解の中から探索することになり，複雑な実装が必要となる．

パラメータ推定のときは，系列予測に対する点予測の簡便さがより際立つ．点予測では，各点を事例として唯一の分類器を学習する．素性とラベルの組の頻度の集計などは容易に並列化できる．また，単語境界が部分的に付与された文の集まり（部分的アノテーションコーパス）からの学習も非常に容易であり，単にアノテーションされた箇所のみから学習データを生成すればよい．学習に要する計算時間は，アノテーションされた箇所数 a に比例する．これに対して，系列予測では，単語境界が付与されていない箇所に対する期待値を計算する必要がある[3]．学習に要する計算時間は，アノテーションされた箇所を含む文の全文字数 l に比例する．さらに，一般的には，定数部分も点予測よりも大きい．単語 n-gram モデルやクラス n-gram モデルの場合でも，EM アルゴリズム[16]を用いることで実現できるが，EM アルゴリズムは繰返し計算である．いずれのモデルでも，系列予測では，特にスパースな部分的アノテーション（1 文に 1 カ所程度のアノテーションなど）からのパラメータ推定の場合に，膨大な記憶領域と計算時間を要する．このように，特に $a \ll l$ の場合に，計算時間の違いが際立つ．以上に述べたような，パラメータ推定における差異が点予測の系列予測に対する最大の優位性である．

4.3　利用可能な言語資源利用

既存手法である系列ラベリングとしての自動単語分割器のパラメータ推定には，一般的に次の 2 つの言語資源のみが利用可能である．

1. フルアノテーションコーパス：このコーパスの各文は，すべての文字間に単語境界情報が付与されている．自動単語分割器の分野適応に際しては，適応分野の文に対して人手により単語境界情報を付与するのが効果的であるが，各文の大部分の箇所は，多くの単語分割基準において利用可能である一般分野のコーパスに頻出している単語や表現であり，文のすべての箇所に情報を付与することは，限られたコストの配分方法として好ましくない．
2. 単語辞書：この辞書の各見出しは，フルアノテーションコーパスと同じ単語分割基準における単語である．これを作成する作業者は，対象分野の言語表現と単語分割基準の両方を熟知している必要がある．

提案手法では，これらの言語資源に加えて次の言語資源が利用可能となる．

3. 部分的アノテーションコーパス：文の一部の文字間にのみ単語境界情報がアノテーションされたコーパスである．単語境界情報には，単語である場合とない場合があり，アノテーションされていない（不明）を含めると各文字間の可能なラベルは 3 つである．

既存手法の単語分割器の分野適応を行う際には，フルアノテーションコーパスを作成する必要があり，作業者は，不明な箇所や判断に自信のない箇所が含まれる文に対しては，その文すべてを棄却するか確信の持てないまま情報付与を行う必要がある．これは，作業速度の低下か質の低下を招くこととなり，言語資源作成の現場では非常に深刻である．このような理由から，確信の持てる箇所のみへのアノテーションを許容する枠組みが渇望されている．

点予測に基づく単語分割器では，部分的アノテーションコーパスも有効に活用することができる．また，この実現方法も，3 章の説明から分かるように，非常に直感的かつ容易である．この特徴により，フルアノテーションコーパスが利用できない分野においても，高速かつ安価な精度向上が可能となる．一般分野の学習コーパスに出現しない 2 文字にのみアノテーションすることで得られる部分的アノテーションコーパスを利用する場合について，定量的評価を次章で述べる．

5.　点予測と系列予測の定量的な比較

提案手法の定量的な評価を行うために，一般分野の学習コーパスからそれぞれの自動単語分割器を構築し，学習コーパスと同じ分野のテストコーパスと学習コーパスと異なる分野の

テストコーパスに対する自動単語分割の精度を測定した．また，学習コーパスと同じ単語分割基準の下で構築された辞書を用いる場合の精度も測定した．

5.1 言語資源

実験に用いた言語資源は，単語分割済みコーパスと辞書である（**表 1** 参照）．単語分割済みコーパスは，現代日本語書き言葉均衡コーパス 2009 年度版のモニタ公開データ[17] のコアデータである．コーパスの出典は，白書と書籍と新聞と Yahoo!知恵袋である．文献 18) によれば，このうち Yahoo!知恵袋の文は，他の出典の文と大きく性質が異なるので，これを適応分野とし，これ以外を一般分野とした．各文は文献 19) で定義される基準で単語に分割されているが，本実験では，活用語の活用語尾を分割し異なる単語とした[⋆1]．両分野のコーパスを先頭からの文番号の 10 で割った余りに応じて文単位で 10 分割し，最初の 9 個を学習コーパス，最後の 1 個をテストコーパスとした．辞書は UniDic-1.3.12[20] である．これに，コーパスと同様の活用語尾の分割を施し，各見出し語はコーパスと同じ単語分割基準を満たすようにした．

表 2 は，テストコーパスのカバー率である．学習コーパスのみを利用する場合の語彙は，学習コーパスに出現するすべての単語の集合である．学習コーパスと UniDic を利用する場合は，これと UniDic の和集合である．UniDic は，Yahoo!知恵袋も含めた BCCWJ を参考に作成されているので，UniDic を用いる場合は適応分野であってもほぼ 100%のカバー率となっている．したがって，UniDic を用いない場合の方が現実的な分野適応の状況といえるであろう．

表 1　言語資源の諸元
Table 1　Specification of the language resources.

言語資源	出典	用途	文数	形態素数	文字数
現代日本語	白書・書籍・新聞	学習	27,338	782,584	1,131,317
書き言葉均	（一般分野）	テスト	3,038	87,458	126,154
衡コーパス	Yahoo!知恵袋	なし	5,800	114,265	158,000
（BCCWJ）	（適応分野）	テスト	645	13,018	17,980
UniDic	—	学習	—	213,174	664,516

活用語の活用語尾を分割し異なる単語としている．

表 2　テストコーパスのカバー率
Table 2　Coverages of the test corpora.

語彙の構成に利用する言語資源	語彙サイズ	一般分野 [%]	適応分野 [%]
学習コーパスのみ	28,315	98.44	96.51
学習コーパスと UniDic	213,454	99.96	99.85

⋆1 単語分割後の様々な処理において小さい語彙で高いカバー率を実現できるようにするためである．

5.2 各自動単語分割器のパラメータ

点予測による方法の分類器には，ロジスティック回帰（LR dual）[21] と線形 SVM[22] を用いた．素性生成の際に参照する文字 n-gram 長の n の上限値，文字種 n-gram 長の n の上限値，窓幅 m はすべて 3 とした．条件付き確率場（CRF）の実装には CRFsuite[23] を用いた[⋆2]．参照する素性は点予測と同じとした．なお，結果的には，次に述べる文献 3) の系列予測による方法が CRFsuite 系列予測の方法よりも高い精度となったので，CRFsuite による系列予測は点予測との計算時間の比較にのみ意味があった．

また，素性の異なる坪井ら[3] の手法による実験も行った．この手法で用いている素性は，提案手法の素性（3 章）と以下の 2 点で異なっている．

1. タグ位置 i を含む文字 n-gram，すなわち $x_k x_{k+1} \cdots x_{k+n-1}$（$k \leq i \leq k+n-1$）およびそれらに対応する文字種 n-gram のみを使用する．
2. 単語辞書素性のうち，タグ位置 i を内包する単語が辞書にあるか否かのフラグを使用しない．また，辞書エントリの長さ別フラグも含まない．

なお，CRFsuite でこの素性変更を行っても精度の改善は見られなかった．

これらを含むハイパーパラメータの決定は，いずれの手法においても，学習コーパスのみを参照する 9 分割交差検定の結果による．

5.3 評価基準

我々が用いた評価基準は，単語境界推定精度と再現率と適合率と F 値である．単語境界推定精度は，各文字間で自動単語分割結果による判断が一致した割合である．ただし，文の両端を含まない．再現率と適合率の定義は次のとおりである．正解コーパスに含まれる延べ単語数を N_{REF}，自動単語分割結果に含まれる延べ単語数を N_{SYS}，双方で分割が一致した延べ単語数を N_{COR} とすると，再現率は N_{COR}/N_{REF} と定義され，適合率は N_{COR}/N_{SYS} と定義される．F 値は，再現率と適合率の調和平均である．

例として，コーパスの内容と解析結果が以下のような場合を考える．

正解コーパス：本部　長　の　クラス　メート

単語分割結果：本　部長　の　クラスメート

この例文の場合，文には 10 文字あり，単語境界か否かの判断をすべき文字間は 9 カ所である．

⋆2 `feature.possible_states`=1 および `feature.possible_transitions`=1 としている．

このうち，自動単語分割による判断が正解コーパスに一致した箇所は 6 カ所なので，単語境界推定精度は 6/9 である．次に，分割が一致した単語は「の」のみであるので，$N_{COR} = 1$ となる．また，正解コーパスには 5 つの単語が含まれ，単語分割結果には 4 つの単語が含まれているので，$N_{REF} = 5$，$N_{SYS} = 4$ である．よって，再現率は $N_{COR}/N_{REF} = 1/5$ となり，適合率は $N_{COR}/N_{SYS} = 1/4$ となる．最後に，F 値は $[\{(1/5)^{-1} + (1/4)^{-1}\}/2]^{-1} = 2/9$ である．

5.4 実験評価

提案手法を評価するために，提案手法と既存手法を一般分野の学習コーパスから構築し，一般分野と適応分野のテストコーパスの単語分割を行った．**表 3** は，提案手法と一部既存手法の計算時間である．単語分割時の速度は大きく変わらないものの，提案手法のパラメータ推定は，分類器として SVM を用いる限り，既存手法に対して圧倒的に速い．能動学習[24]においては，パラメータ推定を何度も行い，その間作業者は待っている（あるいは 1 つ前の分類器で選ばれた箇所を作業する）必要がある．表 3 のパラメータ推定に要する時間を作業者の待ち時間と考えると，既存手法に対する提案手法の優位性はより顕著になる．

なお，ロジスティック回帰を用いる点予測（LR dual）の計算時間が系列予測（CRFsuite）よりも長い理由は，9 分割交差検定の結果選択されたハイパーパラメータの値による影響が大きい．このパラメータを最適値からずらすことで，系列予測（CRFsuite）よりも短いパラメータ推定時間でより高い精度となった．

次に，単語分割の精度についてである．**表 4** と**表 5** は，一般分野のテストコーパスに対する単語分割の精度である．違いは，UniDic を利用するか否かである．同様に，**表 6** と**表 7** は，適応分野のテストコーパスに対する単語分割の精度である．CRF に基づく系列予測の 2 つを比較すると，すべての条件において文献 3) の精度が CRFsuite を用いる場合より高いので，以下では文献 3) の結果を系列予測の代表として評価する．また，提案手法である点予測による方法では，分類器として線形 SVM を用いる場合の精度がロジスティック回帰（LR dual）を用いる場合よりもわずかながら高いので，SVM を用いる場合を提案手法の代表として評価する．

表 3　計算に要した時間（UniDic あり）
Table 3　Time needed for calculation (with UniDic).

手法	パラメータ推定	一般分野の単語分割	適応分野の単語分割
系列予測（CRFsuite）	416.9 s	7.1 s	1.0 s
点予測（LR dual）	924.9 s	7.3 s	2.0 s
点予測（SVM）	79.9 s	6.5 s	1.4 s

用いた計算機の中央演算処理装置は，Intel Xeon W3580（3.33 GHz）である．なお，文献 6) の単語 3-gram モデルとクラス 3-gram モデルは，モデル構築部分が `Perl` で実装されており，文献 3) の系列予測による方法はすべてが `Java` で実装されているため，この表には含めていない．しかしながら，時間を測定した結果，処理系の違いを考慮しても点予測（SVM）に比べて十分に遅かった．

まず，単語 n-gram モデルおよびクラス n-gram モデルと CRF に基づく系列予測との比較についてである．ほぼすべての実験条件において，CRF が n-gram モデルよりも精度が高い．ただし，表 7 の場合はほぼ同じ精度となっている．いずれのテストコーパスに対しても，カバー率が低い（UniDic なしの）場合に CRF の優位性がより顕著になっている．柔軟な素性選択によって未知語に対してより頑健になっていると考えられる．

次に，系列予測と点予測との比較についてである．すべての実験条件において，点予測が系列予測を上回る結果となっている．このことから，点予測の利点は，系列予測に対して同程度以上の解析精度を保ちつつ，部分的アノテーションコーパスなどの言語資源が利用可能になることであるといえる．この利点を例示するために，次のような 2 つの分野適応戦略を比較した．分野適応時には，UniDic のように適応分野のカバー率を大きく上げる辞書が

表 4　一般分野のテストコーパスに対する単語分割精度（UniDic なし）
Table 4　Word segmentation accuracy of the general corpus (without UniDic).

単語分割手法	単語境界推定精度 [%]	単語認識精度		
		適合率 [%]	再現率 [%]	F 値（$\beta = 1$）
単語 3-gram モデル	99.06	98.20	98.43	98.32
クラス 3-gram モデル	99.12	98.40	98.51	98.46
系列予測（CRFsuite）	98.90	97.94	98.02	97.98
系列予測（文献 3)）	99.25	98.57	98.62	98.59
点予測（LR dual）	99.27	98.68	98.68	98.68
点予測（SVM）	99.30	98.72	98.70	98.71

表 5　一般分野のテストコーパスに対する単語分割精度（UniDic あり）
Table 5　Word segmentation accuracy of the general corpus (with UniDic).

単語分割手法	単語境界推定精度 [%]	単語認識精度		
		適合率 [%]	再現率 [%]	F 値（$\beta = 1$）
単語 3-gram モデル	99.49	99.06	99.07	99.07
クラス 3-gram モデル	99.52	99.12	99.10	99.11
系列予測（CRFsuite）	99.42	98.86	98.83	98.84
系列予測（文献 3)）	99.57	99.18	99.15	99.16
点予測（LR dual）	99.67	99.39	99.30	99.35
点予測（SVM）	99.68	99.41	99.33	99.37

表 6　適応分野のテストコーパスに対する単語分割精度（UniDic なし）
Table 6　Word segmentation accuracy of the specific corpus (without UniDic).

単語分割手法	単語境界推定精度 [%]	単語認識精度		
		適合率 [%]	再現率 [%]	F 値（$\beta = 1$）
単語 3-gram モデル	97.82	96.21	96.82	96.52
クラス 3-gram モデル	97.98	96.59	96.96	96.78
系列予測（CRFsuite）	97.68	95.95	96.23	96.09
系列予測（文献 3)）	98.17	96.86	97.00	96.93
点予測（LR dual）	98.14	96.85	97.16	97.00
点予測（SVM）	98.18	96.91	97.20	97.05

表 7　適応分野のテストコーパスに対する単語分割精度（UniDic あり）
Table 7　Word segmentation accuracy of the specific corpus (with UniDic).

単語分割手法	単語境界推定精度 [%]	単語認識精度		
		適合率 [%]	再現率 [%]	F 値（$\beta = 1$）
単語 3-gram モデル	98.60	97.59	97.86	97.73
クラス 3-gram モデル	98.70	97.77	98.00	97.89
系列予測（CRFSuite）	98.52	97.25	97.63	97.44
系列予測（文献 3)）	98.73	97.75	98.03	97.89
点予測（LR dual）	98.97	98.12	98.49	98.31
点予測（SVM）	99.01	98.16	98.53	98.34

利用可能であることはまれであるので，UniDic を用いない場合（表 6）が初期状態である．1 つ目の分野適応戦略は，適応分野のコーパスに対して単語境界情報を先頭から順に付与していくフルアノテーション（既存手法）であり，2 つ目の戦略は，未知の文字 2-gram に 1 カ所ずつ単語境界情報を付与していく部分的アノテーションである．それぞれの方法によって得られる適応分野のコーパスと一般分野の学習コーパから点予測（SVM）による単語分割器を構築し適応分野での精度（F 値）を計算した．**図 2** はその結果である（部分的アノテーションの場合は，未知の文字 2-gram がなくなって終了）．この結果から，部分的アノテーションの利用は，分野適応の効率を大幅に改善することが分かる．

表 7 の場合に対して文献 3) と点予測（SVM）の結果を具体的に調査した．誤りの傾向において明確な差異は見られないが，点予測のほうが曖昧性が大きい多くの例に対して比較的正しく解析していた[⋆1]．たとえば「って」は，分割するか否かが文脈に依存する．点予測

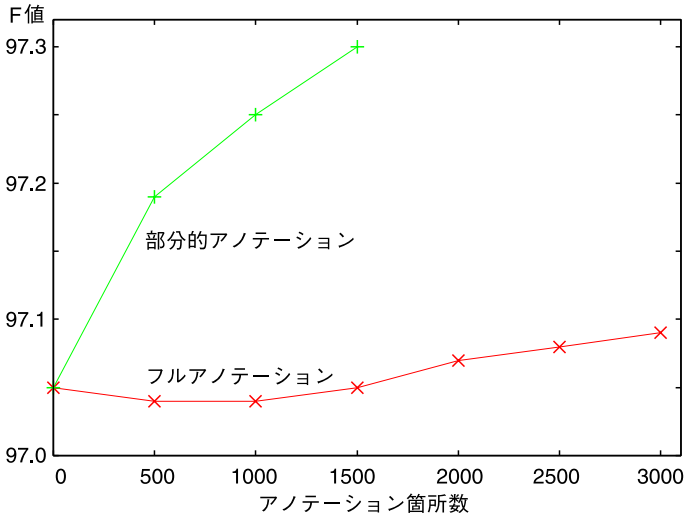

図 2　分野適応時のアノテーション箇所数と F 値の関係（SVM，UniDic なし）
Fig. 2　Relationship between the number of annotations and F measure in a domain adaptation (SVM, without UniDic).

は文献 3) に対してこの判断の精度が高かった．提案手法でも解析を誤るのは主に未知語であった．なかには，「ファミ/〇/ン/の/カセット」[⋆2]のように既知語から仮説を生成するなどある種の方法で正しく分割できそうな誤りも散見された．未知語抽出[25] の研究との融合や未知語モデル[4] の高度化など，包括的な対処が必要である．

最後に，点予測における分類器の違いについてである．ロジスティック回帰を用いると，SVM を用いる場合よりも計算時間が長く，精度はほんのわずかに低下する．しかしながら，各文字間での単語境界確率が得られる．これにより，確率的単語分割[15] のように，単語分割の曖昧性を保持したまま後段の処理に結果を渡すことが可能になるという利点がある．

6. おわりに

本論文では，解析精度やコーパスの準備などを考慮した総合設計として点予測による自動単語分割を提案し，n-gram モデルや系列予測との定量的・定性的比較を行った．提案手法

⋆1 分割の曖昧性が大きい例は，「って」，「られ」，「とも」，「ある」などである．

⋆2 4 文字目の「〇」は原文において伏せ字であり，「コ」であると推察される．

は，部分的単語分割コーパスなどの言語資源を利用して高速かつ安価に分野適応を実現することができるという利点がある．また，自動単語分割の実験結果からは，一般分野と適応分野のそれぞれに対して，提案手法がいずれの既存手法をも上回る精度となることが確認された．以上のことから，提案手法の既存手法に対する優位性が示された．

謝辞　本研究の一部は，科学研究費補助金・若手 A（課題番号：08090047）により行われました．

参 考 文 献

1) 小田裕樹，森　信介，北　研二：文字クラスモデルによる日本語単語分割，自然言語処理，Vol.6, No.7, pp.93–108 (1999).
2) 小田裕樹，北　研二：PPM*言語モデルを用いた日本語単語分割，情報処理学会論文誌，Vol.41, No.3, pp.689–700 (2000).
3) 坪井祐太，森　信介，鹿島久嗣，小田裕樹，松本裕治：日本語単語分割の分野適応のための部分的アノテーションを用いた条件付き確率場の学習，情報処理学会論文誌，Vol.50, No.6, pp.1622–1635 (2009).
4) 永田昌明：統計的言語モデルと N-best 探索を用いた日本語形態素解析法，情報処理学会論文誌，Vol.40, No.9, pp.3420–3431 (1999).
5) 松本裕治：形態素解析システム「茶筌」，情報処理，Vol.41, No.11, pp.1208–1214 (1996).
6) 森　信介，長尾　眞：形態素クラスタリングによる形態素解析精度の向上，自然言語処理，Vol.5, No.2, pp.75–103 (1998).
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 (2004).
8) Ng, H.T. and Low, J.K.: Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?, *Conference on Empirical Methods in Natural Language Processing* (2004).
9) 颯々野学：日本語単語分割を題材としたサポートベクタマシンの能動学習の実験的研究，自然言語処理，Vol.13, No.2, pp.27–41 (2006).
10) 丸山　宏，荻野紫穂，渡辺日出雄：確率的形態素解析，日本ソフトウェア科学会第 8 回大会論文集，pp.177–180 (1991).
11) Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. 18th ICML* (2001).
12) Peng, F., Feng, F. and McCallum, A.: Chinese Segmentation and New Word Detection using Conditional Random Fields, *Proc. 20th International Conference on Computational Linguistics*, pp.562–568 (2004).
13) Brown, P.F., Pietra, V.J.D., deSouza, P.V., Lai, J.C. and Mercer, R.L.: Class-Based *n*-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467–479 (1992).
14) 森　信介，西村雅史，伊東伸泰：クラスに基づく言語モデルのための単語クラスタリング，情報処理学会論文誌，Vol.38, No.11, pp.2200–2208 (1997).
15) 森　信介，宅間大介，倉田岳人：確率的単語分割コーパスからの単語 N-gram 確率の計算，情報処理学会論文誌，Vol.48, No.2, pp.892–899 (2007).
16) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol.39, No.1, pp.1–38 (1977).
17) 前川喜久雄：代表性を有する大規模日本語書き言葉コーパスの構築，人工知能学会誌，Vol.24, No.5, pp.616–622 (2009).
18) Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H. and Den, Y.: Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese, *Proc. 7th International Conference on Language Resources and Evaluation* (2010).
19) 小椋秀樹，小磯花絵，冨士池優美，原　裕：『現代日本語書き言葉均衡コーパス』形態論情報規程集改定版，国立国語研究所内部報告書 edition (2009).
20) 伝　康晴：多様な目的に適した形態素解析システム用電子化辞書，人工知能学会誌，Vol.24, No.5, pp.640–646 (2009).
21) Yu, H.-F., Huang, F.-L. and Lin, C.-J.: Dual coordinate descent methods for logistic regression and maximum entropy models, *Machine Learning* (2010).
22) Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, Vol.9, pp.1871–1874 (2008).
23) Okazaki, N.: CRFsuite A fast implementation of Conditional Random Fields (CRFs) (2010), available from ⟨http://www.chokkan.org/software/crfsuite/⟩.
24) Neubig, G. and Mori, S.: Word-based Partial Annotation for Efficient Corpus Construction, *Proc. 7th International Conference on Language Resources and Evaluation* (2010).
25) 森　信介，長尾　眞：*n* グラム統計によるコーパスからの未知語抽出，情報処理学会論文誌，Vol.39, No.7, pp.2093–2100 (1998).

(平成 23 年 1 月 16 日受付)
(平成 23 年 7 月 8 日採録)

森　　信介（正会員）

1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム（株）入社．2007 年より京都大学学術情報メディアセンター准教授．京都大学博士（工学）．1997 年情報処理学会山下記念研究賞受賞．2010 年情報処理学会論文賞受賞．言語処理学会会員．

ニュービッグ　グラム（正会員）

2005 年米国イリノイ大学アーバナ・シャンペーン校工学部コンピュータ・サイエンス専攻卒業．2010 年京都大学大学院情報学研究科修士課程修了．同年同大学院博士後期課程に進学．現在に至る．自然言語処理に関する研究に従事．

坪井　祐太（正会員）

2002 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年日本アイ・ビー・エム（株）入社．2009 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．日本アイ・ビー・エム東京基礎研究所にてテキストマイニングの研究開発に従事．2010 年情報処理学会論文賞，人工知能学会現場イノベーション賞受賞．言語処理学会会員．